

FOURTH EDITION

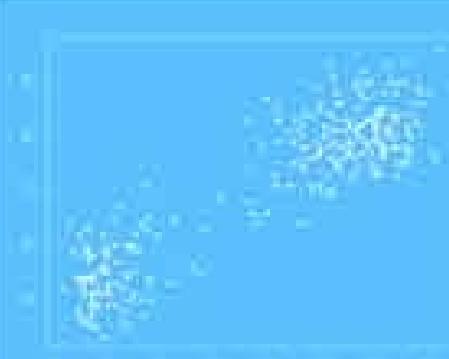


PETER M. LEE

# BAYESIAN STATISTICS

An Introduction

 WILEY



# Contents

[Cover](#)

[Title Page](#)

[Copyright](#)

[Dedication](#)

[Preface](#)

[Preface to the First Edition](#)

[Chapter 1: Preliminaries](#)

[1.1 Probability and Bayes' Theorem](#)

[1.2 Examples on Bayes' Theorem](#)

[1.3 Random variables](#)

[1.4 Several random variables](#)

[1.5 Means and variances](#)

[1.6 Exercises on Chapter 1](#)

[Chapter 2: Bayesian inference for the normal distribution](#)

[2.1 Nature of Bayesian inference](#)

[2.2 Normal prior and likelihood](#)

[2.3 Several normal observations with a normal prior](#)

[2.4 Dominant likelihoods](#)

[2.5 Locally uniform priors](#)

[2.6 Highest density regions](#)

[2.7 Normal variance](#)

[2.8 HDRs for the normal variance](#)

[2.9 The role of sufficiency](#)

[2.10 Conjugate prior distributions](#)

[2.11 The exponential family](#)

[2.12 Normal mean and variance both unknown](#)

[2.13 Conjugate joint prior for the normal distribution](#)

[2.14 Exercises on Chapter 2](#)

## [Chapter 3: Some other common distributions](#)

[3.1 The binomial distribution](#)

[3.2 Reference prior for the binomial likelihood](#)

[3.3 Jeffreys' rule](#)

[3.4 The Poisson distribution](#)

[3.5 The uniform distribution](#)

[3.6 Reference prior for the uniform distribution](#)

[3.7 The tramcar problem](#)

[3.8 The first digit problem; invariant priors](#)

[3.9 The circular normal distribution](#)

[3.10 Approximations based on the likelihood](#)

[3.11 Reference posterior distributions](#)

[3.12 Exercises on Chapter 3](#)

## [Chapter 4: Hypothesis testing](#)

[4.1 Hypothesis testing](#)

[4.2 One-sided hypothesis tests](#)

[4.3 Lindley's method](#)

[4.4 Point \(or sharp\) null hypotheses with prior information](#)

[4.5 Point null hypotheses for the normal distribution](#)

[4.6 The Doogian philosophy](#)

## 4.7 Exercises on Chapter 4

## Chapter 5: Two-sample problems

- 5.1 Two-sample problems – both variances unknown
- 5.2 Variances unknown but equal
- 5.3 Variances unknown and unequal (Behrens–Fisher problem)
- 5.4 The Behrens–Fisher controversy
- 5.5 Inferences concerning a variance ratio
- 5.6 Comparison of two proportions; the \$2\times 2\$ table
- 5.7 Exercises on Chapter 5

## Chapter 6: Correlation, regression and the analysis of variance

- 6.1 Theory of the correlation coefficient
- 6.2 Examples on the use of the correlation coefficient
- 6.3 Regression and the bivariate normal model
- 6.4 Conjugate prior for the bivariate regression model
- 6.5 Comparison of several means – the one way model
- 6.6 The two way layout
- 6.7 The general linear model
- 6.8 Exercises on Chapter 6

## Chapter 7: Other topics

- 7.1 The likelihood principle
- 7.2 The stopping rule principle
- 7.3 Informative stopping rules
- 7.4 The likelihood principle and reference priors
- 7.5 Bayesian decision theory
- 7.6 Bayes linear methods
- 7.7 Decision theory and hypothesis testing

[7.8 Empirical Bayes methods](#)

[7.9 Exercises on Chapter 7](#)

## [Chapter 8: Hierarchical models](#)

[8.1 The idea of a hierarchical model](#)

[8.2 The hierarchical normal model](#)

[8.3 The baseball example](#)

[8.4 The Stein estimator](#)

[8.5 Bayesian analysis for an unknown overall mean](#)

[8.6 The general linear model revisited](#)

[8.7 Exercises on Chapter 8](#)

## [Chapter 9: The Gibbs sampler and other numerical methods](#)

[9.1 Introduction to numerical methods](#)

[9.2 The EM algorithm](#)

[9.3 Data augmentation by Monte Carlo](#)

[9.4 The Gibbs sampler](#)

[9.5 Rejection sampling](#)

[9.6 The Metropolis–Hastings algorithm](#)

[9.7 Introduction to WinBUGS and OpenBUGS](#)

[9.8 Generalized linear models](#)

[9.9 Exercises on Chapter 9](#)

## [Chapter 10: Some approximate methods](#)

[10.1 Bayesian importance sampling](#)

[10.2 Variational Bayesian methods: simple case](#)

[10.3 Variational Bayesian methods: general case](#)

[10.4 ABC: Approximate Bayesian Computation](#)

[10.5 Reversible jump Markov chain Monte Carlo](#)

## 10.6 Exercises on Chapter 10

## Appendix A: Common statistical distributions

[A.1 Normal distribution](#)

[A.2 Chi-squared distribution](#)

[A.3 Normal approximation to chi-squared](#)

[A.4 Gamma distribution](#)

[A.5 Inverse chi-squared distribution](#)

[A.6 Inverse chi distribution](#)

[A.7 Log chi-squared distribution](#)

[A.8 Student's t distribution](#)

[A.9 Normal/chi-squared distribution](#)

[A.10 Beta distribution](#)

[A.11 Binomial distribution](#)

[A.12 Poisson distribution](#)

[A.13 Negative binomial distribution](#)

[A.14 Hypergeometric distribution](#)

[A.15 Uniform distribution](#)

[A.16 Pareto distribution](#)

[A.17 Circular normal distribution](#)

[A.18 Behrens' distribution](#)

[A.19 Snedecor's F distribution](#)

[A.20 Fisher's z distribution](#)

[A.21 Cauchy distribution](#)

[A.22 The probability that one beta variable is greater than another](#)

[A.23 Bivariate normal distribution](#)

[A.24 Multivariate normal distribution](#)

[A.25 Distribution of the correlation coefficient](#)

## Appendix B: Tables

## [Appendix C: R programs](#)

## [Appendix D: Further reading](#)

[D.1 Robustness](#)

[D.2 Nonparametric methods](#)

[D.3 Multivariate estimation](#)

[D.4 Time series and forecasting](#)

[D.5 Sequential methods](#)

[D.6 Numerical methods](#)

[D.7 Bayesian networks](#)

[D.8 General reading](#)

## [References](#)

## [Index](#)

# 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](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***

Lee, Peter M.

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

Includes bibliographical references and index.

ISBN 978-1-11833257-3 (pbk.)

1. Bayesian statistical decision theory. I. Title.

QA279.5.L44 2012

519.5'42–dc23

2012007007

201200/00/

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

ISBN: 9781118332573

*To The Memory of My Mother and of My Father*

## 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 McGrawne (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 hoc’ery’.

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)].<sup>1</sup> 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 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

<sup>1</sup> 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).

# 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  $B$ );

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

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

$A \setminus B = \{x; x \in A, \text{but } x \notin B\}$  (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

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

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

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

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

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

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

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

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

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

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

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

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

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  $H$  is true. It seems reasonable that this measure should abide by the following axioms:

- P1  $P(E|H) = 0$  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  $P(E|FH)P(F|H) = P(EF|H)$ .

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

$$P(C|H) = 12/30,$$

$$P(MC|H) = 10/30.$$

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

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

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 while in certain circumstances we may believe in symmetries or in equally probable cases, we cannot base a general definition of probability on such arguments.

It is convenient to use a stronger form of axiom P3 in many contexts, namely,

$$P3^* \quad P\left(\bigcup_{n=1}^{\infty} E_n | H\right) = \sum_{n=1}^{\infty} P(E_n | H)$$

whenever the  $(E_n)$  are exclusive events given  $H$ . There is no doubt of the mathematical simplifications that result from this assumption, but we are supposed to be modelling our degrees of belief and it is questionable whether these have to obey this form of the axiom. Indeed, one of the greatest advocates of Bayesian theory, Bruno de Finetti, was strongly against the use of P3\*. His views can be found in de Finetti (1972, Section 5.32) or in de Finetti (1974–1975, Section 3.11.3).

There is certainly some arbitrariness about P3\*, which is sometimes referred to as an assumption of  $\sigma$ -additivity, in that it allows additivity over some but not all infinite collections of events (technically over countable but not over uncountable collections). However, it is impossible in a lot of contexts to allow additivity over any (arbitrary) collection of events. Thus, if we want a model for picking a point ‘completely at random’ from the unit interval

$$[0, 1] = \{x; 0 \leq x \leq 1\},$$

it seems reasonable that the probability that the point picked is in any particular sub-interval of the unit interval should equal the length of that sub-interval. However, this clearly implies that the probability of picking *any* one particular  $x$  is zero (since any such  $x$  belongs to intervals of arbitrarily small lengths). But the probability that *some*  $x$  is picked is unity, and it is impossible to get one by adding a lot of zeroes.

Mainly because of its mathematical convenience, we shall assume P3\* while being aware of the problems.

### 1.1.3 ‘Unconditional’ probability

Strictly speaking, there is, in my view, no such thing as an unconditional probability. However, it often happens that many probability statements are made conditional on everything that is part of an individual’s knowledge at a particular time, and when many statements are to be made conditional on the same event, it makes for cumbersome notation to refer to this same conditioning event every time. There are also cases where we have so much experimental data in circumstances judged to be relevant to a particular situation that there is a fairly general agreement as to the probability of an event. Thus, in tossing a coin, you and I both have experience of tossing similar coins many times and so are likely to believe that ‘heads’ is approximately as likely as not, so that the probability of ‘heads’ is approximately  $\frac{1}{2}$  given your knowledge or mine.

In these cases we write

$$P(E) \text{ for } P(E|\Omega),$$

$$P(E|F) \text{ for } P(E|F\Omega),$$

where  $\Omega$  is the set of possibilities consistent with the sum total of data available to the individual or individuals concerned. We usually consider sets  $F$  for which  $F \subset \Omega$ , so that  $F\Omega = F$ . It easily follows from the axioms that

$$0 \leq P(E) \leq 1,$$

$$P(\Omega) = 1, \quad P(\emptyset) = 0,$$

$$P\left(\bigcup_{n=1}^{\infty} E_n\right) = \sum_{n=1}^{\infty} P(E_n)$$

whenever the  $(E_n)$  are exclusive events (or more properly whenever they are exclusive events given  $\Omega$ ), and

$$P(E|F)P(F) = P(EF).$$

Many books begin by asserting that unconditional probability is an intuitive notion and use the latter formula in the form

$$P(E|F) = P(EF)/P(F) \quad (\text{provided } P(F) \neq 0)$$

to define conditional probability.

### 1.1.4 Odds

It is sometimes convenient to use a language more familiar to bookmakers to

express probabilistic statements. We define the *odds on E against F given H* as the ratio

$$\mathsf{P}(E|H)/\mathsf{P}(F|H) \text{ to } 1$$

or equivalently

$$\mathsf{P}(E|H) \text{ to } \mathsf{P}(F|H).$$

A reference to the odds on  $E$  against  $F$  with no mention of  $H$  is to be interpreted as a reference to the odds on  $E$  against  $F$  given  $\Omega$ , where  $\Omega$  is some set of background knowledge as above.

Odds do not usually have properties as simple as probabilities, but sometimes, for example, in connection with Bayesian tests of hypotheses, they are more natural to consider than separate probabilities.

### 1.1.5 Independence

Two events  $E$  and  $F$  are said to be *independent* given  $H$  if

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

From axiom P4, it follows that if  $\mathsf{P}(F|H) \neq 0$  this condition is equivalent to

$$\mathsf{P}(E|FH) = \mathsf{P}(E|H),$$

so that if  $E$  is independent of  $F$  given  $H$  then the extra information that  $F$  is true does not alter the probability of  $E$  from that given  $H$  alone, and this gives the best intuitive idea as to what independence means. However, the restriction of this interpretation to the case where  $\mathsf{P}(F|H) \neq 0$  makes the original equation slightly more general.

More generally, a sequence  $(E_n)$  of events is said to be *pairwise independent* given  $H$  if

$$\mathsf{P}(E_m E_n|H) = \mathsf{P}(E_m|H)\mathsf{P}(E_n|H) \text{ for } m \neq n$$

and is said to consist of *mutually independent* events given  $H$  if for every finite subset of them

$$\mathsf{P}(E_{n_1} E_{n_2} \dots E_{n_k}|H) = \mathsf{P}(E_{n_1}|H)\mathsf{P}(E_{n_2}|H) \dots \mathsf{P}(E_{n_k}|H).$$

You should be warned that pairwise independence does not imply mutual independence and that

$$\mathsf{P}(E_1 E_2 \dots E_n|H) = \mathsf{P}(E_1|H)\mathsf{P}(E_2|H) \dots \mathsf{P}(E_n|H)$$

is not enough to ensure that the finite sequence  $E_1, E_2, \dots, E_n$  consists of mutually independent events given  $H$ .

Naturally, if no conditioning event is explicitly mentioned, the probabilities concerned are conditional on  $\Omega$  as defined above.

## 1.1.0 SOME SIMPLE CONSEQUENCES OF THE AXIOMS, BAYES

### Theorem

We have already noted a few consequences of the axioms, but it is useful at this point to note a few more. We first note that it follows simply from P4 and P2 and the fact that  $HH=H$  that

$$P(E|H) = P(EH|H)$$

and in particular

$$P(E) = P(E\Omega).$$

Next note that if, given  $H$ ,  $E$  implies  $F$ , that is  $EH \subset F$  and so  $EFH=EH$ , then by P4 and the aforementioned equation

$$P(E|FH)P(F|H) = P(EF|H) = P(EFH|H) = P(EH|H) = P(E|H).$$

From this and the fact that  $P(E|FH) \leq 1$  it follows that if, given  $H$ ,  $E$  implies  $F$ , then

$$P(E|H) \leq P(F|H).$$

In particular, if  $E$  implies  $F$  then

$$P(E|F)P(F) = P(E),$$

$$P(E) \leq P(F).$$

For the rest of this subsection, we can work in terms of ‘unconditional’ probabilities, although the results are easily generalized. Let  $(H_n)$  be a sequence of exclusive and exhaustive events, and let  $E$  be any event. Then

$$P(E) = \sum_n P(E|H_n)P(H_n)$$

since by P4 the terms on the right-hand side are  $P(EH_n)$ , allowing us to deduce the result from P3\*. This result is sometimes called the *generalized addition law* or the *law of the extension of the conversation*.

The key result in the whole book is Bayes’ Theorem. This is simply deduced as follows. Let  $(H_n)$  be a sequence of events. Then by P4

$$P(H_n|E)P(E) = P(EH_n) = P(H_n)P(E|H_n),$$

so that provided  $P(E) \neq 0$

$$P(H_n|E) \propto P(H_n)P(E|H_n).$$

This relationship is one of several ways of stating Bayes’ Theorem, and is probably the best way in which to remember it. When we need the constant of proportionality, we can easily see from the above that it is  $1/P(E)$ .

It should be clearly understood that there is nothing controversial about Bayes’ Theorem as such. It is frequently used by probabilists and statisticians, whether or not they are Bayesians. The distinctive feature of Bayesian statistics is the application of the theorem in a wider range of circumstances than is usual in

classical statistics. In particular, Bayesian statisticians are always willing to talk of the probability of a hypothesis, both unconditionally (its *prior probability*) and given some evidence (its *posterior probability*), whereas other statisticians will only talk of the probability of a hypothesis in restricted circumstances.

When  $(H_n)$  consists of exclusive and exhaustive events, we can combine the last two results to see that

$$P(H_n|E) = \frac{P(H_n)P(E|H_n)}{\sum_m P(H_m)P(E|H_m)}.$$

A final result that we will find useful from time to time is the *generalized multiplication law*, which runs as follows. If  $H_1, H_2, \dots, H_n$  are any events then

$$P(H_1 H_2 \dots H_n) = P(H_1)P(H_2|H_1)P(H_3|H_1 H_2) \dots \\ P(H_n|H_1 H_2 \dots H_{n-1})$$

provided all the requisite conditional probabilities are defined, which in practice they will be provided  $P(H_1 H_2 \dots H_{n-1}) \neq 0$ . This result is easily proved by repeated application of P4.

## 1.2 Examples on Bayes' Theorem

### 1.2.1 The Biology of Twins

Twins can be either monozygotic ( $M$ ) (i.e. developed from a single egg) or dizygotic ( $D$ ). Monozygotic twins often look very similar and then are referred to as identical twins, but it is not always the case that one finds very striking similarities between monozygotic twins, while some dizygotic twins can show marked resemblances. Whether twins are monozygotic or dizygotic is not, therefore, a matter which can be settled simply by inspection. However, it is always the case that monozygotic twins are of the same sex, whereas dizygotic twins can be of opposite sex. Hence, assuming that the two sexes are equally probable, if the sexes of a pair of twins are denoted  $GG$ ,  $BB$  or  $GB$  (note  $GB$  is indistinguishable from  $BG$ )

$$P(GG|M) = P(BB|M) = \frac{1}{2}, \quad P(GB|M) = 0,$$

$$P(GG|D) = P(BB|D) = \frac{1}{4}, \quad P(GB|D) = \frac{1}{2}.$$

It follows that

$$\begin{aligned} P(GG) &= P(GG|M)P(M) + P(GG|D)P(D) \\ &= \frac{1}{2}P(M) + \frac{1}{4}\{1 - P(M)\} \end{aligned}$$

from which it can be seen that

$$P(M) = 4P(GG) - 1,$$

so that although it is not easy to be certain whether a particular pair are monozygotic or not, it is easy to discover the *proportion* of monozygotic twins in the whole population of twins simply by observing the sex distribution among *all* twins.

### 1.2.2 A political example

The following example is a simplified version of the situation just before the time of the British national referendum as to whether the United Kingdom should remain part of the European Economic Community which was held in 1975. Suppose that at that date, which was shortly after an election which the Labour Party had won, the proportion of the electorate supporting Labour ( $L$ ) stood at 52%, while the proportion supporting the Conservatives ( $C$ ) stood at 48% (it being assumed for simplicity that support for all other parties was negligible, although this was far from being the case). There were many opinion polls taken at the time, so we can take it as known that 55% of Labour supporters and 85% of Conservative voters intended to vote ‘Yes’ ( $Y$ ) and the remainder intended to vote ‘No’ ( $N$ ). Suppose that knowing all this you met someone at the time who said that she intended to vote ‘Yes’, and you were interested in knowing which political party she supported. If this information were all you had available, you could reason as follows:

$$\begin{aligned} P(L|Y) &= \frac{P(Y|L)P(L)}{P(Y|L)P(L) + P(Y|C)P(C)} \\ &= \frac{(0.55)(0.52)}{(0.55)(0.52) + (0.85)(0.48)} \\ &= 0.41. \end{aligned}$$

### 1.2.3 A warning

In the case of *Connecticut v. Teal* [see DeGroot *et al.* (1986, p. 9)], a case of alleged discrimination on the basis of a test to determine eligibility for promotion was considered. It turned out that of those taking the test 48 were black ( $B$ ) and 259 were white ( $W$ ), so that if we consider a random person taking the test

$$P(B) = 48/307 = 0.16, \quad P(W) = 259/307 = 0.84.$$

Of the blacks taking the test, 26 passed ( $P$ ) and the rest failed ( $F$ ), whereas of the whites, 206 passed and the rest failed, so that altogether 232 people passed.

Hence,

$$P(B|P) = 26/232 = 0.11, \quad P(W|P) = 206/232 = 0.89.$$

There is a temptation to think that these are the figures which indicate the possibility of discrimination. Now there certainly is a case for saying that there was discrimination in this case, *but* the figures that should be considered are

$$P(P|B) = 26/48 = 0.54, \quad P(P|W) = 206/259 = 0.80.$$

It is easily checked that the probabilities here are related by Bayes' Theorem. It is worth while spending a while playing with hypothetical figures to convince yourself that the fact that  $P(B|P)$  is less than  $P(W|P)$  is irrelevant to the real question as to whether  $P(P|B)$  is less than  $P(P|W)$  – it might or might not be depending on the rest of the relevant information, that is on  $P(B)$  and  $P(W)$ . The fallacy involved arises as the first of two well-known fallacies in criminal law which are both well summarized by Aitken (1996) (see also Aitken and Taroni, 2004, and Dawid, 1994) as follows:

Suppose a crime has been committed. Blood is found at the scene for which there is no innocent explanation. It is of a type which is present in 1% of the population. The prosecutor may then state:

'There is a 1% chance that the defendant would have the crime blood type if he were innocent. Thus, there is a 99% chance that he is guilty'.

Alternatively, the defender may state:

'This crime occurred in a city of 800,000 people. This blood type would be found in approximately 8000 people. The evidence has provided a probability of 1 in 8000 that the defendant is guilty and thus has no relevance.'

The first of these is known as the *prosecutor's fallacy* or the *fallacy of the transposed conditional* and, as pointed out above, in essence it consists in quoting the probability  $P(E|I)$  instead of  $P(I|E)$ . The two are, however, equal if and only if the prior probability  $P(I)$  happens to equal  $P(E)$ , which will only rarely be the case.

The second is the *defender's fallacy* which consists in quoting  $P(G|E)$  without regard to  $P(G)$ . In the case considered by Aitken, the prior odds in favour of guilt are

$$P(G)/P(I) = 1/799\,999,$$

while the posterior odds are

$$P(G|E)/P(I|E) = 1/7\,999.$$

Such a large change in the odds is, in Aitken's words 'surely of relevance'. But, again in Aitken's words, 'Of course, it may not be enough to find the suspect

guilty'.

As a matter of fact, Bayesian statistical methods are increasingly used in a legal context. Useful references are Balding and Donnelly (1995), Foreman, Smith and Evett (1997), Gastwirth (1988) and Fienberg (1989).

## 1.3 Random variables

### 1.3.1 Discrete random variables

As explained in Section 1.1, there is usually a set  $\Omega$  representing the possibilities consistent with the sum total of data available to the individual or individuals concerned. Now suppose that with each elementary event  $\omega$  in  $\Omega$ , there is an integer  $\tilde{m}(\omega)$  which may be positive, negative or zero. In the jargon of mathematics, we have a function  $\tilde{m}$  mapping  $\Omega$  to the set  $\mathbb{Z}$  of all (signed) integers. We refer to the function as a *random variable* or an *r.v.*

A case arising in the context of the very first example we discussed, which was about tossing a red die and a blue die, is the integer representing the sum of the spots showing. In this case,  $\omega$  might be ‘red three, blue two’ and then  $\tilde{m}(\omega)$  would be 5. Another case arising in the context of the second (political) example is the Labour majority (represented as a negative integer should the Conservatives happen to win), and here  $\omega$  might be ‘Labour 350, Conservative 250’ in which case  $\tilde{m}(\omega)$  would be 100.

Rather naughtily, probabilists and statisticians tend not to mention the elementary event  $\omega$  of which  $\tilde{m}(\omega)$  is a function and instead just write  $\tilde{m}$  for  $\tilde{m}(\omega)$ . The reason is that what usually matters is the value of  $\tilde{m}$  rather than the nature of the elementary event  $\omega$ , the definition of which is in any case dependent on the context, as noted earlier, in the discussion of elementary events. Thus, we write

$$P(\tilde{m} = m) \text{ for } P(\{\omega; \tilde{m}(\omega) = m\})$$

for the probability that the random variable  $\tilde{m}$  takes the particular value  $m$ . It is a useful convention to use the same lower-case letter and drop the tilde ( $\sim$ ) to denote a particular value of a random variable. An alternative convention used by some statisticians is to use capital letters for random variables and corresponding lower case letters for typical values of these random variables, but in a Bayesian context we have so many quantities that are regarded as random variables that this convention is too restrictive. Even worse than the habit of dropping mention of  $\omega$  is the tendency to omit the tilde and so use the same

notation for a random variable and for a typical value of it. While failure to mention  $\omega$  rarely causes any confusion, the failure to distinguish between random variables and typical values of these random variables can, on occasion, result in real confusion. When there is any possibility of confusion, the tilde will be used in the text, but otherwise it will be omitted. Also, we will use

$$p(m) = P(\tilde{m} = m) = P(\{\omega; \tilde{m}(\omega) = m\})$$

for the probability that the random variable  $\tilde{m}$  takes the value  $m$ . When there is only one random variable we are talking about, this abbreviation presents few problems, but when we have a second random variable  $\tilde{n}$  and write

$$p(n) = P(\tilde{n} = n) = P(\{\omega; \tilde{n}(\omega) = n\})$$

then ambiguity can result. It is not clear in such a case what  $p(5)$  would mean, or indeed what  $p(i)$  would mean (unless it refers to  $p(\tilde{i} = i)$  where  $\tilde{i}$  is yet a third random variable). When it is necessary to resolve such an ambiguity, we will use

$$p_{\tilde{m}}(m) = P(\tilde{m} = m) = P(\{\omega; \tilde{m}(\omega) = m\}),$$

so that, for example,  $p_{\tilde{m}}(5)$  is the probability that  $\tilde{m}$  is 5 and  $p_{\tilde{n}}(i)$  is the probability that  $\tilde{n}$  equals  $i$ . Again, all of this seems very much more confusing than it really is – it is usually possible to conduct arguments quite happily in terms of  $p(m)$  and  $p(n)$  and substitute numerical values at the end if and when necessary.

You could well object that you would prefer a notation that was free of ambiguity, and if you were to do so, I should have a lot of sympathy. But the fact is that constant references to  $\tilde{m}(\omega)$  and  $p_{\tilde{m}}(m)$  rather than to  $m$  and  $p(m)$  would clutter the page and be unhelpful in another way.

We refer to the sequence  $(p(m))$  as the *(probability) density (function)* or *pdf* of the random variable  $m$  (strictly  $\tilde{m}$ ). The random variable is said to have a distribution (of probability) and one way of describing a distribution is by its pdf. Another is by its *(cumulative) distribution function*, or *cdf* or *df*, defined by

$$F(m) = F_{\tilde{m}}(m) = P(\tilde{m} \leq m) = P(\{\omega; \tilde{m}(\omega) \leq m\}) = \sum_{k \leq m} p_{\tilde{m}}(k).$$

Because the pdf has the obvious properties

$$p(m) \geq 0, \quad \sum_m p(m) = 1$$

the df is (weakly) increasing, that is

$$F(m) \leq F(m') \quad \text{if} \quad m \leq m'$$

and moreover

$$\lim_{m \rightarrow -\infty} F(m) = 0, \quad \lim_{m \rightarrow \infty} F(m) = 1.$$

### 1.3.2 The binomial distribution

A simple example of such a distribution is the *binomial distribution* (see Appendix A). Suppose, we have a sequence of trials each of which, independently of the others, results in success ( $S$ ) or failure ( $F$ ), the probability of success being a constant  $\pi$  (such trials are sometimes called Bernoulli trials). Then the probability of any particular sequence of  $n$  trials in which  $k$  result in success is

$$\pi^k(1 - \pi)^{n-k},$$

so that allowing for the  $\binom{n}{k}$  ways in which  $k$  successes and  $n-k$  failures can be ordered, the probability that a sequence of  $n$  trials results in  $k$  successes is

$$\binom{n}{k} \pi^k (1 - \pi)^{n-k} \quad (0 \leq k \leq n).$$

If then  $k$  (strictly  $\tilde{k}$ ) is a random variable defined as the number of successes in  $n$  trials, then

$$p(k) = \binom{n}{k} \pi^k (1 - \pi)^{n-k}.$$

Such a distribution is said to be binomial of index  $n$  and parameter  $\pi$ , and we write

$$k \sim B(n, \pi)$$

[or strictly  $\tilde{k} \sim B(n, \pi)$ ].

We note that it follows immediately from the definition that if  $x$  and  $y$  are independent and  $x \sim B(m, \pi)$  and  $y \sim B(n, \pi)$  then  $x + y \sim B(m + n, \pi)$ .

### 1.3.3 Continuous random variables

So far, we have restricted ourselves to random variables which take only integer values. These are particular cases of *discrete* random variables. Other examples of discrete random variables occur, for example, a measurement to the nearest quarter-inch which is subject to a distribution of error, but these can nearly always be changed to integer-valued random variables (in the given example simply by multiplying by 4). More generally, we can suppose that with each elementary event  $\omega$  in  $\Omega$  there is a real number  $\tilde{x}(\omega)$ . We can define the (cumulative) distribution function, cdf or df of  $\tilde{x}$  by

$$F(x) = P(\tilde{x} \leq x) = P(\{\omega; \tilde{x}(\omega) \leq x\}).$$

As in the discrete case the df is (weakly) increasing, that is

$$F(x) \leq F(x') \quad \text{if } x \leq x'$$

and moreover

$$\lim_{x \rightarrow -\infty} F(x) = 0, \quad \lim_{x \rightarrow \infty} F(x) = 1.$$

It is usually the case that when  $\tilde{x}$  is not discrete there exists a function  $p(x)$ , or more strictly  $p_{\tilde{x}}(x)$ , such that

$$F(x) = \int_{-\infty}^x p_{\tilde{x}}(\xi) d\xi$$

in which case  $p(x)$  is called a (probability) density (function) or pdf. When this is so,  $x$  (strictly  $\tilde{x}$ ) is said to have a continuous distribution (or more strictly an absolutely continuous distribution). Of course, in the continuous case  $p(x)$  is not itself interpretable directly as a probability, but for small  $\delta x$

$$p(x)\delta x \cong P(x < \tilde{x} \leq x + \delta x) = P(\{\omega; x < \tilde{x}(\omega) \leq x + \delta x\}).$$

The quantity  $p(x)\delta x$  is sometimes referred to as the *probability element*. Note that letting  $\delta x \rightarrow 0$  this implies that

$$P(\tilde{x} = x) = 0$$

for every particular value  $x$ , in sharp contrast to the discrete case. We can also use the above approximation if  $y$  is some one-to-one function of  $x$ , for example

$$y = g(x).$$

Then if values correspond in an obvious way

$$P(y < \tilde{y} \leq y + \delta y) = P(x < \tilde{x} \leq x + \delta x)$$

which on substituting in the above relationship gives in the limit

$$p(y)|dy| = p(x)|dx|$$

which is the rule for *change of variable* in probability densities. (It is not difficult to see that, because the modulus signs are there, the same result is true if  $F$  is a strictly decreasing function of  $x$ ). Another way of getting at this rule is by differentiating the obvious equation

$$F(y) = F(x)$$

[strictly  $F_{\tilde{y}}(y) = F_{\tilde{x}}(x)$ ] which holds whenever  $y$  and  $x$  are corresponding values, that is  $y=g(x)$ . We should, however, beware that these results need modification if  $g$  is *not* a one-to-one function. In the continuous case, we can find the density from the df by differentiation, namely

$$p(x) = F'(x) = dF(x)/dx.$$

Although there are differences, there are many similarities between the discrete and the continuous cases, which we try to emphasize by using the same notation in both cases. We note that

$$F(x) = \sum_{\xi \leq x} p_{\tilde{x}}(\xi)$$

in the discrete case, but

$$F(x) = \int_{-\infty}^x p_{\tilde{x}}(\xi) d\xi$$

in the continuous case. The discrete case is slightly simpler in one way in that no complications arise over change of variable, so that

$$p(y) = p(x)$$

if  $y$  and  $x$  are corresponding values, that is  $y=g(x)$ .

### 1.3.4 The normal distribution

The most important example of a continuous distribution is the so-called *normal* or Gaussian distribution. We say that  $z$  has a *standard normal* distribution if

$$p(z) = (2\pi)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}z^2\right)$$

and when this is so we write

$$z \sim N(0, 1).$$

The density of this distribution is the familiar bell-shaped curve, with about two-thirds of the area between  $-1$  and  $1$ , 95% of the area between  $-2$  and  $2$  and almost all of it between  $-3$  and  $3$ . Its distribution function is

$$\Phi(z) = \int_{-\infty}^z (2\pi)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\zeta^2\right) d\zeta.$$

More generally, we say that  $x$  has a normal distribution, denoted

$$x \sim N(\mu, \phi)$$

if

$$x = \mu + z\sqrt{\phi}$$

where  $z$  is as aforementioned, or equivalently if

$$p(x) = (2\pi\phi)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(x - \mu)^2/\phi\right\}.$$

The normal distribution is encountered almost at every turn in statistics. Partly this is because (despite the fact that its density may seem somewhat barbaric at first sight) it is in many contexts the easiest distribution to work with, but this is not the whole story. The *Central Limit Theorem* says (roughly) that if a random variable can be expressed as a sum of a large number of components no one of which is likely to be much bigger than the others, these components being approximately independent, then this sum will be approximately normally distributed. Because of this theorem, an observation which has an error contributed to by many minor causes is likely to be normally distributed. Similar reasons can be found for thinking that in many circumstances we would expect observations to be approximately normally distributed, and this turns out to be the case, although there are exceptions. This is especially useful in cases where we want to make inferences about a population mean.

### 1.5.5 Mixed random variables

While most commonly occurring random variables are discrete or continuous, there are exceptions, for example the time you have to wait until you are served in a queue, which is zero with a positive probability (if the queue is empty when you arrive), but otherwise is spread over a continuous range of values. Such a random variable is said to have a *mixed* distribution.

## 1.4 Several random variables

### 1.4.1 Two discrete random variables

Suppose that with each elementary event  $\omega$  in  $\Omega$ , we can associate a pair of integers  $(\tilde{m}(\omega), \tilde{n}(\omega))$ . We write

$$p(m, n) = \mathbb{P}(\tilde{m} = m, \tilde{n} = n) = \mathbb{P}(\{\omega; \tilde{m}(\omega) = m, \tilde{n}(\omega) = n\}).$$

Strictly speaking,  $p(m, n)$  should be written as  $p_{\tilde{m}, \tilde{n}}(m, n)$  for reasons discussed earlier, but this degree of pedantry in the notation is rarely necessary. Clearly

$$p(m, n) \geq 0, \quad \sum_m \sum_n p(m, n) = 1.$$

The sequence  $(p(m, n))$  is said to be a *bivariate (probability) density (function)* or *bivariate pdf* and is called the *joint pdf* of the random variables  $m$  and  $n$  (strictly  $\tilde{m}$  and  $\tilde{n}$ ). The corresponding *joint distribution function*, *joint cdf* or *joint df* is

$$F(m, n) = \sum_{k \leq m} \sum_{l \leq n} p_{\tilde{m}, \tilde{n}}(k, l).$$

Clearly, the density of  $m$  (called its *marginal density*) is

$$p(m) = \sum_n p(m, n).$$

We can also define a conditional distribution for  $n$  given  $m$  (strictly for  $\tilde{n}$  given  $\tilde{m} = m$ ) by allowing

$$\begin{aligned} p(n|m) &= \mathbb{P}(\tilde{n} = n | \tilde{m} = m) = \mathbb{P}(\{\omega; \tilde{n}(\omega) = n\} | \{\omega; \tilde{m}(\omega) = m\}) \\ &= p(m, n)/p(m), \quad \text{provided } p(m) \neq 0 \end{aligned}$$

to define the *conditional (probability) density (function)* or *conditional pdf*. This represents our judgement as to the chance that  $\tilde{n}$  takes the value  $n$  given that  $\tilde{m}$  is known to have the value  $m$ . If it is necessary to make our notation absolutely precise, we can always write

$$p_{\tilde{m}|\tilde{n}}(m|n),$$

so, for example,  $p_{\tilde{m}|\tilde{n}}(4|3)$  is the probability that  $m$  is 4 given  $\tilde{n}$  is 3, but

$p_{\tilde{n}|\tilde{m}}(4|3)$  is the probability that  $\tilde{n}$  is 4 given that  $\tilde{m}$  takes the value 3, but it should be emphasized that we will not often need to use the subscripts. Evidently

$$p(n|m) \geq 0, \quad \sum p(n|m) = 1,$$

and

$$p(n) = \sum p(m, n) = \sum p(m)p(n|m).$$

We can also define a *conditional distribution function* or *conditional df* by

$$\begin{aligned} F(n|m) &= \mathbb{P}(\tilde{n} \leq n | \tilde{m} = m) = \mathbb{P}(\{\omega; \tilde{n}(\omega) \leq n\} | \{\omega; \tilde{m}(\omega) = m\}) \\ &= \sum_{k \leq n} p_{\tilde{n}|\tilde{m}}(k|m). \end{aligned}$$

## 1.4.2 Two continuous random variables

As in Section 1.4, we have begun by restricting ourselves to integer values, which is more or less enough to deal with any discrete cases that arise. More generally, we can suppose that with each elementary event  $\omega$  in  $\Omega$ , we can associate a pair  $(\tilde{x}(\omega), \tilde{y}(\omega))$  of real numbers. In this case, we define the *joint distribution function* or *joint df* as

$$F(x, y) = \mathbb{P}(\tilde{x} \leq x, \tilde{y} \leq y) = \mathbb{P}(\{\omega; \tilde{x}(\omega) \leq x, \tilde{y}(\omega) \leq y\}).$$

Clearly the df of  $x$  is

$$F(x, +\infty),$$

and that of  $y$  is

$$F(+\infty, y).$$

It is usually the case that when neither  $x$  nor  $y$  is discrete there is a function  $p(x, y)$  such that

$$F(x, y) = \int_{-\infty}^x \int_{-\infty}^y p_{\tilde{x}, \tilde{y}}(\xi, \eta) d\xi d\eta,$$

in which case  $p(x, y)$  is called a *joint (probability) density (function)* or *joint pdf*. When this is so, the joint distribution is said to be *continuous* (or more strictly to be absolutely continuous). We can find the density from the df by

$$p(x, y) = \partial^2 F(x, y) / \partial x \partial y.$$

Clearly,

$$p(x, y) \geq 0, \quad \iint p(x, y) dx dy = 1$$

and

$$p(x) = \int p(x, y) dy.$$

The last formula is the continuous analogue of

$$p(m) = \sum_n p(m, n)$$

in the discrete case.

By analogy with the discrete case, we define the *conditional density* of  $y$  given  $x$  (strictly of  $\tilde{y}$  given  $\tilde{x} = x$ ) as

$$p(y|x) = p(x, y)/p(x),$$

provided  $p(x) \neq 0$ . We can then define the *conditional distribution function* by

$$F(y|x) = \int_{-\infty}^y p(\eta|x) d\eta.$$

There are difficulties in the notion of conditioning on the event that  $\tilde{x} = x$  because this event has probability zero for every  $x$  in the continuous case, and it can help to regard the above distribution as the limit of the distribution which results from conditioning on the event that  $\tilde{x}$  is between  $x$  and  $x + \delta x$ , that is

$$\{\omega; x < \tilde{x}(\omega) \leq x + \delta x\}$$

as  $\delta x \rightarrow 0$ .

### 1.4.3 Bayes' Theorem for random variables

It is worth noting that conditioning the random variable  $y$  by the value of  $x$  does not change the *relative* sizes of the probabilities of those pairs  $(x, y)$  that can still occur. That is to say, the probability  $p(y|x)$  is proportional to  $p(x, y)$  and the constant of proportionality is just what is needed, so that the conditional probabilities integrate to unity. Thus,

$$p(y|x) \geq 0, \quad \int p(y|x) dy = 1.$$

Moreover,

$$p(y) = \int p(x, y) dx = \int p(x)p(y|x) dx.$$

It is clear that

$$p(y|x) = p(x, y)/p(x) = p(y)p(x|y)/p(x),$$

so that

$$p(y|x) \propto p(y)p(x|y).$$

This is, of course, a form of *Bayes' Theorem*, and is in fact the commonest way in which it occurs in this book. Note that it applies equally well if the variables  $x$  and  $y$  are continuous or if they are discrete. The constant of proportionality is

$$1/p(x) = 1 \Big/ \int p(y)p(x|y) dy$$

in the continuous case or

$$1/p(x) = 1 \Big/ \sum_y p(y)p(x|y)$$

in the discrete case.

#### 1.4.4 Example

A somewhat artificial example of the use of this formula in the continuous case is as follows. Suppose  $y$  is the time before the first occurrence of a radioactive decay which is measured by an instrument, but that, because there is a delay built into the mechanism, the decay is recorded as having taken place at a time  $x > y$ . We actually have the value of  $x$ , but would like to say what we can about the value of  $y$  on the basis of this knowledge. We might, for example, have

$$\begin{aligned} p(y) &= e^{-y} && (0 < y < \infty), \\ p(x|y) &= k e^{-k(x-y)} && (y < x < \infty). \end{aligned}$$

Then

$$\begin{aligned} p(y|x) &\propto p(y)p(x|y) \\ &\propto e^{(k-1)y} && (0 < y < x). \end{aligned}$$

Often we will find that it is enough to get a result up to a constant of proportionality, but if we need the constant, it is very easy to find it because we know that the integral (or the sum in the discrete case) must be one. Thus, in this case

$$p(y|x) = \frac{(k-1)e^{(k-1)y}}{e^{(k-1)x} - 1} \quad (0 < y < x).$$

#### 1.4.5 One discrete variable and one continuous variable

We also encounter cases where we have two random variables, one of which is continuous and one of which is discrete. All the aforementioned definitions and formulae extend in an obvious way to such a case provided we are careful, for example, to use integration for continuous variables but summation for discrete variables. In particular, the formulation

$$p(y|x) \propto p(y)p(x|y)$$

for Bayes' Theorem is valid in such a case.

It may help to consider an example (again a somewhat artificial one). Suppose  $k$  is the number of successes in  $n$  Bernoulli trials, so  $k \sim B(n, \pi)$ , but that the value of  $\pi$  is unknown, your beliefs about it being uniformly distributed over the interval  $[0, 1]$  of possible values. Then

$$\begin{aligned} p(k|\pi) &= \binom{n}{k} \pi^k (1-\pi)^{n-k} && (k = 0, 1, \dots, n), \\ p(\pi) &= 1 && (0 \leq \pi \leq 1), \end{aligned}$$

so that

$$p(\pi|k) \propto p(\pi)p(k|\pi) = \binom{n}{k} \pi^k (1-\pi)^{n-k}$$

$$\propto \pi^k (1-\pi)^{n-k}.$$

The constant can be found by integration if it is required. Alternatively, a glance at Appendix A will show that, given  $k, \pi$  has a beta distribution

$$\pi|k \sim \text{Be}(k+1, n-k+1)$$

and that the constant of proportionality is the reciprocal of the beta function  $B(k+1, n-k+1)$ . Thus, this beta distribution should represent your beliefs about  $\pi$  after you have observed  $k$  successes in  $n$  trials. This example has a special importance in that it is the one which Bayes himself discussed.

## 1.4.6 Independent random variables

The idea of independence extends from independence of events to independence of random variables. The basic idea is that  $y$  is independent of  $x$  if being told that  $x$  has any particular value does not affect your beliefs about the value of  $y$ . Because of complications involving events of probability zero, it is best to adopt the formal definition that  $x$  and  $y$  are independent if

$$p(x, y) = p(x)p(y)$$

for all values  $x$  and  $y$ . This definition works equally well in the discrete and the continuous cases (and indeed in the case where one random variable is continuous and the other is discrete). It trivially suffices that  $p(x, y)$  be a product of a function of  $x$  and a function of  $y$ .

All the above generalizes in a fairly obvious way to the case of more than two random variables, and the notions of pairwise and mutual independence go through from events to random variables easily enough. However, we will find that we do not often need such generalizations.

## 1.5 Means and variances

### 1.5.1 Expectations

Suppose that  $m$  is a discrete random variable and that the series

$$\sum mp(m)$$

is absolutely convergent, that is such that

$$\sum |m| p(m) < \infty.$$

Then the sum of the original series is called the *mean* or *expectation* of the random variable, and we denote it

$$\mathbb{E}m \text{ (strictly } \mathbb{E}\tilde{m}).$$

A motivation for this definition is as follows. In a large number  $N$  of trials, we would expect the value  $m$  to occur about  $p(m)N$  times, so that the sum total of the values that would occur in these  $N$  trials (counted according to their multiplicity) would be about

$$\sum mp(m)N,$$

so that the average value should be about

$$\sum mp(m)N/N = \mathbb{E}m.$$

Thus, we can think of expectation as being, at least in some circumstances, a form of very long term average. On the other hand, there are circumstances in which it is difficult to believe in the possibility of arbitrarily large numbers of trials, so this interpretation is not always available. It can also be thought of as giving the position of the ‘centre of gravity’ of the distribution imagined as a distribution of mass spread along the  $x$ -axis.

More generally, if  $g(m)$  is a function of the random variable and  $\sum g(m)p(m)$  is absolutely convergent, then its sum is the expectation of  $g(m)$ . Similarly, if  $h(m, n)$  is a function of two random variables  $m$  and  $n$  and the series  $\sum \sum h(m, n)p(m, n)$  is absolutely convergent, then its sum is the expectation of  $h(m, n)$ . These definitions are consistent in that if we consider  $g(m)$  and  $h(m, n)$  as random variables with densities of their own, then it is easily shown that we get these values for their expectations.

In the continuous case, we define the expectation of a random variable  $x$  by

$$\mathbb{E}x = \int xp(x) dx$$

provided that the integral is absolutely convergent, and more generally define the expectation of a function  $g(x)$  of  $x$  by

$$\mathbb{E}g(x) = \int g(x)p(x) dx$$

provided that the integral is absolutely convergent, and similarly for the expectation of a function  $h(x, y)$  of two random variables. Note that the formulae in the discrete and continuous cases are, as usual, identical except for the use of summation in the one case and integration in the other.

### 1.5.2 The expectation of a sum and of a product

If  $x$  and  $y$  are any two random variables, independent or not, and  $a$ ,  $b$  and  $c$  are constants, then in the continuous case

$$\begin{aligned}\mathbb{E}[ax + by + c] &= \iint (ax + by + c)p(x, y) dx dy \\ &= a \iint xp(x, y) dx dy + b \iint yp(x, y) dx dy + c \iint p(x, y) dx dy \\ &= a \int xp(x) dx + b \int yp(y) dy + c \\ &= a\mathbb{E}x + b\mathbb{E}y + c\end{aligned}$$

and similarly in the discrete case. Yet more generally, if  $g(x)$  is a function of  $x$  and  $h(y)$  a function of  $y$ , then

$$\mathbb{E}[ag(x) + bh(y) + c] = a\mathbb{E}g(x) + b\mathbb{E}h(y) + c.$$

We have already noted that the idea of independence is closely tied up with multiplication, and this is true when it comes to expectations as well. Thus, if  $x$  and  $y$  are independent, then

$$\begin{aligned}\mathbb{E}xy &= \iint xy p(x, y) dx dy \\ &= \iint xy p(x)p(y) dx dy \\ &= \left(\int xp(x) dx\right) \left(\int yp(y) dy\right) \\ &= (\mathbb{E}x)(\mathbb{E}y)\end{aligned}$$

and more generally if  $g(x)$  and  $h(y)$  are functions of independent random variables  $x$  and  $y$ , then

$$\mathbb{E}g(x)h(y) = (\mathbb{E}g(x))(\mathbb{E}h(y)).$$

### 1.5.3 Variance, precision and standard deviation

We often need a measure of how spread out a distribution is, and for most purposes the most useful such measure is the variance  $\mathcal{V}_x$  of  $x$ , defined by

$$\mathcal{V}_x = \mathbb{E}(x - \mathbb{E}x)^2.$$

Clearly if the distribution is very little spread out, then most values are close to one another and so close to their mean, so that  $(x - \mathbb{E}x)^2$  is small with high probability and hence  $\mathcal{V}_x$  is small. Conversely, if the distribution is well spread out then  $\mathcal{V}_x$  is large. It is sometimes useful to refer to the reciprocal of the variance, which is called the *precision*. Further, because the variance is essentially quadratic, we sometimes work in terms of its positive square root, the *standard deviation*, especially in numerical work. It is often useful that

$$\begin{aligned}\mathcal{V}x &= \mathbb{E}(x - \mathbb{E}x)^2 \\ &= \mathbb{E}(x^2 - 2(\mathbb{E}x)x + (\mathbb{E}x)^2) \\ &= \mathbb{E}x^2 - (\mathbb{E}x)^2.\end{aligned}$$

The notion of a variance is analogous to that of a moment of inertia in mechanics, and this formula corresponds to the *parallel axes theorem* in mechanics. This analogy seldom carries much weight nowadays, because so many of those studying statistics took it up with the purpose of avoiding mechanics.

In discrete cases, it is sometimes useful that

$$\mathcal{V}x = \mathbb{E}x(x - 1) + \mathbb{E}x - (\mathbb{E}x)^2.$$

### 1.5.4 Examples

As an example, suppose that  $k \sim B(n, \pi)$ . Then

$$\mathbb{E}k = \sum_{k=0}^n k \binom{n}{k} \pi^k (1-\pi)^{n-k}.$$

After a little manipulation, this can be expressed as

$$\mathbb{E}k = n\pi \sum_{j=0}^{n-1} \binom{n-1}{j} \pi^j (1-\pi)^{n-1-j}.$$

Because the sum is a sum of binomial  $B(n-1, \pi)$  probabilities, this expression reduces to  $n\pi$ , and so

$$\mathbb{E}k = n\pi.$$

Similarly,

$$\mathbb{E}k(k-1) = n(n-1)\pi^2$$

and so

$$\begin{aligned}\mathcal{V}k &= n(n-1)\pi^2 + n\pi - (n\pi)^2 \\ &= n\pi(1-\pi).\end{aligned}$$

For a second example, suppose  $x \sim N(\mu, \phi)$ . Then

$$\begin{aligned}\mathbb{E}x &= \int x (2\pi\phi)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(x-\mu)^2/\phi\right\} dx \\ &= \mu + \int (x-\mu) (2\pi\phi)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(x-\mu)^2/\phi\right\} dx.\end{aligned}$$

The integrand in the last expression is an odd function of  $x - \mu$  and so vanishes, so that

$$\mathbb{E}x = \mu.$$

Moreover,

$$\mathcal{V}x = \int (x-\mu)^2 (2\pi\phi)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(x-\mu)^2/\phi\right\} dx,$$

so that on writing  $z = (x - \mu)/\sqrt{\phi}$

$$\mathcal{V}_x = \phi \int z^2 (2\pi)^{-\frac{1}{2}} \exp(-\frac{1}{2}z^2) dz.$$

Integrating by parts (using  $z$  as the part to differentiate), we get

$$\begin{aligned}\mathcal{V}_x &= \phi \int (2\pi)^{-\frac{1}{2}} \exp(-\frac{1}{2}z^2) dz \\ &= \phi.\end{aligned}$$

### 1.5.5 Variance of a sum; covariance and correlation

Sometimes we need to find the *variance of a sum* of random variables. To do this, note that

$$\begin{aligned}\mathcal{V}(x + y) &= E\{x + y - E(x + y)\}^2 \\ &= E\{(x - Ex) + (y - Ey)\}^2 \\ &= E(x - Ex)^2 + E(y - Ey)^2 + 2E(x - Ex)(y - Ey) \\ &= \mathcal{V}_x + \mathcal{V}_y + 2\mathcal{C}(x, y),\end{aligned}$$

where the *covariance*  $\mathcal{C}(x, y)$  of  $x$  and  $y$  is defined by

$$\begin{aligned}\mathcal{C}(x, y) &= E(x - Ex)(y - Ey) \\ &= Exy - (Ex)(Ey).\end{aligned}$$

More generally,

$$\mathcal{V}(ax + by + c) = a^2\mathcal{V}_x + b^2\mathcal{V}_y + 2ab\mathcal{C}(x, y)$$

for any constants  $a, b$  and  $c$ . By considering this expression as a quadratic in  $a$  for fixed  $b$  or vice versa and noting that (because its value is always positive) this quadratic cannot have two unequal real roots, we see that

$$(\mathcal{C}(x, y))^2 \leq (\mathcal{V}_x)(\mathcal{V}_y).$$

We define the correlation coefficient  $\rho(x, y)$  between  $x$  and  $y$  by

$$\rho(x, y) = \frac{\mathcal{C}(x, y)}{\sqrt{(\mathcal{V}_x)(\mathcal{V}_y)}}.$$

It follows that

$$-1 \leq \rho(x, y) \leq 1$$

and indeed a little further thought shows that  $\rho(x, y) = 1$  if and only if

$$ax + by + c = 0$$

with probability 1 for some constants  $a, b$  and  $c$  with  $a$  and  $b$  having opposite signs, while  $\rho(x, y) = -1$  if and only if the same thing happens except that  $a$  and  $b$  have the same sign. If  $\rho(x, y) = 0$  we say that  $x$  and  $y$  are uncorrelated.

It is easily seen that if  $x$  and  $y$  are independent then

$$\mathcal{C}(x, y) = Exy - (Ex)(Ey) = 0$$

from which it follows that independent random variables are uncorrelated.

The converse is *not* in general true, but it can be shown that if  $x$  and  $y$  have a bivariate normal distribution (as described in Appendix A), then they are independent if and only if they are uncorrelated.

It should be noted that if  $x$  and  $y$  are uncorrelated, and in particular if they are independent

$$\mathcal{V}(x \pm y) = \mathcal{V}x + \mathcal{V}y$$

(observe that there is a plus sign on the right-hand side even if there is a minus sign on the left).

### 1.5.6 Approximations to the mean and variance of a function of a random variable

Very occasionally, it will be useful to have an approximation to the mean and variance of a function of a random variable. Suppose that

$$z = g(x).$$

Then if  $g$  is a reasonably smooth function and  $x$  is not too far from its expectation, Taylor's theorem implies that

$$z \cong g(\mathbf{E}x) + (x - \mathbf{E}x)g'(\mathbf{E}x).$$

It, therefore, seems reasonable that a fair approximation to the expectation of  $z$  is given by

$$\mathbf{E}z = g(\mathbf{E}x)$$

and if this is so, then a reasonable approximation to  $\mathcal{V}z$  may well be given by

$$\mathcal{V}z = \mathcal{V}x\{g'(\mathbf{E}x)\}^2.$$

As an example, suppose that

$$x \sim \text{B}(n, \pi)$$

and that  $z=g(x)$ , where

$$g(x) = \sin^{-1} \sqrt{(x/n)},$$

so that

$$g'(x) = \frac{1}{2}n^{-1}[(x/n)(1 - (x/n))]^{-\frac{1}{2}},$$

and thus  $g'(\mathbf{E}x) = g'(n\pi) = \frac{1}{2}n^{-1}[\pi(1 - \pi)]^{-\frac{1}{2}}$ . The aforementioned argument then implies that

$$\mathbf{E}z \cong \sin^{-1} \sqrt{\pi}, \quad \mathcal{V}z \cong 1/4n.$$

The interesting thing about this transformation, which has a long history [see Eisenhart *et al.* (1947, Chapter 16) and Fisher (1954)], is that, to the extent to which the approximation is valid, the variance of  $z$  does not depend on the parameter  $\pi$ . It is accordingly known as a *variance-stabilizing transformation*.

We will return to this transformation in Section 3.2 on the ‘Reference Prior for the Binomial Distribution’.

### 1.5.7 Conditional expectations and variances

If the reader wishes, the following may be omitted on a first reading and then returned to as needed.

We define the *conditional expectation* of  $y$  given  $x$  by

$$\mathbb{E}(y|x) = \int y p(y|x) dy$$

in the continuous case and by the corresponding sum in the discrete case. If we wish to be pedantic, it can occasionally be useful to indicate what we are averaging over by writing

$$\mathbb{E}_{\tilde{y}|\tilde{x}}(\tilde{y}|x)$$

just as we can write  $p_{\tilde{y}|\tilde{x}}(y|x)$ , but this is rarely necessary (though it can slightly clarify a proof on occasion). More generally, the conditional expectation of a function  $g(y)$  of  $y$  given  $x$  is

$$\mathbb{E}(g(y)|x) = \int g(y) p(y|x) dy.$$

We can also define a conditional variance as

$$\begin{aligned}\mathcal{V}(y|x) &= \mathbb{E}[(y - \mathbb{E}(y|x))^2 | x] \\ &= \mathbb{E}(y^2|x) - \{\mathbb{E}(y|x)\}^2.\end{aligned}$$

Despite some notational complexity, this is easy enough to find since after all a conditional distribution is just a particular case of a probability distribution. If we are really pedantic, then  $\mathbb{E}(\tilde{y}|x)$  is a real number which is a function of the real number  $x$ , while  $\mathbb{E}(\tilde{y}|\tilde{x})$  is a random variable which is a function of the random variable  $\tilde{x}$ , which takes the value  $\mathbb{E}(\tilde{y}|x)$  when  $\tilde{x}$  takes the value  $x$ . However, the distinction, which is hard to grasp in the first place, is usually unimportant.

We may note that the formula

$$p(y) = \int p(y|x)p(x)dx$$

could be written as

$$p(y) = \mathbb{E}p(y|\tilde{x})$$

but we must be careful that it is an expectation over values of  $\tilde{x}$  (i.e.  $\mathbb{E}_{\tilde{x}}$ ) that occurs here.

Very occasionally we make use of results like

$$\begin{aligned} E\tilde{y} &= E_{\tilde{x}}\{E_{\tilde{y}|\tilde{x}}(\tilde{y}|\tilde{x})\}, \\ \mathcal{V}\tilde{y} &= E_{\tilde{x}}\mathcal{V}_{\tilde{y}|\tilde{x}}(\tilde{y}|\tilde{x}) + \mathcal{V}_{\tilde{x}}E_{\tilde{y}|\tilde{x}}(\tilde{y}|\tilde{x}). \end{aligned}$$

The proofs are possibly more confusing than helpful. They run as follows:

$$\begin{aligned} E\{E(\tilde{y}|\tilde{x})\} &= \int E(\tilde{y}|x)p(x)dx \\ &= \int \left( \int yp(y|x) dy \right) p(x)dx \\ &= \iint yp(x,y) dy dx \\ &= \int yp(y) dy \\ &= E\tilde{y}. \end{aligned}$$

Similarly, we get the generalization

$$E\{E(g(\tilde{y})|\tilde{x})\} = Eg(\tilde{y})$$

and in particular

$$E\{E(\tilde{y}^2|\tilde{x})\} = E\tilde{y}^2,$$

hence

$$\begin{aligned} E\mathcal{V}(\tilde{y}|\tilde{x}) &= E\{E(\tilde{y}^2|\tilde{x})\} - E\{E(\tilde{y}|\tilde{x})\}^2 \\ &= E\tilde{y}^2 - E\{E(\tilde{y}|\tilde{x})\}^2 \end{aligned}$$

while

$$\begin{aligned} \mathcal{V}E(\tilde{y}|\tilde{x}) &= E\{E(\tilde{y}|\tilde{x})\}^2 - [E\{E(\tilde{y}|x)\}]^2 \\ &= E\{E(\tilde{y}|\tilde{x})\}^2 - E\tilde{y}^2 \end{aligned}$$

from which it follows that

$$E\mathcal{V}(\tilde{y}|\tilde{x}) + \mathcal{V}E(\tilde{y}|\tilde{x}) = E\tilde{y}^2 - (E\tilde{y})^2 = \mathcal{V}\tilde{y}.$$

## 1.5.8 Medians and modes

The mean is not the only measure of the centre of a distribution. We also need to consider the *median* from time to time, which is defined as any value  $x_0$  such that

$$P(\tilde{x} \leq x_0) \geq \frac{1}{2} \quad \text{and} \quad P(\tilde{x} \geq x_0) \geq \frac{1}{2}.$$

In the case of most continuous random variables there is a unique median such that

$$P(\tilde{x} \geq x_0) = P(\tilde{x} \leq x_0) = \frac{1}{2}.$$

We occasionally refer also to the *mode*, defined as that value at which the pdf is a maximum. One important use we shall have for the mode will be in methods for finding the median based on the approximation

$$\text{mean} - \text{mode} = 3(\text{mean} - \text{median})$$

or equivalently

$\text{median} = (2 \text{ mean} + \text{mode})/3$   
(see the preliminary remarks in Appendix A).

## 1.6 Exercises on Chapter 1

1. A card game is played with 52 cards divided equally between four players, North, South, East and West, all arrangements being equally likely. Thirteen of the cards are referred to as trumps. If you know that North and South have ten trumps between them, what is the probability that all three remaining trumps are in the same hand? If it is known that the king of trumps is included among the other three, what is the probability that one player has the king and the other the remaining two trumps?
2.
  - a. Under what circumstances is an event  $A$  independent of itself?
  - b. By considering events concerned with independent tosses of a red die and a blue die, or otherwise, give examples of events  $A$ ,  $B$  and  $C$  which are not independent, but nevertheless are such that every pair of them is independent.
  - c. By considering events concerned with three independent tosses of a coin and supposing that  $A$  and  $B$  both represent tossing a head on the first trial, give examples of events  $A$ ,  $B$  and  $C$  which are such that  $P(ABC) = P(A)P(B)P(C)$  although no pair of them is independent.
3. Whether certain mice are black or brown depends on a pair of genes, each of which is either  $B$  or  $b$ . If both members of the pair are alike, the mouse is said to be homozygous, and if they are different it is said to be heterozygous  $bb$ . The mouse is brown only if it is homozygous  $bb$ . The offspring of a pair of mice have two such genes, one from each parent, and if the parent is heterozygous, the inherited gene is equally likely to be  $B$  or  $b$ . Suppose that a black mouse results from a mating between two heterozygotes.
  - a. What are the probabilities that this mouse is homozygous and that it is heterozygous?

Now suppose that this mouse is mated with a brown mouse, resulting in seven offspring, all of which turn out to be black.

  - b. Use Bayes' Theorem to find the probability that the black mouse was homozygous  $BB$ .
  - c. Recalculate the same probability by regarding the seven offspring as seven observations made sequentially, treating the posterior after each observation as the prior for the next (cf. Fisher, 1959, Section II.2).

**4.** The example on Bayes' Theorem in Section 1.2 concerning the biology of twins was based on the assumption that births of boys and girls occur equally frequently, and yet it has been known for a very long time that fewer girls are born than boys (cf. Arbuthnot, 1710). Suppose that the probability of a girl is  $p$ , so that

$$\begin{aligned}\mathbb{P}(GG|M) &= p, \quad \mathbb{P}(BB|M) = 1 - p, \quad \mathbb{P}(GB|M) = 0, \\ \mathbb{P}(GG|D) &= p^2, \quad \mathbb{P}(BB|D) = (1 - p)^2, \quad \mathbb{P}(GB|D) = 2p(1 - p).\end{aligned}$$

Find the proportion of monozygotic twins in the whole population of twins in terms of  $p$  and the sex distribution among all twins.

**5.** Suppose a red and a blue die are tossed. Let  $x$  be the sum of the number showing on the red die and twice the number showing on the blue die. Find the density function and the distribution function of  $x$ .

**6.** Suppose that  $k \sim B(n, \pi)$  where  $n$  is large and  $\pi$  is small but  $n\pi = \lambda$  has an intermediate value. Use the exponential limit  $(1 + x/n)^n \rightarrow e^x$  to show that  $\mathbb{P}(k = 0) \cong e^{-\lambda}$  and  $\mathbb{P}(k = 1) \cong \lambda e^{-\lambda}$ . Extend this result to show that  $k$  is such that

$$p(k) \cong \frac{\lambda^k}{k!} \exp(-\lambda)$$

that is,  $k$  is approximately distributed as a Poisson variable of mean  $\lambda$  (cf. Appendix A).

**7.** Suppose that  $\lambda$  and  $\mu$  have independent Poisson distributions of means and respectively (see question 6) and that  $k = m + n$ .

- a. Show that  $\mathbb{P}(k = 0) = e^{-(\lambda + \mu)}$  and  $\mathbb{P}(k = 1) = (\lambda + \mu)e^{-(\lambda + \mu)}$ .
- b. Generalize by showing that  $k$  has a Poisson distribution of mean  $\lambda + \mu$ .
- c. Show that conditional on  $k$ , the distribution of  $m$  is binomial of index  $k$  and parameter  $\lambda/(\lambda + \mu)$ .

**8.** Modify the formula for the density of a one-to-one function  $g(x)$  of a random variable  $x$  to find an expression for the density of  $x^2$  in terms of that of  $x$ , in both the continuous and discrete case. Hence, show that the square of a standard normal density has a chi-squared density on one degree of freedom as defined in Appendix A.

**9.** Suppose that  $x_1, x_2, \dots, x_n$  are independently and all have the same continuous distribution, with density  $f(x)$  and distribution function  $F(x)$ . Find the distribution functions of

$$M = \max\{x_1, x_2, \dots, x_n\} \quad \text{and} \quad m = \min\{x_1, x_2, \dots, x_n\}$$

in terms of  $F(x)$ , and so find expressions for the density functions of  $M$  and

*m.*

**10.** Suppose that  $u$  and  $v$  are independently uniformly distributed on the interval  $[0, 1]$ , so that they divide the interval into three sub-intervals. Find the joint density function of the lengths of the first two sub-intervals.

**11.** Show that two continuous random variables  $x$  and  $y$  are independent (i.e.  $p(x, y) = p(x)p(y)$  for all  $x$  and  $y$ ) if and only if their joint distribution function  $F(x, y)$  satisfies  $F(x, y) = F(x)F(y)$  for all  $x$  and  $y$ . Prove that the same thing is true for discrete random variables. [This is an example of a result which is easier to prove in the continuous case.]

**12.** Suppose that the random variable  $x$  has a negative binomial distribution  $NB(n, \pi)$  of index  $n$  and parameter  $\pi$ , so that

$$p(x) = \binom{n+x-1}{x} \pi^n (1-\pi)^x$$

Find the mean and variance of  $x$  and check that your answer agrees with that given in Appendix A.

**13.** A random variable  $X$  is said to have a chi-squared distribution on  $\nu$  degrees of freedom if it has the same distribution as

$$Z_1^2 + Z_2^2 + \cdots + Z_\nu^2$$

where  $Z_1, Z_2, \dots, Z_\nu$  are independent standard normal variates. Use the facts that  $EZ_i = 0$ ,  $EZ_i^2 = 1$  and  $EZ_i^4 = 3$  to find the mean and variance of  $X$ . Confirm these values using the probability density of  $X$ , which is

$$p(X) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)} X^{\nu/2-1} \exp(-\frac{1}{2}X) \quad (0 < X < \infty)$$

(see Appendix A).

**14.** The *skewness* of a random variable  $x$  is defined as  $\gamma_1 = \mu_3 / (\mu_2)^{\frac{3}{2}}$  where

$$\mu_n = E(x - Ex)^n$$

(but note that some authors work in terms of  $\beta_1 = \gamma_1^2$ ). Find the skewness of a random variable  $X$  with a binomial distribution  $B(n, \pi)$  of index  $n$  and parameter  $\pi$ .

**15.** Suppose that a continuous random variable  $X$  has mean  $\mu$  and variance  $\phi$ . By writing

$$\phi = \int (x - \mu)^2 p(x) dx \geq \int_{\{x; |x - \mu| \geq c\}} (x - \mu)^2 p(x) dx$$

and using a lower bound for the integrand in the latter integral, prove that

$$P(|x - \mu| \geq c) \leq \frac{\phi}{c^2}.$$

Show that the result also holds for discrete random variables. [This result is known as Čebyšev's Inequality (the name is spelt in many other ways, including Chebyshev and Tchebycheff).]

**16.** Suppose that  $x$  and  $y$  are such that

$$\begin{aligned} P(x = 0, y = 1) &= P(x = 0, y = -1) = P(x = 1, y = 0) \\ &= P(x = -1, y = 0) = \frac{1}{4}. \end{aligned}$$

Show that  $x$  and  $y$  are uncorrelated but that they are *not* independent.

**17.** Let  $x$  and  $y$  have a bivariate normal distribution and suppose that  $x$  and  $y$  both have mean 0 and variance 1, so that their marginal distributions are standard normal and their joint density is

$$p(x, y) = \left\{ 2\pi\sqrt{(1-\rho^2)} \right\}^{-1} \exp \left\{ -\frac{1}{2}(x^2 - 2\rho xy + y^2)/(1-\rho^2) \right\}.$$

Show that if the correlation coefficient between  $x$  and  $y$  is  $\rho$ , then that between  $x^2$  and  $y^2$  is  $\rho^2$ .

**18.** Suppose that  $x$  has a Poisson distribution (see question 6)  $P(\lambda)$  of mean  $\lambda$  and that, for given  $x$ ,  $y$ , has a binomial distribution  $B(x, \pi)$  of index  $x$  and parameter  $\pi$ .

a. Show that the unconditional distribution of  $y$  is Poisson of mean

$$\lambda\pi = E_{\tilde{x}} E_{\tilde{y}|\tilde{x}}(\tilde{y}|\tilde{x}).$$

b. Verify that the formula

$$\mathcal{V} \tilde{y} = E_{\tilde{x}} \mathcal{V}_{\tilde{y}|\tilde{x}}(\tilde{y}|\tilde{x}) + \mathcal{V}_{\tilde{x}} E_{\tilde{y}|\tilde{x}}(\tilde{y}|\tilde{x})$$

derived in Section 1.5 holds in this case.

**19.** Define

$$I = \int_0^\infty \exp(-\frac{1}{2}z^2) dz$$

and show (by setting  $z=xy$  and then substituting  $z$  for  $y$ ) that

$$I = \int_0^\infty \exp(-\frac{1}{2}(xy)^2) y dx = \int_0^\infty \exp(-\frac{1}{2}(zx)^2) z dx.$$

Deduce that

$$I^2 = \int_0^\infty \int_0^\infty \exp\{-\frac{1}{2}(x^2 + 1)z^2\} z dz dx.$$

By substituting  $(1+x^2)z^2=2t$ , so that  $z dz = dt/(1+x^2)$  show that  $I = \sqrt{\pi/2}$ , so that the density of the standard normal distribution as defined in Section 1.3 does integrate to unity and so is indeed a density. (This method is due to Laplace, 1812, Section 24.)

# 2

## Bayesian inference for the normal distribution

### 2.1 Nature of Bayesian inference

#### 2.1.1 Preliminary remarks

In this section, a general framework for Bayesian statistical inference will be provided. In broad outline, we take prior beliefs about various possible hypotheses and then modify these prior beliefs in the light of relevant data which we have collected in order to arrive at posterior beliefs. (The reader may prefer to return to this section after reading Section 2.2, which deals with one of the simplest special cases of Bayesian inference.) 2.1.2 Post is prior times likelihood Almost all of the situations we will think of in this book fit into the following pattern. Suppose that you are interested in the values of  $k$  unknown quantities  $\theta = (\theta_1, \theta_2, \dots, \theta_k)$

(where  $k$  can be one or more than one) and that you have some a priori beliefs about their values which you can express in terms of the pdf  $p(\theta)$ .

Now suppose that you then obtain some data relevant to their values. More precisely, suppose that we have  $n$  observations  $X = (X_1, X_2, \dots, X_n)$  which have a probability distribution that depends on these  $k$  unknown quantities as parameters, so that the pdf (continuous or discrete) of the vector  $X$  depends on the vector  $\theta$  in a known way. Usually the components of  $\theta$  and  $X$  will be integers or real numbers, so that the components of  $X$  are random variables, and so the dependence of  $X$  on  $\theta$  can be expressed in terms of a pdf  $p(X|\theta)$ .

You then want to find a way of expressing your beliefs about  $\theta$  taking into account both your prior beliefs and the data. Of course, it is possible that your prior beliefs about  $\theta$  may differ from mine, but very often we will agree on the way in which the data are related to  $\theta$  [i.e. on the form of  $p(X|\theta)$ ]. If this is so, we will differ in our posterior beliefs (i.e. in our beliefs after we have obtained

the data), but it will turn out that if we can collect enough data, then our posterior beliefs will usually become very close.

The basic tool we need is Bayes' Theorem for random variables (generalized to deal with random vectors). From this theorem, we know that  $p(\theta|X) \propto p(\theta) p(X|\theta)$ .

Now we know that  $p(X|\theta)$  considered as a function of  $X$  for fixed  $\theta$  is a density, but we will find that we often want to think of it as a function of  $\theta$  for fixed  $X$ . When we think of it in that way it does not have quite the same properties – for example, there is no reason why it should sum (or integrate) to unity. Thus, in the extreme case where  $p(X|\theta)$  turns out not to depend on  $\theta$ , then it is easily seen that it can quite well sum (or integrate) to  $\infty$ . When we are thinking of  $p(X|\theta)$  as a function of  $\theta$  we call it the *likelihood* function. We sometimes write  $l(\theta|X) = p(X|\theta)$ .

Just as we sometimes write  $p_{X|\theta}(X|\theta)$  to avoid ambiguity, if we really need to avoid ambiguity we write  $l_{\theta|X}(\theta|X)$

but this will not usually be necessary. Sometimes it is more natural to consider the *log-likelihood* function  $L(\theta|X) = \log l(\theta|X)$ .

With this definition and the definition of  $p(\theta)$  as the prior pdf for  $\theta$  and of  $p(\theta|X)$  as the posterior pdf for  $\theta$  given  $X$ , we may think of Bayes' Theorem in the more memorable form Posterior  $\propto$  Prior  $\times$  Likelihood.

This relationship summarizes the way in which we should modify our beliefs in order to take into account the data we have available.

### 2.1.3 Likelihood can be multiplied by any constant

Note that because of the way we write Bayes' Theorem with a proportionality sign, it does not alter the result if we multiply  $l(\theta|X)$  by any constant or indeed more generally by anything which is a function of  $X$  alone. Accordingly, we can regard the definition of the likelihood as being *any constant multiple* of  $p(X|\theta)$  rather than necessarily equalling  $p(X|\theta)$  (and similarly the log-likelihood is

undetermined up to an additive constant). Sometimes the integral  $\int l(\theta|X) d\theta$  (interpreted as a multiple integral  $\iint \dots \int d\theta_1 d\theta_2 \dots d\theta_k$  if  $k > 1$  and interpreted as a summation or multiple summation in the discrete case), taken over the admissible range of  $\theta$ , is finite, although we have already noted that this is not always the case. When it is, it is occasionally convenient to refer to the quantity  $\frac{l(\theta|X)}{\int l(\theta|X) d\theta}$ .

We shall call this the *standardized likelihood*, that is, the likelihood scaled so that it integrates to unity and can thus be thought of as a density.

### 2.1.4 Sequential use of Bayes' Theorem

It should also be noted that the method can be applied sequentially. Thus, if you have an initial sample of observations  $X$ , you have  $p(\theta|X) \propto p(\theta)l(\theta|X)$ .

Now suppose that you have a second set of observations  $Y$  distributed independently of the first sample. Then  $p(\theta|X, Y) \propto p(\theta)l(\theta|X, Y)$ .

But independence implies

$$p(X, Y|\theta) = p(X|\theta)p(Y|\theta)$$

from which it is obvious that

$$l(\theta|X, Y) \propto l(\theta|X)l(\theta|Y)$$

and hence

$$\begin{aligned} p(\theta|X, Y) &\propto p(\theta)l(\theta|X)l(\theta|Y) \\ &\propto p(\theta|X)l(\theta|Y). \end{aligned}$$

So we can find your posterior for  $\theta$  given  $X$  and  $Y$  by treating your posterior given  $X$  as the prior for the observation  $Y$ . This formula will work *irrespective of the temporal order* in which  $X$  and  $Y$  are observed, and this fact is one of the advantages of the Bayesian approach.

### 2.1.5 The predictive distribution

Occasionally (e.g. when we come to consider Bayesian decision theory and empirical Bayes methods), we need to consider the marginal distribution

$$p(X) = \int p(X|\theta)p(\theta) d\theta$$

which is called the *predictive distribution* of  $X$ , since it represents our current predictions of the value of  $X$  taking into account both the uncertainty about the value of  $\theta$  and the residual uncertainty about  $X$  when  $\theta$  is known.

One valuable use of the predictive distribution is in checking your underlying assumptions. If, for example,  $p(X)$  turns out to be small (in some sense) for the observed value of  $X$ , it might suggest that the form of the likelihood you have adopted was suspect. Some people have suggested that another thing you might re-examine in such a case is the prior distribution you have adopted, although there are logical difficulties about this if  $p(\theta)$  just represents your prior beliefs. It might, however, be the case that seeing an observation the possibility of which you had rather neglected causes you to think more fully and thus bring out

beliefs which were previously lurking below the surface.

There are actually two cases in which we might wish to consider a distribution for  $X$  taking into account both the uncertainty about the value of  $\theta$  and the residual uncertainty about  $X$  when  $\theta$  is known, depending on whether the distribution for  $\theta$  under consideration does or does not take into account some current observations, and some authors reserve the term ‘predictive distribution’ for the former case and use the term *preposterior distribution* in cases where we do not yet have any observations to take into account. In this book, the term ‘predictive distribution’ is used in both cases.

### 2.1.6 A warning

The theory described earlier relies on the possibility of specifying the likelihood as a function, or equivalently on being able to specify the density  $p(X|\theta)$  of the observations  $X$  save for the fact that the  $k$  parameters  $\theta_1, \theta_2, \dots, \theta_k$  are unknown. It should be borne in mind that these assumptions about the form of the likelihood may be unjustified, and a blind following of the procedure described earlier can never lead to their being challenged (although the point made earlier in connection with the predictive distribution can be of help). It is all too easy to adopt a model because of its convenience and to neglect the absence of evidence for it.

## 2.2 Normal prior and likelihood

### 2.2.1 Posterior from a normal prior and likelihood

We say that  $x$  is normal of mean  $\theta$  and variance  $\phi$  and write  $x \sim N(\theta, \phi)$  when

$$p(x) = (2\pi\phi)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(x - \theta)^2/\phi\right\}.$$

Suppose that you have an unknown parameter  $\theta$  for which your prior beliefs can be expressed in terms of a normal distribution, so that  $\theta \sim N(\theta_0, \phi_0)$  and suppose also that you have an observation  $x$  which is normally distributed with mean equal to the parameter of interest, that is  $x \sim N(\theta, \phi)$  where  $\theta_0$ ,  $\phi_0$  and  $\phi$  are known. As mentioned in Section 1.3, there are often grounds for suspecting that an observation might be normally distributed, usually related to the Central Limit Theorem, so this assumption is not

$$p(\theta) = (2\pi\phi_0)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\theta - \theta_0)^2/\phi_0\right\}$$

implausible. If these assumptions are valid  $p(x|\theta) = (2\pi\phi)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(x - \theta)^2/\phi\right\}$ , and hence

$$\begin{aligned} p(\theta|x) &\propto p(\theta)p(x|\theta) \\ &= (2\pi\phi_0)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\theta - \theta_0)^2/\phi_0\right\} \times (2\pi\phi)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(x - \theta)^2/\phi\right\} \\ &\propto \exp\left\{-\frac{1}{2}\theta^2(\phi_0^{-1} + \phi^{-1}) + \theta(\theta_0/\phi_0 + x/\phi)\right\} \end{aligned}$$

regarding  $p(\theta|x)$  as a function of  $\theta$ .

It is now convenient to write

$$\phi_1 = \frac{1}{\phi_0^{-1} + \phi^{-1}}$$

$$\theta_1 = \phi_1(\theta_0/\phi_0 + x/\phi),$$

so that

$$\phi_0^{-1} + \phi^{-1} = \phi_1^{-1}$$

$$\theta_0/\phi_0 + x/\phi = \theta_1/\phi_1,$$

and hence,

$$p(\theta|x) \propto \exp\left\{-\frac{1}{2}\theta^2/\phi_1 + \theta\theta_1/\phi_1\right\}.$$

Adding into the exponent

$$-\frac{1}{2}\theta_1^2/\phi_1$$

which is constant as far as  $\theta$  is concerned, we see that

$$p(\theta|x) \propto \exp\left\{-\frac{1}{2}(\theta - \theta_1)^2/\phi_1\right\}$$

from which it follows that as a density must integrate to unity

$$p(\theta|x) = (2\pi\phi_1)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\theta - \theta_1)^2/\phi_1\right\}$$

that is that the posterior density is

$$\theta|x \sim N(\theta_1, \phi_1).$$

In terms of the precision, which we recall can be defined as the reciprocal of the variance, the relationship  $\phi_0^{-1} + \phi^{-1} = \phi_1^{-1}$  can be remembered as Posterior precision = Prior precision + Datum precision.

(It should be noted that this relationship has been derived assuming a normal prior and a normal likelihood.)

The relation for the posterior mean,  $\theta_1$ , is only slightly more complicated. We have

$$\theta_1 = \theta_0 \frac{\phi_0^{-1}}{\phi_0^{-1} + \phi^{-1}} + x \frac{\phi^{-1}}{\phi_0^{-1} + \phi^{-1}}$$

which can be remembered as

Posterior mean = Weighted mean of prior mean and datum value,  
the weights being proportional to their  
respective precisions.

## 2.2.2 Example

According to Kennett and Ross (1983), the first apparently reliable datings for the age of Ennerdale granophyre were obtained from the K/Ar method (which depends on observing the relative proportions of potassium 40 and argon 40 in the rock) in the 1960s and early 1970s, and these resulted in an estimate of  $370 \pm 20$  million years. Later in the 1970s, measurements based on the Rb/Sr method (depending on the relative proportions of rubidium 87 and strontium 87) gave an age of  $421 \pm 8$  million years. It appears that the errors marked are meant to be standard deviations, and it seems plausible that the errors are normally distributed. If then a scientist  $S$  had the K/Ar measurements available in the early 1970s, it could be said that (before the Rb/Sr measurements came in),  $S$ 's prior beliefs about the age of these rocks were represented by  $\theta \sim N(370, 20^2)$ .

We could then suppose that the investigations using the Rb/Sr method result in a measurement

$$x \sim N(\theta, 8^2).$$

We shall suppose for simplicity that the precisions of these measurements are known to be exactly those quoted, although this is not quite true (methods which take more of the uncertainty into account will be discussed later in the book). If we then use the above method, then, noting that the observation  $x$  turned out to be 421, we see that  $S$ 's posterior beliefs about  $\theta$  should be represented by  $\theta | x \sim N(\theta_1, \phi_1)$ ,

where (retaining only one significant figure)

$$\phi_1 = (20^{-2} + 8^{-2})^{-1} = 55 = 7^2,$$

$$\theta_1 = 55(370/20^2 + 421/8^2) = 413.$$

Thus the posterior for the age of the rocks is

$$\theta | x \sim N(413, 7^2)$$

that is  $413 \pm 7$  million years.

Of course, all this assumes that the K/Ar measurements were available. If the Rb/Sr measurements were considered by another scientist  $\hat{S}$  who had no knowledge of these, but had a vague idea (in the light of knowledge of similar rocks) that their age was likely to be  $400 \pm 50$  million years, that is  $\theta \sim N(400, 50^2)$  then  $\hat{S}$  would have a posterior variance

$$\phi_1 = (50^{-2} + 8^{-2})^{-1} = 62 = 8^2$$

and a posterior mean of

$$\theta_1 = 62(400/50^2 + 421/8^2) = 418,$$

so that  $\hat{S}$ 's posterior distribution is

$$\theta | x \sim N(418, 8^2)$$

that is  $418 \pm 8$  million years. We note that this calculation has been carried out assuming that the prior information available is rather vague, and that this is reflected in the fact that the posterior is almost entirely determined by the data.

The situation can be summarized as follows:

Prior distribution	Likelihood from data	Posterior distribution
$S \sim N(370, 20^2)$		$N(413, 7^2)$
	$N(421, 8^2)$	
$\hat{S} \sim N(400, 50^2)$		$N(418, 8^2)$

We note that in numerical work, it is usually more meaningful to think in terms of the standard deviation  $\sqrt{\phi}$ , whereas in theoretical work it is usually easier to work in terms of the variance  $\phi$  itself.

We see that after this single observation the ideas of  $S$  and  $\hat{S}$  about  $\theta$  as represented by their posterior distributions are much closer than before, although they still differ considerably.

### 2.2.3 Predictive distribution

In the case discussed in this section, it is easy to find the predictive distribution, since

$$\tilde{x} = (\bar{x} - \theta) + \theta$$

and, independently of one another,

$$\bar{x} - \theta \sim N(0, \phi),$$

$$\theta \sim N(\theta_0, \phi_0)$$

from which it follows that

$$\tilde{x} \sim N(\theta_0, \phi + \phi_0)$$

using the standard fact that the sum of independent normal variates has a normal distribution. (The fact that the mean is the sum of the means and the variance the sum of the variances is of course true more generally as proved in Section 1.4 on ‘Several Random Variables’.)

2.2.4 The nature of the assumptions made  
Although this example is very simple, it does exhibit the main features of Bayesian inference as outlined in the previous section. We have assumed that the distribution of the observation  $x$  is *known to be normal* but that there is *an*

*unknown parameter*  $\theta$ , in this case the mean of the normal distribution. The assumption that the variance is known is unlikely to be fully justified in a practical example, but it may provide a reasonable approximation. You should, however, beware that it is all too easy to concentrate on the parameters of a well-known family, in this case the normal family, and to forget that the assumption that the density is in that family for *any* values of the parameters may not be valid. The fact that the normal distribution is easy to handle, as witnessed by the way that normal prior and normal likelihood combine to give normal posterior, is a good reason for looking for a normal model when it does provide a fair approximation, but there can easily be cases where it does not.

## 2.3 Several normal observations with a normal prior

### 2.3.1 Posterior distribution

We can generalize the situation in the previous section by supposing that a prior  $\theta \sim N(\theta_0, \phi_0)$

but that instead of having just one observation we have  $n$  independent observations  $x = (x_1, x_2, \dots, x_n)$  such that  $x_i \sim N(\theta, \phi)$ .

We sometimes refer to  $X$  as an  $n$ -sample from  $N(\theta, \phi)$ . Then

$$\begin{aligned} p(\theta|x) &\propto p(\theta)p(x|\theta) = p(\theta)p(x_1|\theta)p(x_2|\theta)\dots p(x_n|\theta) \\ &= (2\pi\phi_0)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\theta - \theta_0)^2/\phi_0\right\} \\ &\quad \times (2\pi\phi)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(x_1 - \theta)^2/\phi\right\} \\ &\quad \times (2\pi\phi)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(x_2 - \theta)^2/\phi\right\} \\ &\quad \times \dots \times (2\pi\phi)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(x_n - \theta)^2/\phi\right\} \\ &\propto \exp\left\{-\frac{1}{2}\theta^2(1/\phi_0 + n/\phi) + \theta\left(\theta_0/\phi_0 + \sum x_i/\phi\right)\right\} \end{aligned}$$

Proceeding just as we did in Section 2.3 when we had only one observation, we see that the posterior distribution is

$$\theta \sim N(\theta_1, \phi_1),$$

where

$$\begin{aligned} \phi_1 &= (1/\phi_0 + n/\phi)^{-1} \\ \theta_1 &= \phi_1 \left( \theta_0/\phi_0 + \sum x_i/\phi \right). \end{aligned}$$

We could alternatively write these formulae as

$$\phi_1 = \{\phi_0^{-1} + (\phi/n)^{-1}\}^{-1}$$

$$\theta_1 = \phi_1\{\theta_0/\phi_0 + \bar{x}/(\phi/n)\}$$

which shows that, assuming a normal prior and likelihood, the result is just the same as the posterior distribution obtained from the single observation of the mean  $\bar{x}$ , since we know that  $\bar{x} \sim N(\theta, \phi/n)$

and the above formulae are the ones we had before with  $\phi$  replaced by  $\phi/n$  and  $x$  by  $\bar{x}$ . (Note that the use of a bar over the  $x$  here to denote a mean is unrelated to the use of a tilde over  $x$  to denote a random variable).

We would of course obtain the same result by proceeding sequentially from  $p(\theta)$  to  $p(\theta|x_1)$  and then treating  $p(\theta|x_1)$  as prior and  $x_2$  as data to obtain  $p(\theta|x_1, x_2)$  and so on. This is in accordance with the general result mentioned in Section 2.1 on ‘Nature of Bayesian Inference’.

### 2.3.2 Example

We now consider a numerical example. The basic assumption in this section is that the variance is *known*, even though in most practical cases, it has to be estimated. There are a few circumstances in which the variance could be known, for example when we are using a measuring instrument which has been used so often that its measurement errors are well known, but there are not many. Later in this book, we will discover two things which mitigate this assumption – firstly, that the numerical results are not much different when we do take into account the uncertainty about the variance, and, secondly, that the larger the sample size is, the less difference it makes.

The data we will consider are quoted by Whittaker and Robinson (1940, Section 97). They consider chest measurements of 10 000 men. Now, based on memories of my experience as an assistant in a gentlemen’s outfitters in my university vacations, I would suggest a prior  $N(38, 9)$ .

Of course, it is open to question whether these men form a random sample from the whole population, but unless I am given information to the contrary I would stick to the prior I have just quoted, except that I might be inclined to increase the variance. Whitaker and Robinson’s data show that the mean turned out to be 39.8 with a standard deviation of 2.0 for their sample of 10 000. If we put the two together, we end with a posterior mean for the chest measurements of men in this population is normal with variance  $\phi_1 = \{9^{-1} + (2^2/10\,000)^{-1}\}^{-1} = 1/2500$  and mean

$$\theta_1 = (1/2500)\{38/9 + 39.8/(2^2/10000)\} = 39.8.$$

Thus, for all practical purposes we have ended up with the distribution

$$N(39.8, 2^2/10000)$$

suggested by the data. You should note that this distribution is

$$N(\bar{x}, \phi/n),$$

the distribution we referred to in Section 2.1 on ‘Nature of Bayesian Inference’ as the *standardized likelihood*. Naturally, the closeness of the posterior to the standardized likelihood results from the large sample size, and whatever my prior had been, unless it were very very extreme, I would have got very much the same result. More formally, the posterior will be close to the standardized likelihood insofar as the weight  $\phi_1/\phi_0$

associated with the prior mean is small, that is insofar as  $\phi_0$  is large compared with  $\phi/n$ . This is reassuring in cases where the prior is not very easy to specify, although of course there are cases where the amount of data available is not enough to get to this comforting position.

### 2.3.3 Predictive distribution

If we consider taking another one observation  $x_{n+1}$ , then the predictive distribution can be found just as in Section 2.3 by writing  $x_{n+1} = (x_{n+1} - \theta) + \theta$  and noting that, independently of one another,

$$(x_{n+1} - \theta) \sim N(0, \phi),$$

$$\theta \sim N(\theta_1, \phi_1),$$

so that

$$x_{n+1} \sim N(\theta_1, \phi + \phi_1).$$

It is easy enough to adapt this argument to find the predictive distribution of an  $m$ -vector  $y = (y_1, y_2, \dots, y_m)$  where  $y_i \sim N(\theta, \psi)$

by writing

$$y = (y - \theta\mathbf{1}) + \theta\mathbf{1}$$

where  $\mathbf{1}$  is the constant vector

$$\mathbf{1} = (1, 1, \dots, 1).$$

Then  $\theta$  has its posterior distribution  $\theta | x$  and the components of the vector  $y - \theta\mathbf{1}$  are  $N(0, \psi)$  variates independent of  $\theta$  and of one another, so that  $y$  has a multivariate normal distribution, although its components are not independent of one another.

### 2.3.4 Robustness

It should be noted that *any* statement of a posterior distribution and *any* inference is conditional not merely on the data, but also on the assumptions made about the likelihood. So, in this section, the posterior distribution ends up being normal as a consequence partly of the prior but also of the assumption that the data was distributed normally, albeit with an unknown mean. We say that an inference is robust if it is not seriously affected by changes in the assumptions on which it is based. The notion of robustness is not one which can be pinned down into a more precise definition, and its meaning depends on the context, but nevertheless the concept is of great importance and increasing attention is paid in statistics to investigations of the robustness of various techniques. We can immediately say that the conclusion that the nature of the posterior is robust against changes in the prior is valid provided that the sample size is large and the prior is a not-too-extreme normal distribution or nearly so. Some detailed exploration of the notion of robustness (or sensitivity analysis) can be found in Kadane (1984).

## 2.4 Dominant likelihoods

### 2.4.1 Improper priors

We recall from the previous section that, when we have several normal observations with a normal prior and the variances are known, the posterior for the mean is  $N(\theta_1, \phi_1)$ ,

where  $\theta_1$  and  $\phi_1$  are given by the appropriate formulae and that this approaches the standardized likelihood  $N(\bar{x}, \phi/n)$

insofar as  $\phi_0$  is large compared with  $\phi/n$ , although this result is only approximate unless  $\phi_0$  is infinite. However, this would mean a prior density  $N(\theta_0, \infty)$  which, whatever  $\theta_0$  were, would have to be uniform over the whole real line, and clearly could not be represented by any *proper* density function. It is basic to the concept of a probability density that it integrates to 1 so, for example,  $p(\theta) = \kappa \quad (-\infty < \theta < \infty)$

cannot possibly represent a probability density whatever  $\kappa$  is, and in particular  $p(\theta) = 0$ , which results from substituting  $\phi = \infty$  into the normal density, cannot be a density. Nevertheless, we shall sometimes find it useful to extend the

concept of a probability density to some cases like this where  $\int_{-\infty}^{\infty} p(\theta) d\theta = \infty$  which we shall call *improper* ‘densities’. The density  $p(\theta) = \kappa$  can then be

regarded as representing a normal density of infinite variance. Another example of an improper density we will have use for later on is  $p(\theta) = \kappa/\theta$  ( $0 < \theta < \infty$ ). It turns out that sometimes when we take an improper prior density then it can combine with an ordinary likelihood to give a posterior which is proper. Thus, if we use the uniform distribution on the whole real line  $p(\theta) = \kappa$  for some  $\kappa \neq 0$ , it is easy to see that it combines with a normal likelihood to give the standardized likelihood as posterior; it follows that the dominant feature of the posterior is the likelihood. The best way to think of an improper density is as an approximation which is valid for some large range of values, but is not to be regarded as truly valid throughout its range. In the case of a physical constant which you are about to measure, you may be very unclear what its value is likely to be, which would suggest the use of a prior that was uniform or nearly so over a large range, but it seems unlikely that you would regard values in the region of, say,  $10^{100}$  as being as likely as, say, values in the region of  $10^{-100}$ . But if you have a prior which is approximately uniform over some (possibly very long) interval and is never very large outside it, then the posterior is close to the standardized likelihood, and so to the posterior which would have resulted from taking an improper prior uniform over the whole real line. [It is possible to formalize the notion of an improper density as part of probability theory – for details, see Rényi (1970).]

## 2.4.2 Approximation of proper priors by improper priors

This result can be made more precise. The following theorem is proved by Lindley (1965, Section 5.2); the proof is omitted.

**Theorem 2.1** A random sample  $x = (x_1, x_2, \dots, x_n)$  of size  $n$  is taken from  $N(\theta, \phi)$  where  $\phi$  is known. Suppose that there exist positive constants  $\alpha, \varepsilon, M$  and  $c$  depending on  $x$  (small values of  $\alpha$  and  $\varepsilon$  are of interest), such that in the interval  $I_\alpha$  defined by  $\bar{x} - \lambda_\alpha \sqrt{(\phi/n)} \leq \theta \leq \bar{x} + \lambda_\alpha \sqrt{(\phi/n)}$ ,

where

$$2\Phi(-\lambda_\alpha) = \alpha$$

the prior density of  $\theta$  lies between  $c(1 - \varepsilon)$  and  $c(1 + \varepsilon)$  and outside  $I_\alpha$  it is bounded by  $Mc$ . Then the posterior density  $p(\theta|x)$  satisfies

$$\frac{(1 - \varepsilon)}{(1 + \varepsilon)(1 - \alpha) + M\alpha} (2\pi\phi/n)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\bar{x} - \theta)^2/(\phi/n)\right\}$$

$$\leq p(\theta|x)$$

$$\leq \frac{(1 + \varepsilon)}{(1 - \varepsilon)(1 - \alpha)} (2\pi\phi/n)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\bar{x} - \theta)^2/(\phi/n)\right\}$$

inside  $I_\alpha$ , and

$$0 \leq p(\theta|x) \leq \frac{M}{(1-\varepsilon)(1-\alpha)} (2\pi\phi/n)^{-\frac{1}{2}} \exp\{-\frac{1}{2}\lambda_\alpha^2\}$$

outside it.

While we are not going to prove the theorem, it might be worth while to give some idea of the sorts of bounds which it implies. Anyone who has worked with the normal distribution is likely to remember that the 1% point is 2.58, that is that if  $\alpha = 0.01$  then  $\lambda_\alpha = 2.58$ , so that  $I_\alpha$  extends 2.58 standard deviations [ $2.58\sqrt{(\phi/n)}$ ] on each side of the sample mean  $\bar{x}$ . Suppose then that *before* you had obtained any data you believed all values in some interval to be equally likely, and that there were no values that you believed to be more than three times as probable as the values in this interval. If then it turns out when you get the data the range  $I_\alpha$  lies entirely in this interval, then you can apply the theorem with  $\alpha = 0.01$ ,  $\varepsilon = 0$ , and  $M = 3$ , to deduce that within  $I_\alpha$  the true density lies within multiples  $(1 - \alpha + Ma)^{-1} = 0.98$  and  $(1 - \alpha)^{-1} = 1.01$  of the normal density. We can regard this theorem as demonstrating how robust the posterior is to changes in the prior. Similar results hold for distributions other than the normal.

It is often sensible to analyze scientific data on the assumption that the likelihood dominates the prior. There are several reasons for this, of which two important ones are as follows. Firstly, even if you and I both have strong prior beliefs about the value of some unknown quantity, we might not agree, and it seems sensible to use a neutral *reference prior* which is dominated by the likelihood and could be said to represent the views of someone who (unlike ourselves) had no strong beliefs a priori. The difficulties of public discourse in a world where different individuals have different prior beliefs constitute one reason why a few people have argued that, in the absence of agreed prior information, we should simply quote the likelihood function [see Edwards, 1992], but there are considerable difficulties in the way of this (see also Section 7.1 on ‘The Likelihood Principle’). Secondly, in many scientific contexts, we would not bother to carry out an experiment unless we thought it was going to increase our knowledge significantly, and if that is the case then the likelihood will presumably dominate the prior.

## 2.5 Locally uniform priors

### 2.0.1 Bayes' Postulate

We have already said that it seems useful to have a reference prior to aid public discourse in situations where prior opinions differ or are not strong. A prior which does not change very much over the region in which the likelihood is appreciable and does not take very large values outside that region is said to be locally uniform. For such a prior  $p(\theta|x) \propto p(x|\theta) = l(\theta|x)$ ,

so that on normalizing the posterior must equal the standardized likelihood.

Bayes himself appears to have thought that, at least in the case where  $\theta$  is an unknown probability between 0 and 1, the situation where we ‘know nothing’ should be represented by taking a uniform prior and this is sometimes known as Bayes’ postulate (as distinct from his theorem).

However, it should be noted that if, for example

$$p(\theta) = 1 \quad (0 < \theta < 1)$$

then on writing

$$\phi = 1/\theta$$

we have according to the usual change-of-variable rule

$$p(\phi) |d\phi| = p(\theta) |d\theta|$$

or

$$\begin{aligned} p(\phi) &= p(\theta) |d\theta/d\phi| \\ &= 1/\phi^2 \quad (1 < \phi < \infty) \end{aligned}$$

(as a check, this density does integrate to unity). Now it has been argued that if we ‘know nothing’ about  $\theta$  then we equally ‘know nothing’ about  $\phi$ , which should surely be represented by the improper prior  $p(\phi) = \text{constant} \quad (1 < \phi < \infty)$

[although one can also argue for a prior proportional to  $\phi^{-1}$  or to  $(\phi - 1)^{-1}$ ], so that the idea that a uniform prior can be used to represent ignorance is not self-consistent. It cannot be denied that this is a serious objection, but it is perhaps not quite as serious as it seems at first sight. With most transformations, the density of the transformed variable will not change very fast over a reasonably short interval. For example, while  $1/\phi^2$  changes quite considerably over long intervals of  $\phi$ , it is sufficiently close to constancy over any moderately short interval that a posterior based a uniform prior is unlikely to differ greatly from one based on the prior with density  $1/\phi^2$ , provided that the amount of data available is not very small. This argument would not necessarily work if you were to consider a very extreme transformation, for example  $\phi = \exp(\exp(\theta))$ , but it could be argued that the mere fact that such an extreme transformation even crossed your mind would suggest that you had really got some prior information which made it sensible, and you should accordingly make use of your prior

information.

## 2.5.2 Data translated likelihoods

Even though it may not make a great deal of difference within broad limits what we treat as our reference prior, provided that it is reasonably flat, there is still a natural urge to look for the ‘right’ scale of measurement in which to have a uniform prior, from which the prior in any other scale of measurement can be deduced. One answer to this is to look for a scale of measurement in which the likelihood is *data translated*. The likelihood is said to be in such a form if  $l(\theta|x) = g(\theta - t(x))$

for some function  $t$  (which we will later note is a *sufficient statistic*). In looking to see whether the likelihood can be expressed in this way, you should bear in mind that the definition of the likelihood function allows you to multiply it by any function of the data  $x$  alone.

For example, if we have an  $n$ -sample from a normal distribution of unknown mean and known variance  $\phi$ , we know that  $l(\theta|x) = \exp\{-\frac{1}{2}(\theta - \bar{x})^2/(\phi/n)\}$  which is clearly of this form. On the other hand, if  $k$  has a binomial distribution of index  $n$  and parameter  $\pi$ , so that  $l(\pi|k) = \pi^k(1 - \pi)^{n-k}$  then the likelihood cannot be put into the form  $g(\pi - t(k))$ .

If the likelihood is in data translated form, then different values of the data will give rise to the same functional form for the likelihood except for a shift in location. Thus, in the case of the normal mean, if we consider two experiments, one of which results in a value of  $\bar{x}$  which is, say, 5 larger than the other, then we get the same likelihood function in both cases except that corresponding values of  $\theta$  differ by 5. This would seem to suggest that the main function of the data is to determine the location of the likelihood. Now if a uniform prior is taken for  $\theta$ , the posteriors are also the same except that corresponding values differ by 5, so that the inferences made do seem simply to represent a determination of location. It is because of this that it seems sensible to adopt a uniform prior when the likelihood is data translated.

## 2.5.3 Transformation of unknown parameters

The next question is what we should do when it is not. Sometimes it turns out that there is a function

$$\psi = \psi(\theta)$$

which is such that we can write

$$l(\theta|x) = g(\psi(\theta) - t(x))$$

in which case the obvious thing to do is to take a prior uniform in  $\psi$  rather than in  $\theta$ , implying a prior for the parameter  $\theta$  given by the usual change-of-variable rule. If, for example,  $x$  has an exponential distribution, that is  $x \sim E(\theta)$  (see under the Gamma distribution in Appendix A), then  $p(x|\theta) = \theta^{-1} \exp(-x/\theta)$ , so that (after multiplying by  $x$  as we are entitled to) we may write

$$\begin{aligned} l(\theta|x) &= (x/\theta) \exp(-x/\theta) \\ &= \exp\{\log x - \log \theta\} - \exp\{\log x - \log \theta\}. \end{aligned}$$

This is in the above with

$$g(y) = \exp\{y - \exp(-y)\}$$

$$t(x) = \log x$$

$$\psi(\theta) = \log \theta.$$

Unfortunately, it is often difficult to see how to express a likelihood function in this form even when it is possible, and it is not always possible. We shall find another case of this when we come to investigate the normal variance in Section 2.7, and a further one when we try to find a reference prior for the uniform distribution in Section 3.6. Sometimes there is a function  $\psi(\theta)$  such that the likelihood is approximately of this form, for example when we have a binomial distribution of known index and unknown parameter (this case will be considered when the reference prior for the binomial parameter is discussed in Section 3.2).

For the moment, we can reflect that this argument strengthens the case for using a uniform (improper) prior for the mean of a normal distribution  $p(\theta) \propto c \quad (-\infty < \theta < \infty)$ .

One way of thinking of the uniform distribution is as a normal distribution of infinite variance or equivalently zero precision. The equations for the case of

$$\text{Posterior precision} = \text{Prior precision} + \text{Datum precision}$$

normal mean and variance      Posterior mean = Weighted mean of prior mean and datum value  
then become

$$\text{Posterior precision} = \text{Datum precision}$$

$$\text{Posterior mean} = \text{Datum value}$$

which accords with the result

$$p(\theta|x) \propto l(\theta|x)$$

for a locally uniform prior. An interesting defence of the notion of a uniform prior can be found in Savage *et al.* (1962, p. 20).

## 2.6 Highest density regions

## 2.6.1 Need for summaries of posterior information

In the case of our example on Ennerdale granophyre, all the information available after the experiment is contained in the posterior distribution. One of the best ways of conveying this information would be to sketch the posterior density (though this procedure is more difficult in cases where we have several parameters to estimate, so that  $\theta$  is multi-dimensional). It is less trouble to the statistician to say simply that  $\theta \sim N(413, 7^2)$

although those without experience may need tables to appreciate what this assertion means.

Sometimes the probability that the parameter lies in a particular interval may be of interest. Thus, there might be geological reasons why, in the above example, we wanted to know the chance that the rocks were less than 400 million years old. If this is the case, the probability required is easily found by use of tables of the normal distribution. More commonly, there are no limits of any special interest, but it seems reasonable to specify an interval in which ‘most of the distribution’ lies. It would appear sensible to look for an interval which is such that the density at any point inside it is greater than the density at any point outside it, and it would also appear sensible to seek (for a given probability level) an interval that is as short as possible (in several dimensions, this means that it should occupy as small a volume as possible). Fortunately, it is clear that these conditions are equivalent. In most common cases, there is one such interval for each probability level.

We shall refer to such an interval as a *highest (posterior) density region* or an *HDR*. Although this terminology is used by several authors, there are other terms in use, for example *Bayesian confidence interval* (cf. Lindley 1965, Section 5.2) and *credible interval* (cf. Edwards *et al.* 1963, Section 5). In the particular example referred to aforementioned text, we could use the well-known fact that 95% of the area of a normal distribution lies within  $\pm 1.96$  standard deviations of the mean to say that  $413 \pm 1.96 \times 7$ , that is  $(399, 427)$  is a 95% HDR for the age  $\theta$  given the data.

## 2.6.2 Relation to classical statistics

The traditional approach, sometimes called *classical statistics* or *sampling theory statistics* would lead to similar conclusions in this case. From either standpoint  $(\theta - x)/\sqrt{\phi} \sim N(0, 1)$

and in either case the interval (399, 427) is used at a 95% level. However, in the classical approach, it is  $x$  that is regarded as random and giving rise to a random interval which has a probability 0.95 of containing the fixed (but unknown) value  $\theta$ . By contrast, the Bayesian approach regards  $\theta$  as random in the sense that we have certain beliefs about its value, and think of the interval as fixed once the datum is available. Perhaps the tilde notation for random variables helps. With this, the classical approach amounts to saying that  $|(\theta - \tilde{x})/\sqrt{\phi}| < 1.96$  with probability 0.95, while the Bayesian approach amounts to saying that

$$|(\tilde{\theta} - x)/\sqrt{\phi}| < 1.96$$

with probability 0.95.

Although there is a simple relationship between the conclusions that classical and Bayesian statisticians would arrive at in this case, there will be cases later on in which there is no great similarity between the conclusions arrived at.

## 2.7 Normal variance

### 2.7.1 A suitable prior for the normal variance

Suppose that we have an  $n$ -sample  $x = (x_1, x_2, \dots, x_n)$  from  $N(\mu, \phi)$  where the variance  $\phi$  is unknown but the mean  $\mu$  is known. Then clearly  $p(x|\phi) = (2\pi\phi)^{-\frac{1}{2}} \exp\{-\frac{1}{2}(x_1 - \mu)^2/\phi\}$

$$\times \cdots \times (2\pi\phi)^{-\frac{1}{2}} \exp\{-\frac{1}{2}(x_n - \mu)^2/\phi\}$$

$$\propto \phi^{-n/2} \exp\left\{-\frac{1}{2} \sum (x_i - \mu)^2/\phi\right\}.$$

On writing

$$S = \sum (x_i - \mu)^2$$

(remember that  $\mu$  is known; we shall use a slightly different notation when it is not), we see that  $p(x|\phi) \propto \phi^{-n/2} \exp(-\frac{1}{2}S/\phi)$ .

In principle, we might have any form of prior distribution for the variance  $\phi$ . However, if we are to be able to deal easily with the posterior distribution (and, e.g. to be able to find HDRs easily from tables), it helps if the posterior distribution is of a ‘nice’ form. This will certainly happen if the prior is of a similar form to the likelihood, namely,  $p(\phi) \propto \phi^{-\kappa/2} \exp(-\frac{1}{2}S_0/\phi)$ ,

where  $\kappa$  and  $S_0$  are suitable constants. For reasons which will emerge, it is convenient to write  $\kappa = \nu + 2$ , so that  $p(\phi) \propto \phi^{-\nu/2-1} \exp(-\frac{1}{2}S_0/\phi)$

leading to the posterior distribution

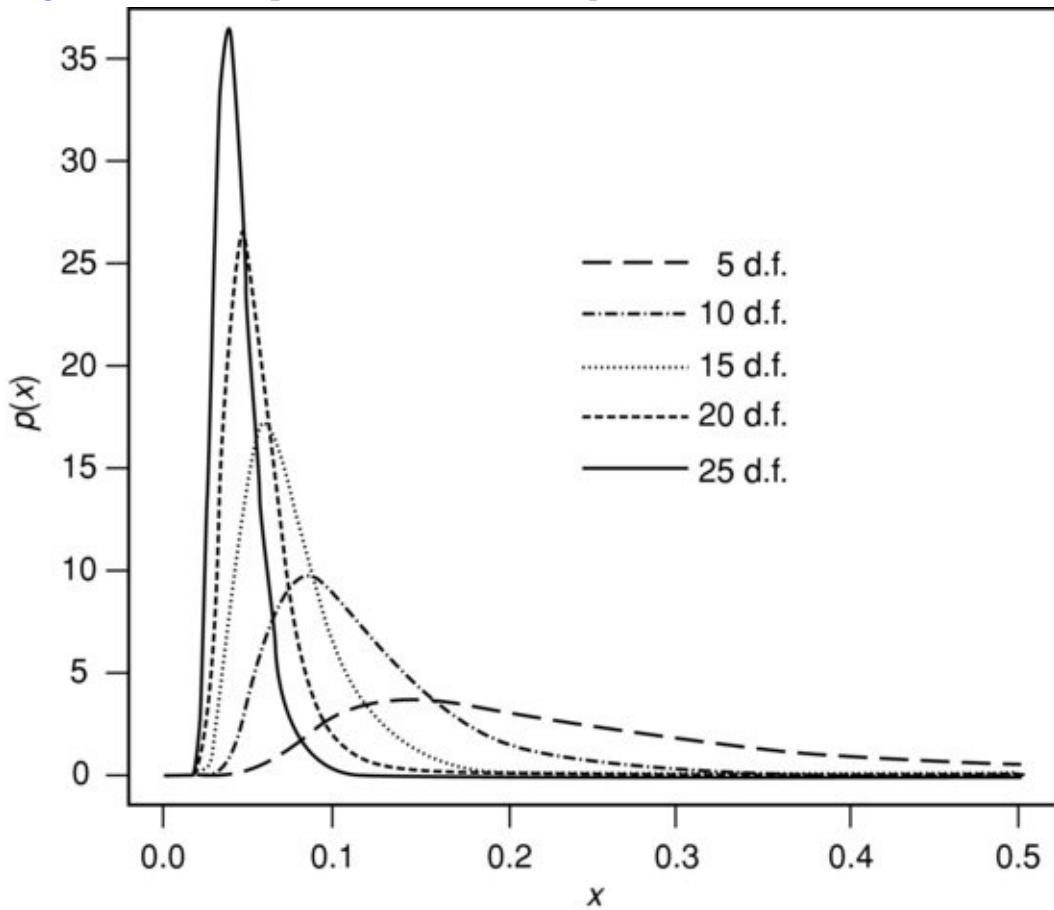
$$p(\phi|x) \propto p(\phi)p(x|\phi)$$

$$\propto \phi^{-(v+n)/2-1} \exp\{-\frac{1}{2}(S_0 + S)/\phi\}.$$

Although it is unlikely that our true prior beliefs are exactly represented by such a density, it is quite often the case that they can be reasonably well approximated by something of this form, and when this is so the calculations become notably simpler.

This distribution is, in fact, closely related to one of the best known continuous distributions in statistics (after the normal distribution), namely, the *chi-squared distribution* or  $\chi^2$  distribution. This is seen more clearly if we work in terms of the precision  $\lambda = 1/\phi$  instead of in terms of the variance  $\phi$ , since, using the change of variable rule introduced in Section 1.3 on ‘Random Variables’,

**Figure 2.1** Examples of inverse chi-squared densities.



$$p(\lambda|x) \propto p(\phi|x) |\mathrm{d}\phi / \mathrm{d}\lambda|$$

$$\propto \lambda^{(v+n)/2+1} \exp\{-\frac{1}{2}(S_0 + S)\lambda\} \times \lambda^{-2}$$

$$\propto \lambda^{(v+n)/2-1} \exp\{-\frac{1}{2}(S_0 + S)\lambda\}.$$

This is now very close to the form of the chi-squared distribution (see Appendix

A or indeed any elementary textbook on statistics). It is, in fact, a very simple further step (left as an exercise!) to check that (for given  $X$ )  $(S_0 + S)\lambda \sim \chi_{v+n}^2$  that is  $(S_0 + S)\lambda$  has a  $\chi^2$  distribution on  $v + n$  degrees of freedom. [The term ‘degrees of freedom’ is hallowed by tradition, but is just a name for a parameter.] We usually indicate this by writing  $\phi \sim (S_0 + S)\chi_{v+n}^{-2}$ , and saying that  $\phi$  has (a multiple of) an *inverse chi-squared distribution*.

Clearly the same argument can be applied to the prior distribution, so our prior assumption is that

$$\phi \sim S_0 \chi_v^{-2}.$$

It may be that you cannot find suitable values of the parameters  $v$  and  $S_0$ , so that a distribution of this type represents your prior beliefs, but clearly if values can be chosen, so that they are reasonably well approximated, it is convenient. Usually, the approximation need not be too close since, after all, the chances are that the likelihood will dominate the prior. In fitting a plausible prior one possible approach is to consider the mean and variance of your prior distribution

$$E\phi = S_0/(v - 2),$$

and then choose  $v$  and  $S_0$ , so that  $V\phi = \frac{2S_0^2}{(v - 2)^2(v - 4)}$ .

Inverse chi-squared distributions (and variables which have such a distribution apart from a constant multiple) often occur in Bayesian statistics, although the inverse chi-squared (as opposed to the chi-squared) distribution rarely occurs in classical statistics. Some of its properties are described in Appendix A, and its density for typical values of  $v$  (and  $S_0=1$ ) is illustrated in [Figure 2.1](#)

## 2.7.2 Reference prior for the normal variance

The next thing to investigate is whether there is something which we can regard as a reference prior by finding a scale of measurement in which the likelihood is data translated. For this purpose (and others), it is convenient to define the sample standard deviation  $s$  by  $s^2 = S/n = \sum(x_i - \mu)^2/n$

(again we shall use a different definition when  $\mu$  is unknown). Then

$$\begin{aligned} l(\phi|x) &\propto p(x|\phi) \propto \phi^{-n/2} \exp(-\frac{1}{2}S/\phi) \\ &\propto s^n \phi^{-n/2} \exp(-\frac{1}{2}ns^2/\phi) \\ &= \exp[-\frac{1}{2}n(\log \phi - \log s^2) - \frac{1}{2}ns^2/\phi] \\ &= \exp[-\frac{1}{2}n(\log \phi - \log s^2) - \frac{1}{2}n \exp\{-(\log \phi - \log s^2)\}]. \end{aligned}$$

This is of data translated form

$$l(\phi|x) = g(\psi(\phi) - t(x))$$

(cf. Section 2.4 on ‘Dominant likelihoods’) with

$$g(y) = \exp[-\frac{1}{2}ny - \frac{1}{2}n \exp(-y)]$$

$$t(x) = \log s^2$$

$$\psi(\phi) = \log \phi.$$

The general argument about data translated likelihoods now suggests that we take as reference prior an improper density which is locally uniform in  $\psi = \log \phi$ , that is  $p(\psi) \propto c$ , which, in terms of  $\phi$  corresponds to  $p(\phi) = p(\psi) |d\psi/d\phi| \propto d \log \phi / d\phi$  and so to

$$p(\phi) \propto 1/\phi.$$

(Although the above argument is complicated, and a similarly complicated example will occur in the case of the uniform distribution in 3.6, there will be no other difficult arguments about data translated likelihoods.) This prior (which was first mentioned in Section 2.4 on dominant likelihoods) is, in fact, a particular case of the priors of the form  $S_0 \chi_{\nu}^{-2}$ , which we were considering earlier, in which  $\nu = 0$  and  $S_0 = 0$ . Use of the reference prior results in a posterior distribution  $\phi \propto S \chi_n^{-2}$

which again is a particular case of the distribution found before, and is quite easy to use.

You should perhaps be warned that inferences about variances are not as robust as inferences about means if the underlying distribution turns out to be only approximately normal, in the sense that they are more dependent on the precise choice of prior distribution.

## 2.8 HDRs for the normal variance

### 2.8.1 What distribution should we be considering?

It might be thought that as the normal variance has (under the assumptions we are making) a distribution which is a multiple of the inverse chi-squared distribution we should be using tables of HDRs for the inverse chi-squared distribution to give intervals in which most of the posterior distribution lies. This procedure is, indeed, recommended by, for example, Novick and Jackson (1974,

Section 7.3) and Schmitt (1969, Section 6.3). However, there is another procedure which seems to be marginally preferable.

The point is that we chose a reference prior which was uniform in  $\log \phi$ , so that the density of  $\log \phi$  was constant and no value of  $\log \phi$  was more likely than any other a priori. Because of this, it seems natural to use  $\log \phi$  in the posterior distribution and thus to look for an interval inside which the *posterior density of  $\log \phi$*  is higher than anywhere outside. It might seem that this implies the use of tables of HDRs of log chi-squared, but in practice it is more convenient to use tables of the corresponding values of chi-squared, and such tables can be found in the Appendix. In fact, it does not make much difference whether we look for regions of highest density of the inverse chi-squared distribution or of the log chi-squared distribution, but insofar as there is a difference it seems preferable to base inferences on the log chi-squared distribution.

## 2.8.2 Example

When we considered the normal distribution with unknown mean but known variance, we had to admit that this was a situation which rarely occurred in real-life examples. This is even more true when it comes to the case where the mean is known and the variance unknown, and it should really be thought of principally as a building block towards the structure we shall erect to deal with the more realistic case where both mean and variance are unknown.

We shall, therefore, consider an example in which the mean was in fact unknown, but treat it *as if* the mean were known. The following numbers give the uterine weight (in mg) of 20 rats drawn at random from a large stock:

9	18	21	26
14	18	22	27
15	19	22	29
15	19	24	30
16	20	24	32

It is easily checked that  $n=20$ ,  $\sum x_i = 420$ ,  $\sum x_i^2 = 9484$ , so that  $\bar{x} = 21.0$  and  $S = \sum (x_i - \bar{x})^2 = \sum x_i^2 - (\sum x_i)^2 / n = 664$ .

In such a case, we do not know that the mean is 21.0 (or at least it is difficult to imagine circumstances in which we could have this information). However, we shall exemplify the methodology for the case where the mean is known by analyzing this data as if we knew that the mean were  $\mu = 21.0$ . If this were so, then we would be able to assert that  $\phi \propto 664 \chi_{20}^{-2}$ .

All the information we have about the variance  $\phi$  is contained in this statement,

but of course it is not necessarily easy to interpret from the point of view of someone inexperienced with the use of statistical methods (or even of someone who is but does not know about the inverse chi-squared distribution). Accordingly, it may be useful to give some idea of the distribution if we look for a HDR. From the tables in the Appendix, we see that the values of chi-squared corresponding to a 95% HDR for log chi-squared are 9.958 and 35.227, so that the interval for  $\phi$  is from  $664/35.227$  to  $664/9.958$ , that is is the interval (19, 67). (We note that it is foolish to quote too many significant figures in your conclusions, though it may be sensible to carry through extra significant figures in intermediate calculations.) It may be worth comparing this with the results from looking at HDRs for the inverse chi-squared distribution itself. From the tables in the Appendix A, 95% HDR for the inverse chi-squared distribution on 20 degrees of freedom lies between 0 025 and 0 094, so that the interval for  $\phi$  is from  $664 \times 0.025$  to  $664 \times 0.094$ , that is is the interval (17, 62). It follows that the two methods do not give notably different answers.

## 2.9 The role of sufficiency

### 2.9.1 Definition of sufficiency

When we considered the normal variance with known mean, we found that the posterior distribution depended on the data only through the single number  $S$ . It often turns out that the data can be reduced in a similar way to one or two numbers, and as long as we know them we can forget the rest of the data. It is this notion that underlies the formal definition of sufficiency.

Suppose observations  $x = (x_1, x_2, \dots, x_n)$  are made with a view to gaining knowledge about a parameter  $\theta$ , and that  $t = t(x)$

is a function of the observations. We call such a function a *statistic*. We often suppose that  $t$  is real valued, but it is sometimes vector valued. Using the formulae in Section 1.4 on ‘Several Random Variables’ and the fact that once we know  $x$  we automatically know the value of  $t$ , we see that for any statistic  $t$   $p(x|\theta) = p(x, t|\theta) = p(t|\theta) p(x|t, \theta)$ .

However, it sometimes happens that

$$p(x|t, \theta)$$

does not depend on  $\theta$ , so that

$$p(x|\theta) = p(t|\theta) p(x|t).$$

If this happens, we say that  $t$  is a *sufficient statistic* for  $\theta$  given  $X$ , often

abbreviated by saying that  $t$  is sufficient for  $\theta$ . It is occasionally useful to have a further definition as follows: a statistic  $u = u(x)$  whose density  $p(u|\theta) = p(u)$  does not depend on  $\theta$  is said to be *ancillary* for  $\theta$ .

## 2.9.2 Neyman's factorization theorem

The following theorem is frequently used in finding sufficient statistics:

**Theorem 2.2** A statistic  $t$  is sufficient for  $\theta$  given  $X$  if and only if there are functions  $f$  and  $g$  such that  $p(x|\theta) = f(t, \theta) g(x)$ ,  
where  $t = t(x)$ .

*Proof.* If  $t$  is sufficient for  $\theta$  given  $x$ , we may take  $f(t, \theta) = p(t|\theta)$  and  $g(x) = p(x|t)$ . Conversely, if the condition holds, then, in the discrete case, since once  $x$  is known then  $t$  is known,  $p(x, t|\theta) = f(t, \theta)g(x)$ .

We may then sum both sides of the equation over all values of  $X$  such that  $t(x) = t$  to get  $p(t|\theta) = f(t, \theta) G(t)$

where  $G(t)$  is obtained by summing  $g(x)$  over all these values of  $X$ , using the formula  $p(t) = \sum p(x, t)$ .

In the continuous case, write

$$A = \{x; t(x) = \tau\} \quad \text{and} \quad B = \{x; t(x) = t\}.$$

Then

$$P(\tilde{t} \leq t|\theta) = \int_{-\infty}^t \int_A f(\tau, \theta) g(x) dx d\tau,$$

so that on differentiating with respect to  $t$  we find that

$$\begin{aligned} p(t|\theta) &= \int_B f(t, \theta) g(x) dx \\ &= f(t, \theta) \int_B g(x) dx. \end{aligned}$$

Writing  $G(t)$  for the last integral, we get the same result as in the discrete case, *viz.*

$$p(t|\theta) = f(t, \theta) G(t).$$

From this it follows that

$$f(t, \theta) = p(t|\theta)/G(t).$$

Considering now any one value of  $X$  such that  $t(x) = t$  and substituting in the equation in the statement of the theorem we obtain  $p(x|\theta) = p(t|\theta) g(x)/G(t)$ .

Since whether  $t$  is sufficient or not

$$p(x|t, \theta) = p(x, t|\theta)/p(t|\theta) = p(x|\theta)/p(t|\theta)$$

we see that

$$p(x|t, \theta) = g(x)/G(t).$$

Since the right-hand side does not depend on  $\theta$ , it follows that  $t$  is indeed sufficient, and the theorem is proved. ■

### 2.9.3 Sufficiency principle

**Theorem 2.3** A statistic  $t$  is sufficient for  $\theta$  given  $X$  if and only if  $l(\theta|x) \propto l(\theta|t)$  whenever  $t = t(x)$  (where the constant of proportionality does not, of course, depend on  $\theta$ ).

*Proof.* If  $t$  is sufficient for  $\theta$  given  $X$  then  $l(\theta|x) \propto p(x|\theta) = p(t|\theta)p(x|t) \propto p(t|\theta) \propto l(\theta|t)$ .

Conversely, if the condition holds then

$$p(x|\theta) \propto l(\theta|x) \propto l(\theta|t) \propto p(t|\theta),$$

so that for some function  $g(x)$

$$p(x|\theta) = p(t|\theta)g(x).$$

The theorem now follows from the Factorization Theorem. ■

**Corollary 2.1** For any prior distribution, the posterior distribution of  $\theta$  given  $X$  is the same as the posterior distribution of  $\theta$  given a sufficient statistic  $t$ .

*Proof.* From Bayes' Theorem  $p(\theta|x)$  is proportional to  $p(\theta|t)$ ; they must then be equal as they both integrate or sum to unity. ■

**Corollary 2.2** If a statistic  $t = t(x)$  is such that  $l(\theta|x) \propto l(\theta|x')$  whenever  $t(x) = t(x')$ , then it is sufficient for  $\theta$  given  $X$ .

*Proof.* By summing or integrating over all  $X$  such that  $t(x) = t$ , it follows that  $l(\theta|t) \propto p(t|\theta) = \sum p(x'|\theta) \propto \sum l(\theta|x') \propto l(\theta|x)$

the summations being over all  $x'$  such that  $t(x') = t(x)$ . The result now follows from the theorem. ■

### 2.9.4 Examples

*Normal variance.* In the case where the  $x_i$  are normal of known mean  $\mu$  and unknown variance  $\phi$ , we noted that  $p(x|\phi) \propto \phi^{-n/2} \exp(-\frac{1}{2}S/\phi)$

where  $S = \sum(x_i - \mu)^2$ . It follows from the Factorization Theorem that  $S$  is sufficient for  $\mu$  given  $X$ . Moreover, we can verify the Sufficiency Principle as follows. If we had simply been given the value of  $S$  without being told the values of  $x_1, x_2, \dots, x_n$  separately, we could have noted that for each  $i$   $(x_i - \mu)/\sqrt{\phi} \sim N(0, 1)$ ,

so that  $S/\phi$  is a sum of squares of  $n$  independent  $N(0, 1)$  variables. Now a  $\chi^2_n$

distribution is often *defined* as being the distribution of the sum of squares of  $n$  random variables with an  $N(0, 1)$  distribution, and the density of  $\chi_n^2$  can be deduced from this. It follows that  $S/\phi \sim \chi_n^2$  and hence if  $y = S/\phi$  then  $p(y) \propto y^{n/2-1} \exp(-\frac{1}{2}y)$ .

Using the change of variable rule it is then easily seen that

$$p(S|\phi) \propto S^{n/2-1} \phi^{-n/2} \exp(-\frac{1}{2}S/\phi).$$

We can thus verify the Sufficiency Principle in this particular case because

$$l(\phi|x) \propto p(x|\phi) \propto S^{-n/2+1} p(S|\phi) \propto p(S|\phi) \propto l(\phi|S).$$

*Exponential distribution.* Let us suppose that  $x_i$  ( $i = 1, 2, \dots, n$ ) are independently distributed with an exponential distribution (see under the gamma distribution in Appendix A) so that  $x_i \sim E(\theta)$  or equivalently  $x_i \sim G(1, \theta)$ . Then

$$\begin{aligned} p(x|\theta) &= \theta^{-n} \exp\left(-\sum x_i/\theta\right) \\ &= \theta^{-n} \exp(-S/\theta) \end{aligned}$$

where  $S = \sum x_i$ . It follows from the Factorization Theorem that  $S$  is sufficient for  $\theta$  given  $S$ . It is also possible to verify the Sufficiency Principle in this case. In this case it is not hard to show that  $S \sim G(n, \theta)$  so that  $p(S|\theta) = \frac{1}{\theta^n \Gamma(n)} S^{n-1} \exp(-S)$  and we find

$$l(\theta|x) = \theta^{-n} \exp(-S/\theta) \propto \frac{1}{\theta^n \Gamma(n)} S^{n-1} \exp(-S) \propto p(S|\theta) \propto l(\theta|S).$$

*Poisson case.* Recall that the integer-valued random variable  $x$  is said to have a Poisson distribution of mean  $\lambda$  [denoted  $x \sim P(\lambda)$ ] if  $p(x|\lambda) = (\lambda^x/x!) \exp(-\lambda)$  ( $x = 0, 1, 2, \dots$ ).

We shall consider the Poisson distribution in more detail later in the book. For the moment, all that matters is that it often serves as a model for the number of occurrences of a rare event, for example for the number of times the King's Arms on the riverbank at York is flooded in a year. Then if  $x_1, x_2, \dots, x_n$  have independent Poisson distributions with the same mean (so could, e.g. represent the numbers of floods in several successive years), it is easily seen that  $p(x|\lambda) \propto \lambda^T \exp(-n\lambda)$ ,

where

$$T = \sum x_i.$$

It follows from the Factorization Theorem that  $T$  is sufficient for  $\lambda$  given  $X$ . Moreover, we can verify the Sufficiency Principle as follows. If we had simply been given the value of  $T$  without being given the values of the  $x_i$  separately, we could have noted that a sum of independent Poisson distributions has a Poisson distribution with mean the sum of the means (see question 7 on Chapter 1), so

that  $T \sim P(n\lambda)$ ,

and hence,

$$l(\lambda|T) \propto p(T|\lambda) = \{(n\lambda)^T / T!\} \exp(-n\lambda) \propto p(x|\lambda) \propto l(\lambda|x)$$

in accordance with the sufficiency principle.

## 2.9.5 Order statistics and minimal sufficient statistics

It may be noted that it is easy to see that whenever  $x = (x_1, x_2, \dots, x_n)$  consists of independently identically distributed observations whose distribution depends on a parameter  $\theta$ , then the order statistic  $x_{(O)} = (x_{(1)}, x_{(2)}, \dots, x_{(n)})$

which consists of the values of the  $x_i$  arranged in increasing order, so that  $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$

is sufficient for  $\theta$  given  $X$ .

This helps to underline the fact that there is, in general, no such thing as a *unique* sufficient statistic. Indeed, if  $t$  is sufficient for  $\theta$  given  $X$ , then so is  $(t, u)$  for *any* statistic  $u(x)$ . If  $t$  is a function of all other sufficient statistics that can be constructed, so that no further reduction is possible, then  $t$  is said to be *minimal sufficient*. Even a minimal sufficient statistic is not unique, since any one function of such a statistic is itself minimal sufficient.

It is not obvious that a minimal sufficient statistic always exists, but in fact, it does. Although the result is more important than the proof, we shall now prove this. We define a statistic  $u(x)$  which is a *set*, rather than a real number or a vector, by  $u(x) = \{x'; l(\theta|x') \propto l(\theta|x)\}$ .

Then it follows from Corollary 2.2 to the Sufficiency Principle that  $u$  is sufficient. Further, if  $v = v(x)$  is any other sufficient statistic, then by the same principle whenever  $v(x') = v(x)$ , we have  $l(\theta|x') \propto l(\theta|v) \propto l(\theta|x)$ ,

and hence,  $u(x') = u(x)$ , so that  $u$  is a function of  $v$ . It follows that  $u$  is minimal sufficient. We can now conclude that the condition that  $t(x') = t(x)$  if and only if

$$l(\theta|x') \propto l(\theta|x)$$

is equivalent to the condition that  $t$  is minimal sufficient.

## 2.9.6 Examples on minimal sufficiency

*Normal variance.* In the case where the  $x_i$  are independently  $N(\mu, \phi)$  where  $\mu$  is

known but  $\phi$  is unknown, then  $S$  is not merely sufficient but minimal sufficient.

*Poisson case.* In the case where the  $x_i$  are independently  $P(\lambda)$ , then  $T$  is not merely sufficient but minimal sufficient.

*Cauchy distribution.* We say that  $x$  has a Cauchy distribution with location parameter  $\theta$  and scale parameter 1, denoted  $x \sim \text{C}(\theta, 1)$  if it has density  $p(x|\theta) = \pi^{-1}\{1 + (x - \theta)^2\}^{-1}$ .

It is hard to find examples of real data which follow a Cauchy distribution, but the distribution often turns up in counter-examples in theoretical statistics (e.g. a mean of  $n$  variables with a  $\text{C}(\theta, 1)$  distribution has itself a  $\text{C}(\theta, 1)$  distribution and does not tend to normality as  $n$  tends to infinity in apparent contradiction of the Central Limit Theorem). Suppose that  $x_1, x_2, \dots, x_n$  are independently  $\text{C}(\theta, 1)$ . Then if  $l(\theta|x') \propto l(\theta|x)$ , we must have  $\prod\{1 + (x'_k - \theta)^2\} \propto \prod\{1 + (x_k - \theta)^2\}$ .

By comparison of the coefficients of  $\theta^{2n}$  the constant of proportionality must be 1 and by comparison of the zeroes of both sides considered as polynomials in  $\theta$ , namely,  $x'_k \pm i$  and  $x_k \pm i$ , respectively, we see that the  $x'_k$  must be a permutation of the  $x_k$  and hence the order statistics  $x'_{(O)}$  and  $x_{(O)}$  of  $x'$  and  $X$  are equal. It follows that the order statistic  $x_{(O)}$  is a minimal sufficient statistic, and in particular there is no one-dimensional sufficient statistic. This sort of situation is unusual with the commoner statistical distributions, but you should be aware that it can arise, even if you find the above proof confusing.

A useful reference for advanced workers in this area is Huzurbazar (1976).

## 2.10 Conjugate prior distributions

### 2.10.1 Definition and difficulties

When the normal variance was first mentioned, it was stated said that it helps if the prior is of such that the posterior is of a ‘nice’ form, and this led to the suggestion that if a reasonable approximation to your prior beliefs could be managed by using (a multiple of) an inverse chi-squared distribution, it would be sensible to employ this distribution. It is this thought which leads to the notion of conjugate families. The usual definition adopted is as follows: Let  $l$  be a likelihood function  $l(\theta|x)$ . A class  $\Pi$  of prior distributions is said to form a *conjugate family* if the posterior density  $p(\theta|x) \propto p(\theta)l(\theta|x)$  is in the class  $\Pi$  for all  $X$  whenever the prior density is in  $\Pi$ .

There is actually a difficulty with this definition, as was pointed out by Diaconis and Ylvisaker (1979 and 1985). If  $\Pi$  is a conjugate family and  $q(\theta)$  is any fixed function, then the family  $\Psi$  of densities proportional to  $q(\theta)p(\theta)$  for  $p \in \Pi$  is also a conjugate family. While this is a logical difficulty, we are in practice only interested in ‘natural’ families of distributions which are at least simply related to the standard families that are tabulated. In fact, there is a more precise definition available when we restrict ourselves to the exponential family (discussed in Section 2.11), and there are not many cases discussed in this book that are not covered by that definition. Nevertheless, the usual definition gives the idea well enough.

## 2.10.2 Examples

*Normal mean.* In the case of several normal observations of known variance with a normal prior for the mean (discussed in Section 2.3), where  $l(\theta|x) \propto \exp\{-\frac{1}{2}(\theta - \bar{x})^2/(\phi/n)\}$

we showed that if the prior  $p(\theta)$  is  $N(\theta_0, \phi_0)$  then the posterior  $p(\theta|x)$  is  $N(\theta_1, \phi_1)$  for suitable  $\theta_1$  and  $\phi_1$ . Consequently if  $\Pi$  is the class of all normal distributions, then the posterior is in  $\Pi$  for all  $X$  whenever the prior is in  $\Pi$ . Note, however, that it would not do to let  $\Pi$  be the class of all normal distributions with any mean but fixed variance  $\phi$  (at least unless we regard the sample size as fixed once and for all);  $\Pi$  must in some sense be ‘large enough.’

*Normal variance.* In the case of the normal variance, where

$$l(\phi|x) \propto \phi^{-n/2} \exp(-\frac{1}{2}S/\phi),$$

we showed that if the prior  $p(\phi)$  is  $S_0 \chi_{\nu}^{-2}$  then the posterior  $p(\phi|x)$  is  $(S_0 + S) \chi_{\nu+n}^{-2}$ . Consequently, if  $\Pi$  is the class of distributions of constant multiples of inverse chi-squares, then the posterior is in  $\Pi$  whenever the prior is. Again, it is necessary to take  $\Pi$  as a two-parameter rather than a one-parameter family.

*Poisson distribution.* Suppose  $x = (x_1, x_2, \dots, x_n)$  is a sample from the Poisson distribution of mean  $\lambda$ , that is  $x_i \sim P(\lambda)$ . Then as we noted in the last section  $l(\lambda|x) \propto \lambda^T \exp(-n\lambda)$ ,

where  $T = \sum x_i$ . If  $\lambda$  has a prior distribution of the form

$$p(\lambda) \propto \lambda^{\nu/2-1} \exp(-\frac{1}{2}S_0\lambda)$$

that is  $\lambda \sim S_0^{-1} \chi_{\nu}^2$ , so that  $\lambda$  is a multiple of a chi-squared random variable, then

$$p(\lambda|x) \propto p(\lambda)l(\lambda|x)$$

the posterior is  $\propto \lambda^{(v+2T)/2-1} \exp\{-\frac{1}{2}(S_0 + 2n)\lambda\}$ .

Consequently, if  $\Pi$  is the class distributions of constant multiples of chi-squared random variables, then the posterior is in  $\Pi$  whenever the prior is. There are three points to be drawn to your attention. Firstly, this family is closely related to, but different from, the conjugate family in the previous example. Secondly, the conjugate family consists of a family of continuous distributions although the observations are discrete; the point is that this discrete distribution depends on a continuous parameter. Thirdly, the conjugate family in this case is usually referred to in terms of the gamma distribution, but the chi-squared distribution is preferred here in order to minimize the number of distributions you need to know about and because when you need to use tables, you are likely to refer to tables of chi-squared in any case; the two descriptions are of course equivalent.

*Binomial distribution.* Suppose that  $k$  has a binomial distribution of index  $n$  and parameter  $\pi$ . Then  $l(\pi|k) \propto \pi^k(1-\pi)^{n-k}$ .

We say that  $\pi$  has a beta distribution with parameters  $\alpha$  and  $\beta$ , denoted  $Be(\alpha, \beta)$  if its density is of the form  $p(\pi) \propto \pi^{\alpha-1}(1-\pi)^{\beta-1}$

(the fact that  $\alpha - 1$  and  $\beta - 1$  appear in the indices rather than  $\alpha$  and  $\beta$  is for technical reasons). The beta distribution is described in more detail in Appendix A. If, then,  $\pi$  has a beta prior density, it is clear that it has a beta posterior density, so that the family of beta densities forms a conjugate family. It is a simple extension that this family is still conjugate if we have a sample of size  $k$  rather than just one observation from a binomial distribution.

### 2.10.3 Mixtures of conjugate densities

Suppose we have a likelihood  $l(\theta|x)$  and  $p_1(\theta)$  and  $p_2(\theta)$  are both densities in a conjugate family  $\Pi$  which give rise to posteriors  $p_1(\theta|x)$  and  $p_2(\theta|x)$ , respectively. Let  $\alpha$  and  $\beta$  be any non-negative real numbers summing to unity, and write  $p(\theta) = \alpha p_1(\theta) + \beta p_2(\theta)$ .

Then (taking a little more care with constants of proportionality than usual) it is easily seen that the posterior corresponding to the prior  $p(\theta)$  is

$$\begin{aligned}
p(\theta|x) &= p(x|\theta)p(\theta)/p(x) \\
&= \frac{\alpha p(x|\theta)p_1(\theta) + \beta p(x|\theta)p_2(\theta)}{\alpha \int p(x|\theta)p_1(\theta) d\theta + \beta \int p(x|\theta)p_2(\theta) d\theta} \\
&= \frac{\alpha p_1(\theta|x) \int p(x|\theta)p_1(\theta) d\theta + \beta p_2(\theta|x) \int p(x|\theta)p_2(\theta) d\theta}{\alpha \int p(x|\theta)p_1(\theta) d\theta + \beta \int p(x|\theta)p_2(\theta) d\theta} \\
&= \alpha' p_1(\theta|x) + \beta' p_2(\theta|x),
\end{aligned}$$

where

$$\alpha' \propto \alpha \int p(x|\theta)p_1(\theta) d\theta \quad \text{and} \quad \beta' \propto \beta \int p(x|\theta)p_2(\theta) d\theta$$

with the constant of proportionality being such that

$$\alpha' + \beta' = 1.$$

More generally, it is clearly possible to take any convex combination of more than two priors in  $\Pi$  and get a corresponding convex combination of the respective posteriors. Strictly in accordance with the definition given, this would allow us to extend the definition of  $\Pi$  to include all such convex combinations, but this would not retain the ‘naturalness’ of families such as the normal or the inverse chi-squared.

The idea can, however, be useful if, for example, you have a bimodal prior distribution. An example quoted by Diaconis and Ylvisaker (1985) is as follows. To follow this example, it may help to refer to Section 3.1 on ‘The binomial distribution’, or to return to it after you have read that section. Diaconis and Ylvisaker observe that there is a big difference between spinning a coin on a table and tossing it in the air. While tossing often leads to about an even proportion of ‘heads’ and ‘tails’, spinning often leads to proportions like  $\frac{1}{3}$  or  $\frac{2}{3}$ ; we shall write  $\pi$  for the proportion of heads. They say that the reasons for this bias are not hard to infer, since the shape of the edge will be a strong determining factor – indeed magicians have coins that are slightly shaved; the eye cannot detect the shaving but the spun coin always comes up ‘heads’. Assuming that they were not dealing with one of the said magician’s coins, they thought that a fifty-fifty mixture (i.e.  $\alpha = \beta = \frac{1}{2}$ ) of two beta densities, namely,  $Be(10, 20)$  (proportional to  $\pi^{10-1}(1-\pi)^{20-1}$ ) and  $Be(20, 10)$  (proportional to  $\pi^{20-1}(1-\pi)^{10-1}$ ), would seem a reasonable prior (actually they consider other possibilities as well). This is a bimodal distribution, which of course no beta density is, having modes, that is maxima of the density, near to the modes  $\pi = 0.32$  and at  $\pi = 0.68$  of the components.

They then spun a coin ten times, getting ‘heads’ three times. This gives a likelihood proportional to  $\pi^3(1-\pi)^7$  and so

$$\alpha' \propto \frac{1}{2} \int \pi^{13-1} (1-\pi)^{27-1} d\pi, \quad \beta' \propto \frac{1}{2} \int \pi^{23-1} (1-\pi)^{17-1} d\pi$$

that is,

$$\alpha' \propto B(13, 27), \quad \beta' \propto B(23, 17)$$

or, since  $13+27=23+17$ ,

$$\alpha' \propto \Gamma(13)\Gamma(27), \quad \beta' \propto \Gamma(23)\Gamma(17).$$

From the fact that  $\Gamma(n) = (n-1)!$ , it is easily deduced that

$$\alpha' = 115/129, \quad \beta' = 14/129.$$

We can deduce some properties of this posterior from those of the component betas. For example, the probability that  $\pi$  is greater than 0.5 is the sum 115/129 times the probability that a  $Be(13, 27)$  is greater than 0.5 and 14/129 times the probability that a  $Be(27, 13)$  is greater than 0.5; and similarly the mean is an appropriately weighted average of the means.

These ideas are worth bearing in mind if you have a complicated prior which is not fully dominated by the data, and yet want to obtain a posterior about which at least something can be said without complicated numerical integration.

## 2.10.4 Is your prior really conjugate?

The answer to this question is, almost certainly, ‘No’. Nevertheless, it is often the case that the family of conjugate priors is large enough that there is one that is sufficiently close to your real prior beliefs that the resulting posterior is barely distinguishable from the posterior that comes from using your real prior. When this is so, there are clear advantages in using a conjugate prior because of the greater simplicity of the computations. You should, however, be aware that cases can arise when no member of the conjugate family is, in the aforementioned sense, close enough, and then you may well have to proceed using numerical integration if you want to investigate the properties of the posterior.

# 2.11 The exponential family

## 2.11.1 Definition

It turns out that many of the common statistical distributions have a similar form. This leads to the definition that a density is from the *one-parameter exponential family* if it can be put into the form  $p(x|\theta) = g(x)h(\theta) \exp\{t(x)\psi(\theta)\}$  or equivalently if the likelihood of  $n$  independent observations  $x = (x_1, x_2, \dots, x_n)$

from this distribution is  $l(\theta|x) \propto h(\theta)^n \exp\left\{\sum t(x_i)\psi(\theta)\right\}$ .

It follows immediately from Neyman's Factorization Theorem that  $\Sigma t(x_i)$  is sufficient for  $\theta$  given  $X$ .

## 2.11.2 Examples

*Normal mean.* If  $x \sim N(\theta, \phi)$  with  $\phi$  known then

$$p(x|\theta) = (2\pi\phi)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(x-\theta)^2/\phi\right\}$$

$$= [(2\pi\phi)^{-\frac{1}{2}} \exp(-\frac{1}{2}x^2/\phi)] \exp(-\frac{1}{2}\theta^2/\phi) \exp(x\theta/\phi)$$

which is clearly of the above form.

*Normal variance.* If  $x \sim N(\theta, \phi)$  with  $\theta$  known then we can express the density in the appropriate form by writing  $p(x|\phi) = (2\pi)^{-\frac{1}{2}}\phi^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(x-\theta)^2/\phi\right\}$ .

*Poisson distribution.* In the Poisson case, we can write

$$p(x|\lambda) = (1/x!) \exp(-\lambda) \exp\{x(\log \lambda)\}.$$

*Binomial distribution.* In the binomial case we can write

$$p(x|\pi) = \binom{n}{x} (1-\pi)^n \exp[x \log \pi / (1-\pi)].$$

## 2.11.3 Conjugate densities

When a likelihood function comes from the exponential family, so

$$l(\theta|x) \propto h(\theta)^n \exp\left\{\sum t(x_i)\psi(\theta)\right\}$$

there is an unambiguous definition of a conjugate family – it is defined to be the family  $\Pi$  of densities such that

$$p(\theta) \propto h(\theta) \exp\{\tau\psi(\theta)\}.$$

This definition does fit in with the particular cases we have discussed before. For example, if  $x$  has a normal distribution  $N(\theta, \phi)$  with unknown mean but known variance, the conjugate family as defined here consists of densities such that  $p(\theta) \propto \{\exp(-\frac{1}{2}\theta^2/\phi)\}^\nu \exp(\tau\theta/\phi)$ .

If we set  $\tau = \nu\theta_0$ , we see that

$$\begin{aligned} p(\theta) &\propto \exp\{-\frac{1}{2}\nu(\theta^2 - 2\theta\theta_0)/\phi\} \\ &\propto (2\pi\phi)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\theta - \theta_0)^2/(\phi/\nu)\right\} \end{aligned}$$

which is a normal  $N(\theta_0, \phi/\nu)$  density. Although the notation is slightly different, the end result is the same as the one we obtained earlier.

#### 2.11.4 Two-parameter exponential family

The one-parameter exponential family, as its name implies, only includes densities with one unknown parameter (and not even all of those which we shall encounter). There are a few cases in which we have two unknown parameters, most notably when the mean and variance of a normal distribution are both unknown, which will be considered in detail in Section 2.12. It is this situation which prompts us to consider a generalization. A density is from the two-parameter exponential family if it is of the form  $p(x|\theta, \phi) = g(x)h(\theta, \phi)\exp\{t(x)\psi(\theta, \phi) + u(x)\chi(\theta, \phi)\}$

or equivalently if, given  $n$  independent observations  $x = (x_1, x_2, \dots, x_n)$ , the likelihood takes the form  $l(\theta, \phi|x) \propto h(\theta, \phi)^n \exp\left\{\sum t(x_i)\psi(\theta, \phi) + \sum u(x_i)\chi(\theta, \phi)\right\}$ .

Evidently the two-dimensional vector  $(\sum t(x_i), \sum u(x_i))$  is sufficient for the two-dimensional vector  $(\theta, \phi)$  of parameters given  $X$ . The family of densities conjugate to such a likelihood takes the form  $p(\theta, \phi) \propto h(\theta, \phi)^n \exp\{\tau\psi(\theta, \phi) + v\chi(\theta, \phi)\}$ .

While the case of the normal distribution with both parameters unknown is of considerable theoretical and practical importance, there will not be many other two-parameter families we shall encounter. The idea of the exponential family can easily be extended to a  $k$ -parameter exponential family in an obvious way, but there will be no need for more than two parameters in this book.

## 2.12 Normal mean and variance both unknown

### 2.12.1 Formulation of the problem

It is much more realistic to suppose that both parameters of a normal distribution are unknown rather than just one. So we consider the case where we have a set of observations  $x = (x_1, x_2, \dots, x_n)$  which are  $N(\theta, \phi)$  with  $\theta$  and  $\phi$  both unknown.

$$p(x|\theta, \phi) = (2\pi\phi)^{-\frac{1}{2}} \exp\{-\frac{1}{2}(x - \theta)^2/\phi\}$$

Clearly,  $= \{(2\pi)^{-\frac{1}{2}}\}\{\phi^{-\frac{1}{2}}\} \exp(-\frac{1}{2}\theta^2/\phi) \exp(x\theta/\phi - \frac{1}{2}x^2/\phi)$

from which it follows that the density is in the two-parameter exponential family as defined above. Further

$$\begin{aligned}
l(\theta, \phi|x) &\propto p(x|\theta, \phi) \\
&= (2\pi\phi)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(x_1 - \theta)^2/\phi\right\} \\
&\quad \times \cdots \times (2\pi\phi)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(x_n - \theta)^2/\phi\right\} \\
&\propto \phi^{-n/2} \exp\left[-\frac{1}{2}\left\{\sum(x_i - \theta)^2 + n(\bar{x} - \theta)^2\right\}/\phi\right] \\
&= \phi^{-n/2} \exp[-\frac{1}{2}\{S + n(\bar{x} - \theta)^2\}/\phi]
\end{aligned}$$

wherein we define

$$S = \sum(x_i - \bar{x})^2$$

(rather than as  $\sum(x_i - \mu)^2$  as in the case where the mean is known to be equal to  $\mu$ ). It is also convenient to define  $s^2 = S/(n - 1)$

(rather than  $s^2 = S/n$  as in the case where the mean is known).

It is worth noting that the two-dimensional vector  $(\bar{x}, S)$  or equivalently  $(\bar{x}, s^2)$  is clearly sufficient for  $(\theta, \phi)$  given  $X$ .

Because this case can get quite complicated, we shall first consider the case of an indifference or ‘reference’ prior. It is usual to take  $p(\theta, \phi) \propto 1/\phi$  which is the product of the reference prior  $p(\theta) \propto 1$  for  $\theta$  and the reference prior  $p(\phi) \propto 1/\phi$  for  $\phi$ . The justification for this is that it seems unlikely that if you knew very little about either the mean or the variance, then being given information about the one would affect your judgements about the other. (Other possible priors will be discussed later.) If we do take this reference prior, then  $p(\theta, \phi|x) \propto \phi^{-n/2-1} \exp[-\frac{1}{2}\{S + n(\bar{x} - \theta)^2\}/\phi]$ .

For reasons which will appear later, it is convenient to set

$$v = n - 1$$

in the power of  $\phi$ , but not in the exponential, so that

$$p(\theta, \phi|x) \propto \phi^{-(v+1)/2-1} \exp[-\frac{1}{2}\{S + n(\bar{x} - \theta)^2\}/\phi].$$

## 2.12.2 Marginal distribution of the mean

Now in many real problems what interests us is the mean  $\theta$ , and  $\phi$  is what is referred to as a *nuisance parameter*. In classical (sampling theory) statistics, nuisance parameters can be a real nuisance, but there is (at least in principle) no problem from a Bayesian viewpoint. All we need to do is to find the *marginal* (posterior) distribution of  $\theta$ , and you should recall from Section 1.4 on ‘Several

$$p(\theta|x) = \int p(\theta, \phi|x) d\phi$$

$$\propto \int_0^\infty \phi^{-(v+1)/2-1} \exp[-\frac{1}{2}\{S + n(\bar{x} - \theta)^2\}/\phi] d\phi.$$

Random Variables' that

This integral is not too bad – all you need to do is to substitute

$$x = \frac{1}{2}A/\phi,$$

where

$$A = \{S + n(\bar{x} - \theta)^2\},$$

and it reduces to a standard gamma function integral

$$\left( \int_0^\infty x^{(v+1)/2-1} \exp(-x) dx \right) / A^{(v+1)/2}.$$

It follows that

$$p(\theta|x) \propto \{S + n(\bar{x} - \theta)^2\}^{-(v+1)/2}$$

which is the required posterior distribution of  $\theta$ . However, this is not the most convenient way to express the result. It is usual to define  $t = \frac{\theta - \bar{x}}{s/\sqrt{n}}$ ,

where (as defined earlier)  $s^2 = S/(n-1) = S/\nu$ . Because the Jacobian  $|d\theta/dt|$  of the transformation from  $\theta$  to  $t$  is a constant, the posterior density of  $t$  is given by

$$p(t|x) \propto \{v s^2 + (st)^2\}^{-(v+1)/2}$$

$$\propto \{1 + t^2/v\}^{-(v+1)/2}.$$

A glance at Appendix A will show that this is the density of a random variable with Student's distribution on  $v$  degrees of freedom, so that we can write  $t \sim t_v$ . The fact that the distribution of  $t$  depends on the single parameter  $v$  makes it sensible to express the result in terms of this distribution rather than that of  $\theta$  itself, which depends on  $\bar{x}$  and  $S$  as well as on  $v$ , and is consequently more complicated to tabulate. Note that as  $v \rightarrow \infty$  the standard exponential limit shows that the density of  $t$  is ultimately proportional to  $\exp(-\frac{1}{2}t^2)$ , which is the standard normal form. On the other hand, if  $v = 1$ , we see that  $t$  has a standard Cauchy distribution  $C(0, 1)$ , or equivalently that  $\theta \sim C(\bar{x}, s^2/n)$ .

Because the density of Student's  $t$  is symmetric about the origin, an HDR is also symmetric about the origin, and so can be found simply from a table of percentage points.

## 2.12.3 Example of the posterior density for the mean

Consider the data on uterine weight of rats introduced earlier in Section 2.8 on ‘HDRs for the Normal Variance.’ With those data,  $n=20$ ,  $\sum x_i = 420$ , and

$$S = \sum (x_i - \bar{x})^2 = \sum x_i^2 - \left( \sum x_i \right)^2 / n = 664,$$

$$\sum x_i^2 = 9484, \text{ so that } \bar{x} = 21.0 \text{ and } s/\sqrt{n} = \sqrt{664/(19 \times 20)} = 1.32.$$

We can deduce that the posterior distribution of the true mean  $\theta$  is given by

$$\frac{\theta - 21}{1.32} \sim t_{19}.$$

In principle, this tells us all we can deduce from the data if we have no very definite prior knowledge. It can help to understand what this means by looking for highest density regions. From tables of the t distribution the value exceeded by  $t_{19}$  with probability 0.025 is  $t_{19,0.025} = 2.093$ . It follows that a 95% HDR for  $\theta$  is  $21.0 \pm 1.32 \times 2.093$ , that is the interval (18, 24).

## 2.12.4 Marginal distribution of the variance

If we require knowledge about  $\phi$  rather than  $\theta$ , we use

$$\begin{aligned} p(\phi|x) &= \int p(\theta, \phi|x) d\theta \\ &= \int_{-\infty}^{\infty} \phi^{-(\nu+1)/2-1} \exp[-\frac{1}{2}\{S + n(\theta - \bar{x})^2\}/\phi] d\theta \\ &\propto \phi^{-\nu/2-1} \exp(-\frac{1}{2}S/\phi) \int_{-\infty}^{\infty} (2\pi\phi/n)^{-\frac{1}{2}} \exp\{-\frac{1}{2}n(\theta - \bar{x})^2/\phi\} d\theta \\ &= \phi^{-\nu/2-1} \exp(-\frac{1}{2}S/\phi) \end{aligned}$$

as the last integral is that of a normal density. It follows that the posterior density of the variance is  $S\chi_{\nu}^{-2}$ . Except that  $n$  is replaced by  $\nu = n - 1$  the conclusion is the same as in the case where the mean is known. Similar considerations to those which arose when the mean was known make it preferable to use HDRs based on log chi-squared, though with a different number of degrees of freedom.

## 2.12.5 Example of the posterior density of the variance

With the same data as before, if the mean is not known (which in real life it almost certainly would not be), the posterior density for the variance  $\phi$  is  $664\chi_{19}^{-2}$ . Some idea of the meaning of this can be got from looking for a 95% HDR.

Because values of  $\chi^2_{19}$  corresponding to an HDR for  $\log \chi^2_{19}$  are found from the tables in the Appendix to be 9.267 and 33.921, a 95% HDR lies between 664/33.921 and 664/9.267, that is the interval (20, 72). It may be worth noting that this does not differ all that much from the interval (19, 67) which we found on the assumption that the mean was known.

## 2.12.6 Conditional density of the mean for given variance

We will find it useful in Section 2.13 to write the posterior in the form

$$p(\theta, \phi | \mathbf{x}) = p(\phi | \mathbf{x}) p(\theta | \phi, \mathbf{x}).$$

Since

$$p(\theta, \phi | \mathbf{x}) \propto \phi^{-(v+1)/2-1} \exp[-\frac{1}{2}\{S + n(\bar{x} - \theta)^2\}/\phi],$$

$$p(\phi | \mathbf{x}) = \phi^{-v/2-1} \exp(-\frac{1}{2}S/\phi)$$

this implies that

$$p(\theta | \phi, \mathbf{x}) \propto \phi^{-\frac{1}{2}} \exp\{-\frac{1}{2}n(\theta - \bar{x})^2/\phi\}$$

which as the density integrates to unity implies that

$$p(\theta | \phi, \mathbf{x}) = (2\pi\phi/n)^{-\frac{1}{2}} \exp\{-\frac{1}{2}n(\theta - \bar{x})^2/\phi\}$$

that is that, for given  $\phi$  and  $X$ , the distribution of the mean  $\theta$  is  $N(\bar{x}, \phi/n)$ . This is the result we might have expected from our investigations of the case where the variance is known, although this time we have arrived at the result from conditioning on the variance in the case where neither parameter is truly known.

A distribution for the two-dimensional vector  $(\theta, \phi)$  of this form, in which  $\phi$  has (a multiple of an) inverse chi-squared distribution and, for given  $\phi$ ,  $\theta$  has a normal distribution, will be referred to as a *normal/chi-squared distribution*, although it is more commonly referred to as normal gamma or normal inverse gamma. (The chi-squared distribution is used to avoid unnecessary complications.) It is possible to try to look at the joint posterior density of  $\theta$  and  $\phi$ , but two-dimensional distributions can be hard to visualize in the absence of independence, although numerical techniques can help. Some idea of an approach to this can be got from Box and Tiao (1992, Section 2.4).

## 2.13 Conjugate joint prior for the normal distribution

In Section 2.12, we considered a reference prior for a normal distribution with both parameters unknown, whereas in this section we shall consider a conjugate prior for this situation. It is, in fact, rather difficult to determine *which* member of the conjugate family to use when substantial prior information is available, and hence in practice the reference prior is often used in the hope that the likelihood dominates the prior. It is also the case that the manipulations necessary to deal with the conjugate prior are a bit involved, although the end results are, of course, similar to those when we use a reference prior, with some of the parameters altered slightly. Part of the problem is the unavoidable notational complexity. Further, the notation is not agreed among the different writers on the subject. A new notation is introduced below.

We first recall that the likelihood is

$$l(\theta, \phi | x) \propto p(x|\theta, \phi) \propto \phi^{-n/2} \exp\left[-\frac{1}{2}\{S + n(\theta - \bar{x})^2\}/\phi\right].$$

Now suppose that your prior distribution of  $\phi$  is (a multiple of) an inverse chi-squared on  $v_0$  degrees of freedom. It may be convenient to think of  $v_0$  as  $l_0 - 1$ , so that your prior knowledge about  $\phi$  is in some sense worth  $l_0$  observations. Thus,  $p(\phi) \propto \phi^{-v_0/2-1} \exp(-\frac{1}{2}S_0/\phi)$ .

Now suppose that, conditional on  $\phi$ , your prior distribution for  $\theta$  is normal of mean  $\theta_0$  and variance  $\phi/n_0$ , so that your prior knowledge is worth  $n_0$  observations of variance  $\phi$  or their equivalent. It is not necessarily the case that  $n_0 = l_0$ . Then  $p(\theta|\phi) = (2\pi\phi/n_0)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\theta - \theta_0)^2/(\phi/n_0)\right\}$ .

Thus, the joint prior is a case of a *normal/chi-squared distribution*, which was referred to briefly at the end of the last section. Its joint density is  $p(\theta, \phi) = p(\phi)p(\theta|\phi)$

$$\begin{aligned} &\propto \phi^{-(v_0+1)/2-1} \exp\left[-\frac{1}{2}\{S_0 + n_0(\theta - \theta_0)^2\}/\phi\right] \\ &= \phi^{-(v_0+1)/2-1} \exp\left\{-\frac{1}{2}Q_0(\theta)/\phi\right\}, \end{aligned}$$

where  $Q_0(\theta)$  is the quadratic

$$Q_0(\theta) = n_0\theta^2 - 2(n_0\theta_0)\theta + (n_0\theta_0^2 + S_0).$$

It should be clear that by suitable choice of the parameters  $n_0$ ,  $\theta_0$  and  $S_0$ , the quadratic  $Q_0(\theta)$  can be made into any non-negative definite quadratic form.

We note that the marginal prior density for  $\theta$  is

$$p(\theta) \propto \{S_0 + n_0(\theta - \theta_0)^2\}^{-(v_0+1)/2}$$

(cf. Section 2.12), so that

$$\frac{\theta - \theta_0}{\sqrt{S_0/n_0 v_0}} \sim t_{v_0}$$

and it follows from Appendix A on that the unconditional prior mean and variance of  $\theta$  are  $\theta_0$  and  $S_0/n_0(v_0 - 2)$ .

By taking

$$v_0 = -1 \text{ (i.e. } l_0 = 0) \quad n_0 = 0, \quad \text{and} \quad S_0 = 0$$

(so that the quadratic vanishes identically, that is  $Q_0(\theta) \equiv 0$ ) we get the reference prior  $p(\theta, \phi) \propto 1/\phi$ .

It should be noted that if  $n_0 \neq 0$  then  $p(\theta, \phi)$  is not a product of a function of  $\theta$  and  $\phi$ , so that  $\theta$  and  $\phi$  are not independent a priori. This does not mean that it is impossible to use priors other than the reference prior in which the mean and the variance are independent a priori, but that such a prior will not be in the conjugate family, so that the posterior distribution will be complicated and it may need a lot of numerical investigation to find its properties.

We shall deal with priors for  $\theta$  and  $\phi$  which are independent when we come to consider numerical methods in Chapter 9.

## 2.13.2 Derivation of the posterior

If the prior is of a normal/chi-squared form, then the posterior is

$$\begin{aligned} p(\theta, \phi | \mathbf{x}) &\propto p(\theta, \phi) l(\theta, \phi | \mathbf{x}) \\ &\propto \phi^{-(v_0+n+1)/2-1} \\ &\quad \times \exp[-\frac{1}{2}\{(S_0 + S) + n_0(\theta - \theta_0)^2 + n(\theta - \bar{x})^2\}/\phi] \\ &= \phi^{-(v_1+1)/2-1} \exp\{-\frac{1}{2}Q_1(\theta)/\phi\}, \end{aligned}$$

where

$$v_1 = v_0 + n$$

and  $Q_1(\theta)$  is another quadratic in  $\theta$ , namely,

$$\begin{aligned} Q_1(\theta) &= (S_0 + S) + n_0(\theta - \theta_0)^2 + n(\theta - \bar{x})^2 \\ &= (n_0 + n)\theta^2 - 2(n_0\theta_0 + n\bar{x})\theta \\ &\quad + (n_0\theta_0^2 + n\bar{x}^2 + S_0 + S) \end{aligned}$$

which is in the form in which the prior was expressed, that is

$$\begin{aligned} Q_1(\theta) &= S_1 + n_1(\theta - \theta_1)^2 \\ &= n_1\theta^2 - 2(n_1\theta_1)\theta + (n_1\theta_1^2 + S_1) \end{aligned}$$

if we define

$$\begin{aligned}
n_1 &= n_0 + n \\
\theta_1 &= (n_0\theta_0 + n\bar{x})/n_1 \\
S_1 &= S_0 + S + n_0\theta_0^2 + n\bar{x}^2 - n_1\theta_1^2 \\
&= S_0 + S + (n_0^{-1} + n^{-1})^{-1}(\theta_0 - \bar{x})^2.
\end{aligned}$$

(The second formula for  $S_1$  follows from the first after a little manipulation – its importance is that it is less subject to rounding errors.) This result has finally vindicated the claim that if the prior is normal/chi-squared, then so is the posterior, so that the normal/chi-squared family is conjugate to the normal likelihood with both mean and variance unknown. Thus, the posterior for  $\phi$  is  $\phi \sim S_1 \chi_{v_1}^{-2}$

and that for  $\theta$  given  $\phi$  is

$$\theta | \phi \sim N(\theta_1, \phi/n).$$

Clearly, we can adapt the argument used when we considered the reference prior to find the marginal distribution for  $\theta$ . Thus, the posterior distribution of  $t = \frac{\theta - \theta_1}{S_1/\sqrt{n_1}}$ ,

where

$$s_1^2 = S_1/v_1$$

is a Student's t distribution on  $v_1$  degrees of freedom, that is  $t \sim t_{v_1}$ .

It follows that if you use a conjugate prior, then your inferences should proceed as with the reference prior *except* that you have to replace  $v$  by  $v_1$ ,  $S$  by  $S_1$ ,  $n$  by  $n_1$  and  $\bar{x}$  by  $\theta_1$ .

### 2.13.3 Example

An experimental station has had experience with growing wheat which leads it to believe that the yield per plot is more or less normally distributed with mean 100 and standard deviation 10. The station then wished to investigate the effect of a growth hormone on the yield per plot. In the absence of any other information, the prior distribution for the variance on the plots might be taken to have mean 300 and standard deviation 160. As for the mean, it is expected to be about 110, and this information is thought to be worth about 15 observations. To fit a prior of a normal/chi-squared form first equate the mean and variance of (a multiple of) an inverse chi-squared distribution to 300 and 160<sup>2</sup>, so that

$$S_0/(v_0 - 2) = 300, \quad \text{so} \quad 2S_0^2/(v_0 - 2)^2 = 180000, \quad \frac{2S_0^2}{(v_0 - 2)^2(v_0 - 4)} = 25600$$

from which  $v_0 - 4 = 7$  and hence  $v_0 = 11$  and  $S_0 = 2700$ . The other information gives  $\theta_0 = 110$  and  $n_0 = 15$ .

Twelve plots treated with the hormone gave the following yields:  
 141, 102, 73, 171, 137, 91, 81, 157, 146, 69, 121, 134,  
 so that  $n = 12$ ,  $\sum x_i = 1423$ ,  $\sum x_i^2 = 181789$ , and so  $\bar{x} = 119$ ,  
 $S = \sum (x_i - \bar{x})^2 = \sum x_i^2 - (\sum x_i)^2 / n = 13045$ .

Using the rounded values found earlier, the parameters of the posterior come to

$$\nu_1 = \nu_0 + n = 23,$$

$$n_1 = n_0 + n = 27,$$

$$\theta_1 = (n_0\theta_0 + n\bar{x})/n_1 = 114,$$

$$S_1 = S_0 + S + (n_0^{-1} + n^{-1})^{-1}(\theta_0 - \bar{x})^2 = 16285,$$

$$s/\sqrt{n_1} = \sqrt{16285/(23 \times 27)} = 5.1.$$

It follows that a posteriori

$$\frac{\theta - \theta_1}{s/\sqrt{n_1}} \sim t_{\nu_1}$$

$$\phi \sim S_1 \chi_{\nu_1}^{-2}.$$

In particular,  $\phi$  is somewhere near  $S_1/\nu_1 = 708$  [actually the exact mean of  $\phi$  is  $S_1/(\nu_1 - 2) = 775$ ]. Using tables of on 23 degrees of freedom, a 95% HDR for  $\theta$  is  $\theta_1 \pm 2.069s/\sqrt{n_1}$ , that is the interval (103, 125). The chi-squared distribution can also be approximated by a normal distribution (see Appendix A).

This example will be considered further in Chapter 9.

## 2.13.4 Concluding remarks

While Bayesian techniques are, in principle, just as applicable when there are two or even more unknown parameters as when there is only one unknown parameter, the practical problems are considerably increased. The computational problems can be quite severe if the prior is not from the conjugate family, but even more importantly it is difficult to convince yourself that you have specified the prior to your satisfaction. In the case of the normal distribution, the fact that if the prior is taken from the conjugate family the mean and variance are not usually independent makes it quite difficult to understand the nature of the assumption you are making. Of course, the more data you have, the less the prior matters and hence some of the difficulties become less important. Fuller consideration will be given to numerical methods in Chapter 9.

## 2.14 Exercises on Chapter 2

1. Suppose that  $k \sim B(n, \pi)$ . Find the standardized likelihood as a function of  $\pi$  for given  $k$ . Which of the distributions listed in Appendix A does this represent?

2. Suppose we are given the 12 observations from a normal distribution:

15.644, 16.437, 17.287, 14.448, 15.308, 15.169,

18.123, 17.635, 17.259, 16.311, 15.390, 17.252,

and we are told that the variance  $\phi = 1$ . Find a 90% HDR for the posterior distribution of the mean assuming the usual reference prior.

3. With the same data as in the previous question, what is the predictive distribution for a possible future observation  $x$ ?

4. A random sample of size  $n$  is to be taken from an  $N(\theta, \phi)$  distribution where  $\phi$  is known. How large must  $n$  be to reduce the posterior variance of  $\phi$  to the fraction  $\phi/k$  of its original value (where  $k > 1$ )?

5. Your prior beliefs about a quantity  $\theta$  are such that

$$p(\theta) = \begin{cases} 1 & (\theta \geq 0) \\ 0 & (\theta < 0). \end{cases}$$

A random sample of size 25 is taken from an  $N(\theta, 1)$  distribution and the mean of the observations is observed to be 0.33. Find a 95% HDR for  $\theta$ .

6. Suppose that you have prior beliefs about an unknown quantity  $\theta$  which can be approximated by an  $N(\lambda, \phi)$  distribution, while my beliefs can be approximated by an  $N(\mu, \psi)$  distribution. Suppose further that the reasons that have led us to these conclusions do not overlap with one another. What distribution should represent our beliefs about  $\theta$  when we take into account all the information available to both of us?

7. Prove the theorem quoted without proof in Section 2.4.

8. Under what circumstances can a likelihood arising from a distribution in the exponential family be expressed in data translated form?

9. Suppose that you are interested in investigating how variable the performance of schoolchildren on a new mathematics test, and that you begin by trying this test out on children in 12 similar schools. It turns out that the average standard deviation is about 10 marks. You then want to try the test on a thirteenth school, which is fairly similar to those you have already investigated, and you reckon that the data on the other schools gives

you a prior for the variance in this new school which has a mean of 100 and is worth eight direct observations on the school. What is the posterior distribution for the variance if you then observe a sample of size 30 from the school of which the standard deviation is 13.2? Give an interval in which the variance lies with 90% posterior probability.

**10.** The following are the dried weights of a number of plants (in g) from a batch of seeds:

$$4.17, 5.58, 5.18, 6.11, 4.50, 4.61, 5.17, 4.53, 5.33, 5.14.$$

Give 90% HDRs for the mean and variance of the population from which they come.

**11.** Find a sufficient statistic for  $\mu$  given an  $n$ -sample  $x = (x_1, x_2, \dots, x_n)$  from the exponential distribution  $p(x|\mu) = \mu^{-1} \exp(-x/\mu)$  ( $0 < x < \infty$ )

where the parameter  $\mu$  can take any value in  $0 < \mu < \infty$ .

**12.** Find a (two-dimensional) sufficient statistic for  $(\alpha, \beta)$  given an  $n$ -sample  $x = (x_1, x_2, \dots, x_n)$  from the two-parameter gamma distribution  $p(x|\alpha, \beta) = \{\beta^\alpha \Gamma(\alpha)\}^{-1} x^{\alpha-1} \exp(-x/\beta)$  ( $0 < x < \infty$ )

where the parameters  $\alpha$  and  $\beta$  can take any values in  $0 < \alpha < \infty, 0 < \beta < \infty$ .

**13.** Find a family of conjugate priors for the likelihood  $l(\beta|x) = p(x|\alpha, \beta)$ , where  $p(x|\alpha, \beta)$  is as in the previous question, but  $\alpha$  is known.

**14.** Show that the tangent of a random angle (i.e. one which is uniformly distributed on  $[0, 2\pi]$ ) has a Cauchy distribution  $C(0, 1)$ .

**15.** Suppose that the vector  $x = (x, y, z)$  has a trinomial distribution depending on the index  $n$  and the parameter  $\pi = (\pi, \rho, \sigma)$ , where  $\pi + \rho + \sigma = 1$

$$p(x|\pi) = \frac{n!}{x! y! z!} \pi^x \rho^y \sigma^z \quad (x + y + z = n).$$

Show that this distribution is in the two-parameter exponential family.

**16.** Suppose that the results of a certain test are known, on the basis of general theory, to be normally distributed about the same mean  $\mu$  with the same variance  $\phi$ , neither of which is known. Suppose further that your prior beliefs about  $(\mu, \phi)$  can be represented by a normal/chi-squared distribution with  $v_0 = 4, S_0 = 350, n_0 = 1$  and  $\theta_0 = 85$ .

Now suppose that 100 observations are obtained from the population with mean 89 and sample variance  $s^2=30$ . Find the posterior distribution of  $(\mu, \phi)$ . Compare 50% prior and posterior HDRs for  $\mu$ .

**17.** Suppose that your prior for  $\theta$  is a  $\frac{2}{3} : \frac{1}{3}$  mixture of  $N(0, 1)$  and  $N(1, 1)$  and that a single observation  $x \sim N(\theta, 1)$  turns out to equal 2. What is your

posterior probability that  $\theta > 1$ ?

**18.** Establish the formula

$$(n_0^{-1} + n^{-1})^{-1}(\bar{x} - \theta_0)^2 = n\bar{x}^2 + n_0\theta_0^2 - n_1\theta_1^2,$$

where  $n_1 = n_0 + n$  and  $\theta_1 = (n_0\theta_0 + n\bar{x})/n_1$ , which was quoted in Section 2.13 as providing a formula for the parameter  $S_1$  of the posterior distribution in the case where both mean and variance are unknown which is less susceptible to rounding errors.

# 3

## Some other common distributions

### 3.1 The binomial distribution

#### 3.1.1 Conjugate prior

In this section, the parameter of interest is the probability  $\pi$  of success in a number of trials which can result in success ( $S$ ) or failure ( $F$ ), the trials being independent of one another and having the same probability of success. Suppose that there is a fixed number of  $n$  trials, so that you have an observation  $x$  (the number of successes) such that  $x \sim B(n, \pi)$

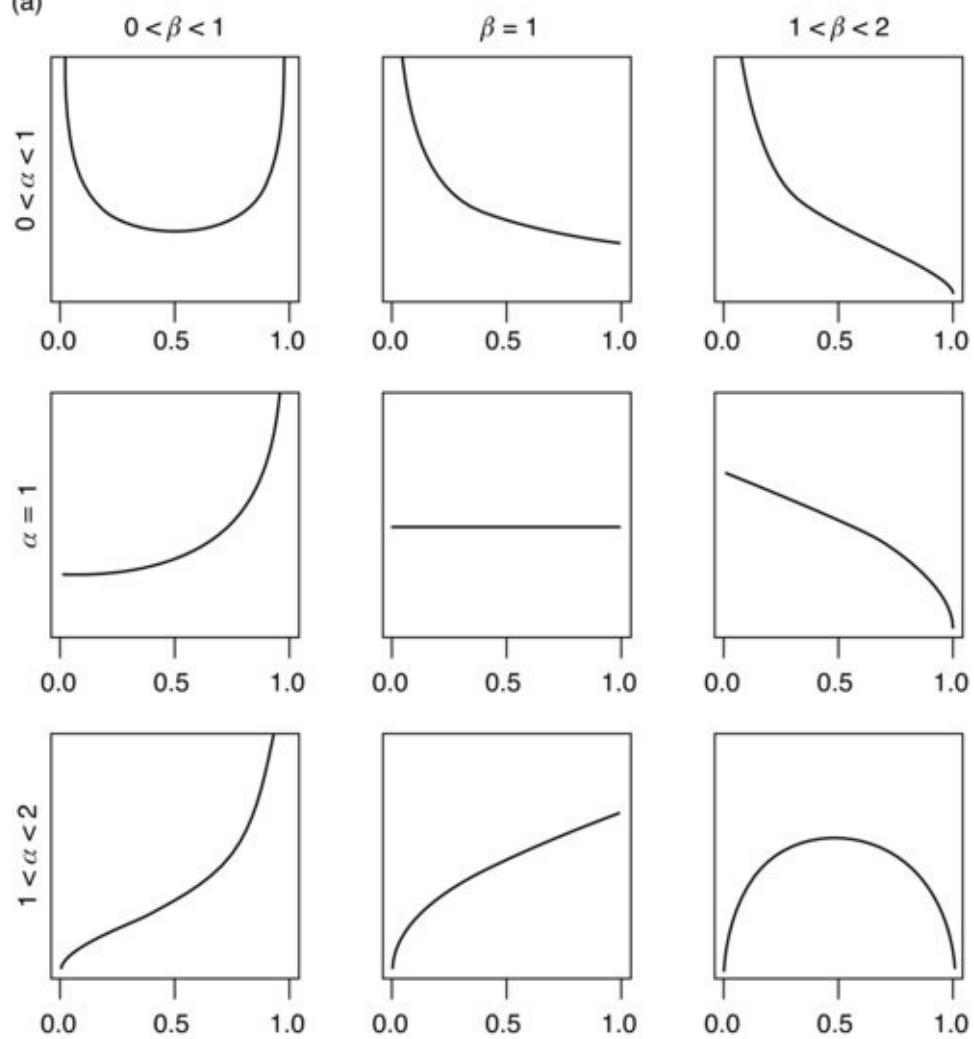
from a binomial distribution of index  $n$  and parameter  $\pi$ , and so

$$p(x|\pi) = \binom{n}{x} \pi^x (1-\pi)^{n-x} \quad (x = 0, 1, \dots, n)$$
$$\propto \pi^x (1-\pi)^{n-x}.$$

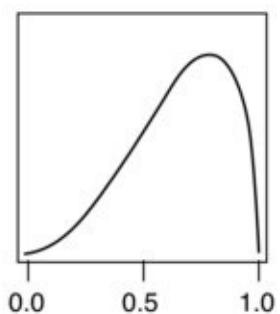
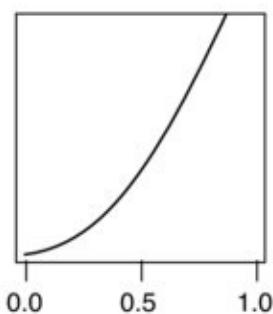
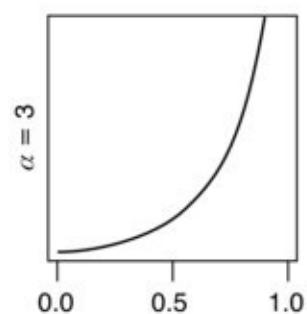
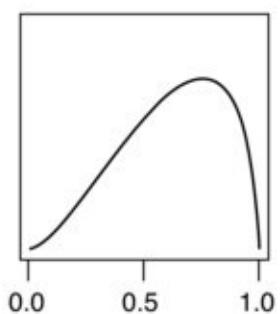
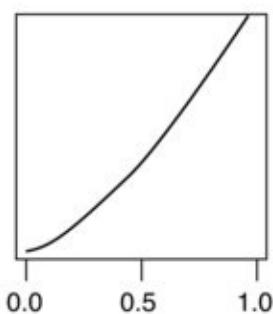
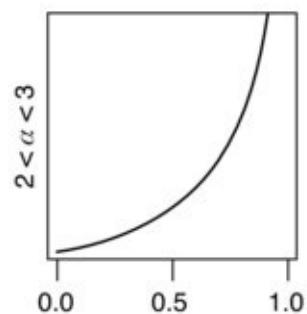
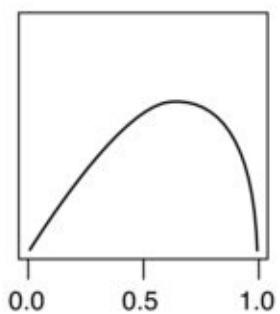
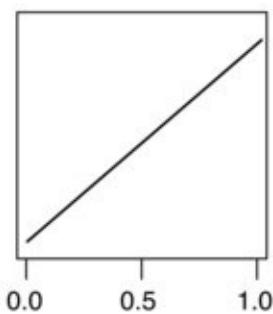
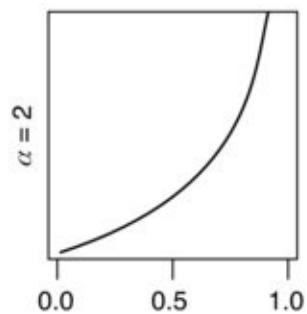
The binomial distribution was introduced in Section 1.3 on ‘Random Variables’ and its properties are of course summarized in Appendix A.

**Figure 3.1** Examples of beta densities.

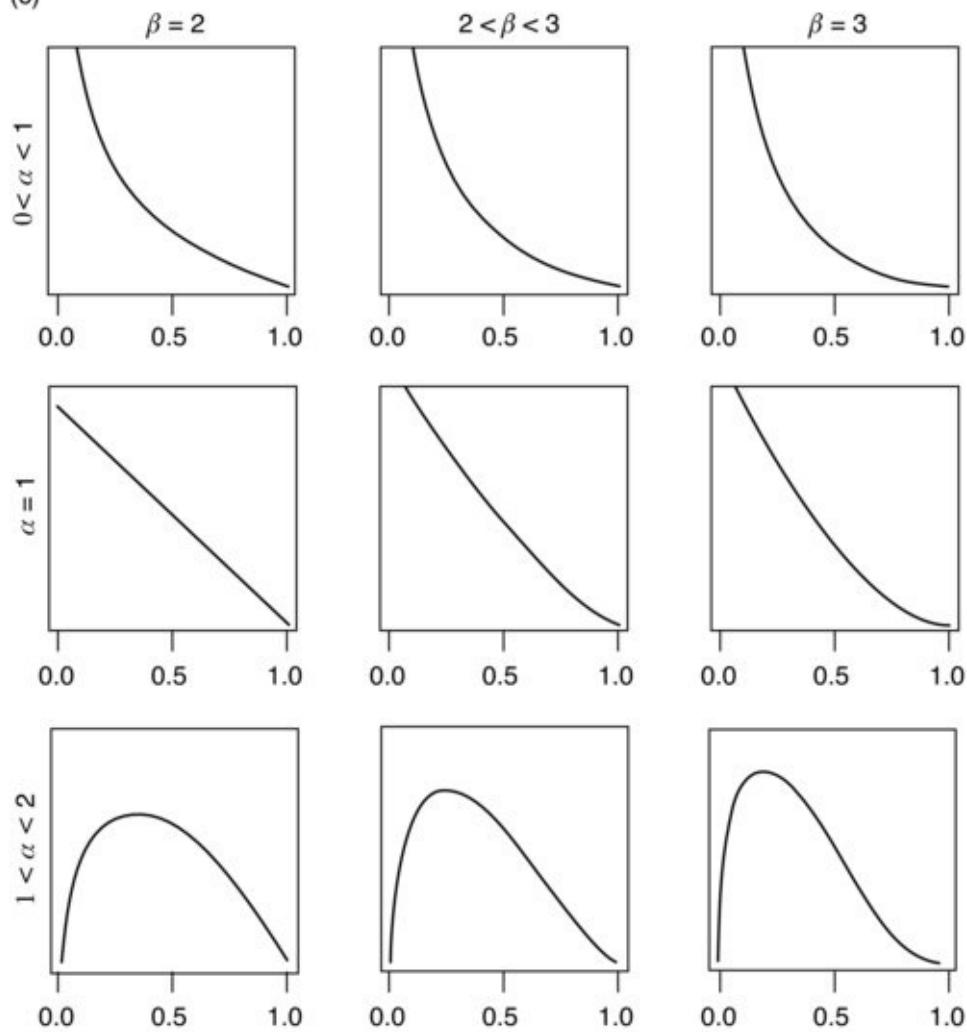
(a)



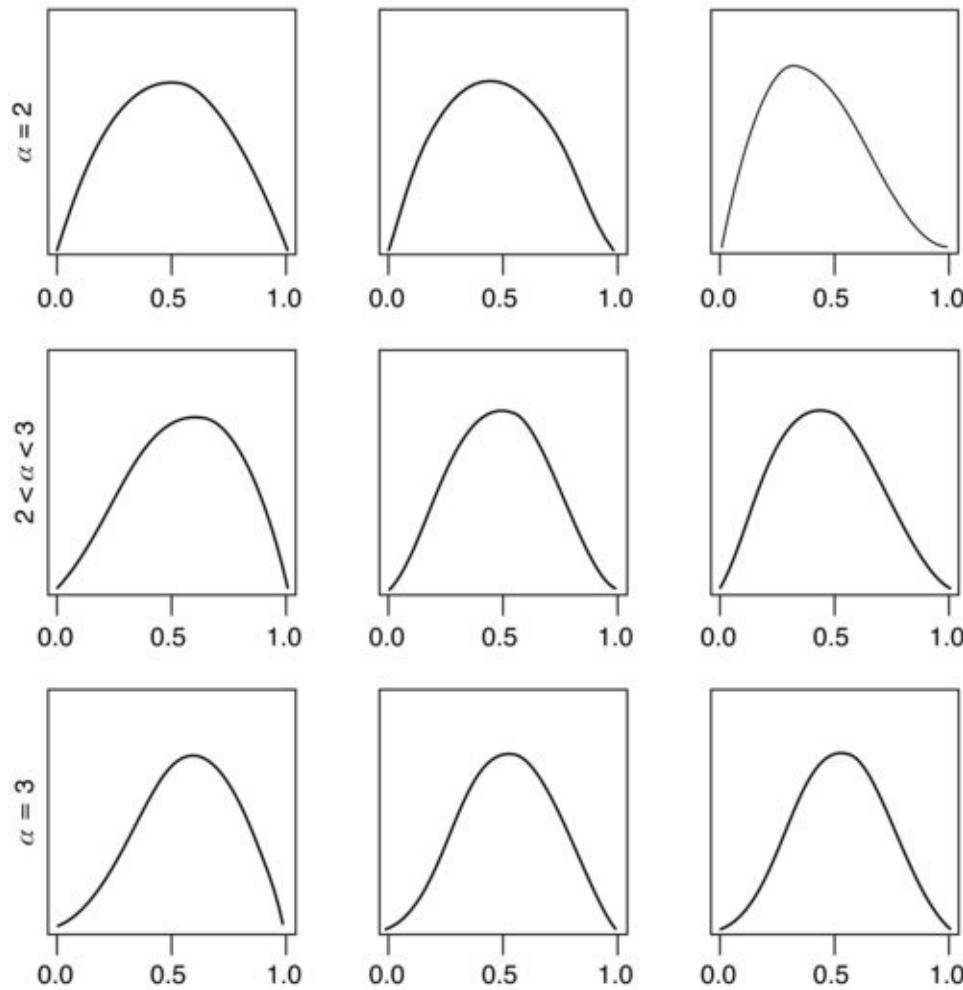
(b)



(c)



(d)



If your prior for  $\pi$  has the form

$$p(\pi) \propto \pi^{\alpha-1}(1-\pi)^{\beta-1} \quad (0 \leq \pi \leq 1)$$

that is, if

$$\pi \sim \text{Be}(\alpha, \beta)$$

has a beta distribution (which is also described in the same Appendix), then the posterior evidently has the form

$$p(\pi|x) \propto \pi^{\alpha+x-1}(1-\pi)^{\beta+n-x-1}$$

that is

$$\pi|x \sim \text{Be}(\alpha + x, \beta + n - x).$$

It is immediately clear that the family of beta distributions is conjugate to a binomial likelihood.

The family of beta distributions is illustrated in [Figure 3.1](#). Basically, any reasonably smooth unimodal distribution on  $[0, 1]$  is likely to be reasonably well approximated by some beta distribution, so that it is very often possible to

approximate your prior beliefs by a member of the conjugate family, with all the simplifications that this implies. In identifying an appropriate member of the family, it is often useful to equate the mean  $E\pi = \alpha/(\alpha + \beta)$  of  $\text{Be}(\alpha, \beta)$  to a value which represents your belief and  $\alpha + \beta$  to a number which in some sense represents the number of observations which you reckon your prior information to be worth. (It is arguable that it should be  $\alpha + \beta + 1$  or  $\alpha + \beta + 2$  that should equal this number, but in practice this will make no real difference). Alternatively, you could equate the mean to a value which represents your beliefs about the location of  $\pi$  and the variance  $V\pi = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$  of  $\text{Be}(\alpha, \beta)$  to a value which represents how spread out your beliefs are.

### 3.1.2 Odds and log-odds

We sometimes find it convenient to work in terms of the odds on success against failure, defined by

$$\lambda = \pi/(1 - \pi),$$

so that  $\pi = \lambda/(1 + \lambda)$ . One reason for this is the relationship mentioned in Appendix A that if  $\pi \sim \text{Be}(\alpha, \beta)$  then  $\frac{\beta}{\alpha}\lambda = \frac{\beta\pi}{\alpha(1 - \pi)} \sim F_{2\alpha, 2\beta}$  has Snedecor's F distribution. Moreover, the log-odds

$$\Lambda = \log \lambda = \log\{\pi/(1 - \pi)\}$$

is close to having Fisher's z distribution; more precisely

$$\frac{1}{2}\Lambda + \frac{1}{2}\log(\beta/\alpha) \sim z_{2\alpha, 2\beta}.$$

It is then easy to deduce from the properties of the z distribution given in Appendix A that

$$E\Lambda \cong \log\{(\alpha - \frac{1}{2})/(\beta - \frac{1}{2})\}$$

$$V\Lambda \cong \alpha^{-1} + \beta^{-1}.$$

One reason why it is useful to consider the odds and log-odds is that tables of the F and z distributions are more readily available than tables of the beta distribution.

### 3.1.3 Highest density regions

Tables of HDRs of the beta distribution are available [see, Novick and Jackson (1974, Table A.15) or Isaacs *et al.* (1974, Table 43)], but it is not necessary or particularly desirable to use them. (The reason is related to the reason for not

using HDRs for the inverse chi-squared distribution as such.) In Section 3.2, we shall discuss the choice of a reference prior for the unknown parameter  $\pi$ . It turns out that there are several possible candidates for this honour, but there is at least a reasonably strong case for using a prior  $p(\pi) \propto \pi^{-1}(1-\pi)^{-1}$ .

Using the usual change-of-variable rule  $p(\Lambda) \propto p(\pi)|d\pi/d\Lambda|$ , it is easily seen that this implies a uniform prior  $p(\Lambda) \propto 1$

in the log-odds  $\Lambda$ . As argued in Section 2.8 on ‘HDRs for the normal variance’, this would seem to be an argument in favour of using an interval in which the *posterior distribution of*  $\Lambda = \log \lambda$  is higher than anywhere outside. The Appendix includes tables of values of  $F$  corresponding to HDRs for  $\log F$ , and the distribution of  $\Lambda$  as deduced earlier is clearly very nearly that of  $\log F$ . Hence in seeking for, for example, a 90% interval for  $\pi$  when  $\pi \sim \text{Be}(\alpha, \beta)$ , we should first look up values  $\underline{F}$  and  $\bar{F}$  corresponding to a 90% HDR for  $\log F_{2\alpha, 2\beta}$ . Then a suitable interval for values of the odds  $\lambda$  is given by  $\underline{F} \leq \beta\lambda/\alpha \leq \bar{F}$

from which it follows that a suitable interval of values of  $\pi$  is

$$\frac{\alpha\underline{F}}{\beta + \alpha\underline{F}} \leq \pi \leq \frac{\alpha\bar{F}}{\beta + \alpha\bar{F}}.$$

If the tables were going to be used solely for this purpose, they could be better arranged to avoid some of the arithmetic involved at this stage, but as they are used for other purposes and do take a lot of space, the minimal extra arithmetic is justifiable.

Although this is not the reason for using these tables, a helpful thing about them is that we need not tabulate values of  $\underline{F}$  and  $\bar{F}$  for  $\beta > \alpha$ . This is because if  $F$  has an  $F_{\alpha, \beta}$  distribution then  $F^{-1}$  has an  $F_{\beta, \alpha}$  distribution. It follows that if an HDR for  $\log F$  is  $(\log \underline{F}, \log \bar{F})$  then an HDR for  $\log F^{-1}$  is  $(-\log \bar{F}, -\log \underline{F})$ , and so if  $(\alpha, \beta)$  is replaced by  $(\beta, \alpha)$  then the interval  $(\underline{F}, \bar{F})$  is simply replaced by  $(1/\bar{F}, 1/\underline{F})$ . There is no such simple relationship in tables of HDRs for  $F$  itself or in tables of HDRs for the beta distribution.

### 3.1.4 Example

It is my guess that about 20% of the best known (printable) limericks have the same word at the end of the last line as at the end of the first. However, I am not very sure about this, so I would say that my prior information was only ‘worth’ some nine observations. If I seek a conjugate prior to represent my beliefs, I

$$\alpha/(\alpha + \beta) = 0.20$$

need to take  $\alpha + \beta = 9$ .

These equations imply that  $\alpha = 1.8$  and  $\beta = 7.2$ . There is no particular reason to restrict  $\alpha$  and  $\beta$  to integer values, but on the other hand prior information is rarely very precise, so it seems simpler to take  $\alpha = 2$  and  $\beta = 7$ . Having made these conjectures, I then looked at one of my favourite books of light verse, Silcock (1952), and found that it included 12 limericks, of which two (both by Lear) have the same word at the ends of the first and last lines. This leads me to a posterior which is  $\text{Be}(4, 17)$ . I can obtain some idea of what this distribution is like by looking for a 90% HDR. From interpolation in the tables in the Appendix, values of  $F$  corresponding to a 90% HDR for  $\log F_{34,8}$  are  $\underline{F} = 0.42$  and  $\bar{F} = 2.85$ . It follows that an appropriate interval of values of  $F_{8,34}$  is  $(1/\bar{F}, 1/\underline{F})$ , that is  $(0.35, 2.38)$ , so that an appropriate interval for  $\pi$  is

$$\frac{4 \times 0.35}{17 + 4 \times 0.35} \leq \pi \leq \frac{4 \times 2.38}{17 + 4 \times 2.38}$$

that is  $(0.08, 0.36)$ .

If for some reason, we want HDRs for  $\pi$  itself, instead of for  $\Lambda = \log \lambda$  and insist on using HDRs for  $\pi$  itself, then we can use the tables quoted earlier [namely, Novick and Jackson (1974, Table A.15) or Isaacs *et al.* (1974, Table 43)]. Alternatively, Novick and Jackson (1974, Section 5.5), point out that a reasonable approximation can be obtained by finding the median of the posterior distribution and looking for a 90% interval such that the probability of being between the lower bound and the median is 45% and the probability of being between the median and the upper bound is 45%. The usefulness of this procedure lies in the ease with which it can be followed using tables of the percentage points of the beta distribution alone, should tables of HDRs be unavailable. It can even be used in connection with the nomogram which constitutes Table 17 of Pearson and Hartley (ed.) (1966), although the accuracy resulting leaves something to be desired. On the whole, the use of the tables of values of  $F$  corresponding to HDRs for  $\log F$ , as described earlier, seems preferable.

### 3.1.5 Predictive distribution

The posterior distribution is clearly of the form  $\text{Be}(\alpha, \beta)$  for some  $\alpha$  and  $\beta$  (which, of course, include  $x$  and  $n-x$ , respectively), so that the predictive distribution of the next observation  $y \sim \text{B}(m, \pi)$  after we have the single observation  $x$  on top of our previous background information is

$$\begin{aligned}
p(y|x) &= \int p(y|\pi)p(\pi|x) d\pi \\
&= \int \binom{m}{y} \pi^y (1-\pi)^{m-y} B(\alpha, \beta)^{-1} \pi^{\alpha-1} (1-\pi)^{\beta-1} d\pi \\
&= \binom{m}{y} B(\alpha, \beta)^{-1} B(\alpha + y, \beta + m - y).
\end{aligned}$$

This distribution is known as the *beta-binomial distribution*, or sometimes as the *Pólya distribution* [see Calvin, 1984]. We shall not have a great deal of use for it in this book, although we will refer to it briefly in Chapter 9. It is considered, for example, in Raiffa and Schlaifer (1961, Section 7.11). We shall encounter a related distribution, the *beta-Pascal distribution* in Section 7.3 when we consider informative stopping rules.

## 3.2 Reference prior for the binomial likelihood

### 3.2.1 Bayes' postulate

The Rev. Thomas Bayes himself in Bayes (1763) put forward arguments in favour of a uniform prior

$$p(\pi) = \begin{cases} 1 & (0 < \pi < 1) \\ 0 & (\text{otherwise}) \end{cases}$$

(which, unlike the choice of a prior uniform over  $-\infty < \theta < \infty$ , is a proper density in that it integrates to unity) as the appropriate one to use when we are ‘completely ignorant’. This choice of prior has long been known as *Bayes’ postulate*, as distinct from his theorem. The same prior was used by Laplace (1774). It is a member of the conjugate family, to wit  $\text{Be}(1, 1)$ .

Bayes’ arguments are quite intricate, and still repay study. Nevertheless, he seems to have had some doubts about the validity of the postulate, and these doubts appear to have been partly responsible for the fact that his paper was not published in his lifetime, but rather communicated posthumously by his friend Richard Price.

The postulate seems intuitively reasonable, in that it seems to treat all values on a level and thus reflect the fact that you so no reason for preferring any one value to any other. However, you should not be too hasty in endorsing it because ignorance about the value of  $\pi$  presumably implies ignorance about the value of any function of  $\pi$ , and yet when the change of variable rule is used a uniform prior for  $\pi$  will not usually imply a uniform prior for any function of  $\pi$ .

One possible argument for it is as follows. A ‘natural’ estimator for the parameter  $\pi$  of a binomial distribution of index  $n$  is the observed proportion  $x/n$  of successes, and it might seem a sensible estimator to use when we have no prior information. It is in fact the *maximum likelihood* estimator, that is, the value of  $\pi$  for which the likelihood  $l(\pi|x) \propto \pi^x(1-\pi)^{n-x}$  is a maximum. In classical or sampling theory statistics it is also commended for various reasons which do not usually carry much weight with Bayesians, for example that it is *unbiased*, that is,  $E(x/n) = \pi$  (the expectation being taken over repeated sampling) whatever the value of  $\pi$  is. Indeed, it is not hard to show that it is a *minimum variance unbiased estimator (MVUE)*.

Now if you have a  $Be(\alpha, \beta)$  prior and so get a posterior which is  $Be(\alpha + x, \beta + n - x)$ , it might seem natural to say that a good estimator for  $\pi$  would be obtained by finding that value at which the posterior density is a maximum, that is, the posterior mode. This procedure is clearly related to the idea of maximum likelihood. Since the posterior mode occurs at  $(\alpha + x - 1)/(\alpha + \beta + n - 2)$  as is easily checked by differentiation, this posterior mode coincides with  $x/n$  if and only if  $\alpha = \beta = 1$ , that is, the prior is uniform.

Jeffreys (1961, Section 3.1) argued that ‘Again, is there not a preponderence at the extremes. Certainly if we take the Bayes-Laplace rule right up to the extremes we are lead to results that do not correspond to anybody’s way of thinking.’

### 3.2.2 Haldane’s prior

Another suggestion, due to Haldane (1931), is to use a  $Be(0, 0)$  prior, which has density

$$p(\pi) \propto \pi^{-1}(1-\pi)^{-1}$$

which is an improper density and is equivalent (by the usual change of variable argument) to a prior uniform in the log-odds

$$\Lambda = \log\{\pi/(1-\pi)\}.$$

An argument for this prior based on the ‘naturalness’ of the estimator  $x/n$  when  $x \sim B(n, \pi)$  is that the mean of the posterior distribution for  $\pi$ , namely,  $Be(\alpha + x, \beta + n - x)$ , is  $(\alpha + x)/(\alpha + \beta + n)$ ,

which coincides with  $x/n$  if and only if  $\alpha = \beta = 0$ . (There is a connection here with the classical notion of the unbiasedness of  $x/n$ .) Another argument that has

been used for this prior is that since any observation always increases either  $\alpha$  or  $\beta$ , it corresponds to the greatest possible ignorance to take  $\alpha$  and  $\beta$  as small as possible. For a beta density to be proper (i.e. to have a finite integral and so be normalizable, so that its integral is unity) it is necessary and sufficient that  $\alpha$  and  $\beta$  should both be strictly greater than 0. This can then be taken as an indication that the right reference prior is  $\text{Be}(0, 0)$ .

A point against this choice of prior is that if we have one observation with probability of success  $\pi$ , then use of this prior results in a posterior which is  $\text{Be}(1, 0)$  if that observation is a success and  $\text{Be}(0, 1)$  if it is a failure. However, a  $\text{Be}(1, 0)$  distribution gives infinitely more weight to values near 1 than to values away from 1, and so it would seem that a sample with just one success in it would lead us to conclude that all future observations will result in successes, which seems unreasonable on the basis of so small an amount of evidence.

### 3.2.3 The arc-sine distribution

A possible compromise between  $\text{Be}(1, 1)$  and  $\text{Be}(0, 0)$  is  $\text{Be}(\frac{1}{2}, \frac{1}{2})$ , that is, the (proper) density  $p(\pi) \sim \pi^{-\frac{1}{2}}(1 - \pi)^{-\frac{1}{2}}$ .

This distribution is sometimes called the *arc-sine distribution* (cf. Feller 1968, 1; III.4). In Section 3.3, we will see that a general principle known as *Jeffreys' rule* suggests that this is the correct reference prior to use. However, Jeffreys' rule is a guideline which cannot be followed blindly, so that in itself does not settle the matter.

The  $\text{Be}(\frac{1}{2}, \frac{1}{2})$  prior can easily be shown (by the usual change-of-variable rule  $p(\psi) = p(\pi)|d\pi/d\psi|$ ) to imply a uniform prior for  $z = \sin^{-1}\sqrt{\pi}$ .

This transformation is related to the transformation of the data when

$$x \sim \text{B}(n, \pi)$$

in which  $z$  is defined by

$$z = \sin^{-1}\sqrt{(x/n)}.$$

This transformation was first introduced in 1.5 on ‘Means and Variances’, where we saw that it results in the approximations

$$\mathbb{E}z \cong \sin^{-1}\sqrt{\pi} = \psi,$$

$$\mathcal{V}z \cong 1/4n.$$

Indeed it turns out that

$$z \approx \mathbf{N}(\psi, 1/4n)$$

where the symbol  $\approx$  means ‘is approximately distributed as’ (see Section 3.10 on ‘Approximations based on the Likelihood’). To the extent that this is so, it

follows that the transformation  $\psi = \psi(\pi)$  puts the likelihood in data translated form, and hence that a uniform prior in  $\psi$ , that is, a  $\text{Be}(\frac{1}{2}, \frac{1}{2})$  prior for  $\pi$ , is an appropriate reference prior.

### 3.2.4 Conclusion

The three aforementioned possibilities are not the only ones that have been suggested. For example, Zellner (1977) suggested the use of a prior  $p(\pi) \propto \pi^\pi (1 - \pi)^{1-\pi}$

[see also the references in Berger (1985, Section 3.3.4)]. However, this is difficult to work with because it is not in the conjugate family.

In fact, the three suggested conjugate priors  $\text{Be}(0, 0)$ ,  $\text{Be}(\frac{1}{2}, \frac{1}{2})$  and  $\text{Be}(1, 1)$  (and for that matter Zellner's prior) do not differ enough to make much difference with even a fairly small amount of data, and the aforementioned discussion on the problem of a suitable reference prior may be too lengthy, except for the fact that the discussion does underline the difficulty in giving a precise meaning to the notion of a prior distribution that represents 'knowing nothing'. It may be worth your while trying a few examples to see how little difference there is between the possible priors in particular cases.

In practice, the use of  $\text{Be}(0, 0)$  is favoured here, although it must be admitted that one reason for this is that it ties in with the use of HDRs found from tables of values of  $F$  based on HDRs for  $\log F$  and hence obviates the need for a separate set of tables for the beta distribution. But in any case, we could use the method based on these tables and the results would not be very different from those based on any other appropriate tables.

## 3.3 Jeffreys' rule

### 3.3.1 Fisher's information

In Section 2.1 on the nature of Bayesian inference, the log-likelihood function was defined as

$$L(\theta|x) = \log l(\theta|x).$$

In this section, we shall sometimes write  $l$  for  $l(\theta|x)$ ,  $L$  for  $L(\theta|x)$  and  $p$  for the probability density function  $p(x|\theta)$ . The fact that the likelihood can be multiplied by any constant implies that the log-likelihood contains an arbitrary additive

constant.

An important concept in classical statistics which arises, for example, in connection with the Cramér-Rao bound for the variance of an unbiased estimator, is that of the information provided by an experiment which was defined by Fisher (1925a) as  $I(\theta|x) = -\mathbb{E} \partial^2(\log p)/\partial\theta^2$ ,

the expectation being taken over all possible values of  $x$  for fixed  $\theta$ . It is important to note that the information depends on the distribution of the data rather than on any particular value of it, so that if we carry out an experiment and observe, for example, that  $\tilde{x} = 3$ , then the information is no different from the information if  $\tilde{x} = 5$ ; basically it is to do with what can be expected from an experiment before rather than after it has been performed. It may be helpful to note that strictly speaking it should be denoted  $I(\theta|\tilde{x})$ .

Because the log-likelihood differs from  $\log p(x|\theta)$  by a constant, all their derivatives are equal, and we can equally well define the information by  $I(\theta|x) = -\mathbb{E} \partial^2 L/\partial\theta^2$ .

It is useful to prove two lemmas. In talking about these, you may find it useful to use a terminology frequently employed by classical statisticians. The first derivative  $\partial L/\partial\theta$  of the log-likelihood is sometimes called the *score*; see Lindgren (1993, Section 4.5.4).

### **Lemma 3.1**

$$\mathbb{E} \partial L/\partial\theta = 0.$$

*Proof.* From the definition

$$\begin{aligned}\mathbb{E} \partial L/\partial\theta &= \int \{\partial(\log l)/\partial\theta\} p \, dx = \int \{\partial(\log p)/\partial\theta\} p \, dx \\ &= \int \{(\partial p/\partial\theta)/p\} p \, dx = \int (\partial p/\partial\theta) \, dx \\ &= \frac{d}{d\theta} \int p \, dx = \frac{d}{d\theta} 1 = 0.\end{aligned}$$

since in any reasonable case it makes no difference whether differentiation with respect to  $\theta$  is carried out inside or outside the integral with respect to  $x$ . ■

### **Lemma 3.2**

$$I(\theta|x) = \mathbb{E}(\partial L/\partial\theta)^2$$

*Proof.* Again differentiating under the integral sign

$$\begin{aligned}
I(\theta|x) &= -\mathbb{E}\partial^2(\log l)/\partial\theta^2 = -\int\{\partial^2(\log p)/\partial\theta^2\}p \, dx \\
&= -\int \frac{\partial}{\partial\theta} \left( \frac{\partial p/\partial\theta}{p} \right) p \, dx \\
&= -\int \left( \frac{\partial^2 p/\partial\theta^2}{p} \right) p \, dx + \int \left( \frac{(\partial p/\partial\theta)^2}{p^2} \right) p \, dx \\
&= -\int (\partial^2 p/\partial\theta^2) \, dx + \int \{\partial(\log p)/\partial\theta\}^2 p \, dx \\
&= -\frac{d^2}{d\theta^2} 1 + \int (\partial L/\partial\theta)^2 p \, dx \\
&= \mathbb{E}(\partial L/\partial\theta)^2
\end{aligned}$$

as required. ■

### 3.3.2 The information from several observations

If we have  $n$  independent observations  $x = (x_1, x_2, \dots, x_n)$ , then the probability densities multiply, so the log-likelihoods add. Consequently, if we define  $I(\theta|x) = -\mathbb{E}\partial^2 L(\theta|x)/\partial\theta^2$ ,

then by linearity of expectation

$$I(\theta|x) = nI(\theta|x)$$

where  $x$  is any one of the  $x_i$ . This accords with the intuitive idea that  $n$  times as many observations should give us  $n$  times as much information about the value of an unknown parameter.

### 3.3.3 Jeffreys' prior

In a Bayesian context, the important thing to note is that if we transform the unknown parameter  $\theta$  to  $\psi = \psi(\theta)$  then  $\frac{\partial\{\log l(\psi|x)\}}{\partial\psi} = \frac{\partial\{\log l(\theta|x)\}}{\partial\theta} \frac{d\theta}{d\psi}$ .

Squaring and taking expectations over values of  $x$  (and noting that  $d\theta/d\psi$  does not depend on  $x$ ), it follows that  $I(\psi|x) = I(\theta|x)(d\theta/d\psi)^2$ .

It follows from this that if a prior density

$$p(\theta) \propto \sqrt{I(\theta|x)}$$

is used, then by the usual change-of-variable rule

$$p(\psi) \propto \sqrt{I(\psi|x)}.$$

It is because of this property that Jeffreys (1961, Section 3.10) suggested that the density

$$p(\theta) \propto \sqrt{I(\theta|x)}$$

provided a suitable reference prior (the use of this prior is sometimes called *Jeffreys' rule*). This rule has the valuable property that the prior is *invariant* in that, whatever scale we choose to measure the unknown parameter in, the same prior results when the scale is transformed to any particular scale. This seems a highly desirable property of a reference prior. In Jeffreys' words, ‘any arbitrariness in the choice of parameters could make no difference to the results’.

### 3.3.4 Examples

*Normal mean.* For the normal mean with known variance, the log-likelihood is

$$L(\theta|x) = -\frac{1}{2}(x - \theta)^2/\phi + \text{constant},$$

so that

$$\partial^2 L / \partial \theta^2 = -1/\phi$$

which does not depend on  $x$ , so that

$$I(\theta|x) = 1/\phi$$

implying that we should take a prior

$$p(\theta) \propto 1/\sqrt{\phi} = \text{constant}$$

which is the rule suggested earlier for a reference prior.

*Normal variance.* In the case of the normal variance

$$L(\phi|x) = -\frac{1}{2} \log \phi - \frac{1}{2}(x - \theta)^2/\phi + \text{constant},$$

so that

$$\partial^2 L / \partial \phi^2 = \frac{1}{2}\phi^{-2} - (x - \theta)^2/\phi^3.$$

Because  $E(x - \theta)^2 = \mathbb{V}x = \phi$ , it follows that

$$I(\phi|x) = -\frac{1}{2}\phi^{-2} + \phi/\phi^3 = \frac{1}{2}\phi^{-2}$$

implying that we should take a prior

$$p(\phi) \propto 1/\phi$$

which again is the rule suggested earlier for a reference prior.

*Binomial parameter.* In this case,

$$L(\pi|x) = x \log \pi + (n - x) \log(1 - \pi) + \text{constant},$$

so that

$$\partial^2 L / \partial \pi^2 = -x/\pi^2 - (n - x)/(1 - \pi)^2.$$

Because  $E\pi = n\pi$ , it follows that

$$I(\pi|x) = n\pi/\pi^2 + (n - n\pi)/(1 - \pi)^2 = n\pi^{-1}(1 - \pi)^{-1}$$

implying that we should take a prior

$$p(\pi) \propto \pi^{-\frac{1}{2}}(1 - \pi)^{-\frac{1}{2}}$$

that is  $\pi \sim \text{Be}(\frac{1}{2}, \frac{1}{2})$ , that is  $\pi$  has an arc-sine distribution, which is one of the

rules suggested earlier as possible choices for the reference prior in this case.

### 3.3.5 Warning

While Jeffreys' rule is suggestive, it cannot be applied blindly. Apart from anything else, the integral defining the information can diverge; it is easily seen to do so for the Cauchy distribution  $C(\theta, 1)$ , for example. It should be thought of as a guideline that is well worth considering, particularly if there is no other obvious way of finding a prior distribution. Generally speaking, it is less useful if there are more unknown parameters than one, although an outline of the generalization to that case is given later for reference.

### 3.3.6 Several unknown parameters

If there are several unknown parameters  $\theta = (\theta_1, \theta_2, \dots, \theta_k)$ , the information  $I(\theta|x)$  provided by a single observation is defined as a *matrix*, the element in row  $i$ , column  $j$ , of which is  $(I(\theta|x))_{i,j} = -\mathbb{E}(\partial^2 L/\partial\theta_i \partial\theta_j)$ .

As in the one parameter case, if there are several observations  $x = (x_1, x_2, \dots, x_n)$ , we get

$$I(\theta|x) = n I(\theta|x).$$

If we transform to new parameters  $\psi = (\psi_1, \psi_2, \dots, \psi_k)$  where  $\psi = \psi(\theta)$ , we see that if  $J$  is the matrix the element in row  $i$ , column  $j$  of which is  $J_{i,j} = \partial\theta_i/\partial\psi_j$  then it is quite easy to see that

$$I(\psi|x) = J I(\theta|x) J^T,$$

where  $J^T$  is the transpose of  $J$ , and hence that the determinant  $\det I$  of the information matrix satisfies  $\det I(\psi|x) = \{\det I(\theta|x)\}(\det J)^2$ .

Because  $\det J$  is the Jacobian determinant, it follows that

$$p(\theta) \propto \sqrt{\det I(\theta|x)}$$

provides an invariant prior for the multi-parameter case.

### 3.3.7 Example

*Normal mean and variance both unknown.* In this case, the log-likelihood is

$$L(\theta, \phi|x) = -\frac{1}{2} \log \phi - \frac{1}{2}(x - \theta)^2/\phi + \text{constant},$$

so that

$$\partial L/\partial\theta = (x - \theta)/\phi \text{ and } \partial L/\partial\phi = -\frac{1}{2}\phi^{-1} + \frac{1}{2}(x - \theta)^2/\phi^2$$

and hence

$$\partial^2 L / \partial \theta^2 = -1/\phi;$$

$$\partial^2 L / \partial \theta \partial \phi = -(x - \theta)/\phi^2;$$

$$\partial^2 L / \partial \phi^2 = \frac{1}{2}\phi^{-2} - (x - \theta)^2/\phi^3.$$

Because  $E(x) = \theta$  and  $E(x - \theta)^2 = Vx = \phi$ , it follows that

$$I(\theta, \phi|x) = \begin{pmatrix} \phi^{-1} & 0 \\ 0 & \frac{1}{2}\phi^{-2} \end{pmatrix}$$

and, so that

$$\det I(\theta, \phi|x) = \frac{1}{2}\phi^{-3}.$$

This implies that we should use the reference prior

$$p(\theta, \phi|x) \propto \phi^{-3/2}.$$

It should be noted that this is not the same as the reference prior recommended earlier for use in this case, namely,

$$p(\theta, \phi|x) \propto \phi^{-1}.$$

However, I would still prefer to use the prior recommended earlier. The invariance argument does not take into account the fact that in most such problems your judgement about the mean would not be affected by anything you were told about the variance or vice versa, and on those grounds it seems reasonable to take a prior which is the product of the reference priors for the mean and the variance separately.

The example underlines the fact that we have to be rather careful about the choice of a prior in multi-parameter cases. It is also worth mentioning that it is very often the case that when there are parameters which can be thought of as representing ‘location’ and ‘scale’, respectively, then it would usually be reasonable to think of these parameters as being independent *a priori*, just as suggested earlier in the normal case.

## 3.4 The Poisson distribution

### 3.4.1 Conjugate prior

A discrete random variable  $x$  is said to have a Poisson distribution of mean  $\lambda$  if it has the density  $p(x|\lambda) = \frac{\lambda^x}{x!} \exp(-\lambda)$ .

This distribution often occurs as a limiting case of the binomial distribution as the index  $n \rightarrow \infty$  and the parameter  $\pi \rightarrow 0$  but their product  $n\pi \rightarrow \lambda$  (see Exercise 6 in Chapter 1). It is thus a useful model for rare events, such as the

number of radioactive decays in a fixed time interval, when we can split the interval into an arbitrarily large number of sub-intervals in any of which a particle might decay, although the probability of a decay in any particular sub-interval is small (though constant).

Suppose that you have  $n$  observations  $x = (x_1, x_2, \dots, x_n)$  from such a distribution, so that the likelihood is  $L(\lambda|x) \propto \lambda^T \exp(-n\lambda)$ , where  $T$  is the sufficient statistic

$$T = \sum x_i.$$

We have already seen in Section 2.10 on ‘Conjugate Prior Distributions’ that the appropriate conjugate density is

$$p(\lambda) \propto \lambda^{v/2-1} \exp\left(-\frac{1}{2}S_0\lambda\right)$$

that is,  $\lambda \sim S_0^{-1} \chi_{v'}^2$ , so that  $\lambda$  is a multiple of a chi-squared random variable. Then the posterior density is  $p(\lambda|x) \propto \lambda^{(v+2T)/2-1} \exp\left\{-\frac{1}{2}(S_0 + 2n)\lambda\right\}$ , that is

$$\lambda | x \sim S_1^{-1} \chi_{v'}^2,$$

where

$$S_1 = S_0 + 2n, \quad v' = v + 2T.$$

### 3.4.2 Reference prior

This is a case where we can try to use Jeffreys’ rule. The log-likelihood resulting from a single observation  $x$  is  $L(\lambda|x) = x \log \lambda - \lambda + \text{constant}$ ,

so that

$$\partial^2 L / \partial \lambda^2 = -x/\lambda^2,$$

and hence,

$$I(\lambda|x) = \lambda/\lambda^2 = 1/\lambda.$$

Consequently Jeffreys’ rule suggests the prior

$$p(\lambda) \propto \lambda^{-\frac{1}{2}},$$

which corresponds to  $v = 1$ ,  $S_0 = 0$  in the conjugate family, and is easily seen to be equivalent to a prior uniform in  $\psi = \sqrt{\lambda}$ . It may be noted that there is a sense in which this is intermediate between a prior uniform in  $\log \lambda$  and one uniform in  $\lambda$  itself, since as  $k \rightarrow 0$

$$\frac{\lambda^k - 1}{k} = \frac{\exp(k \log \lambda) - 1}{k} \rightarrow \log \lambda,$$

so that there is a sense in which the transformation from  $\lambda$  to  $\log \lambda$  can be regarded as a ‘zeroth power’ transformation (cf. Box and Cox, 1964).

On the other hand, it could be argued that  $\lambda$  is a scale parameter between 0 and  $+\infty$  and that the right reference prior should therefore be  $p(\lambda) \propto 1/\lambda$  which is uniform in  $\log \lambda$  and corresponds to  $\nu = 0$ ,  $S_0 = 0$  in the conjugate family. However, the difference this would make in practice would almost always be negligible.

### 3.4.3 Example

The numbers of misprints spotted on the first few pages of an early draft of this book were

3, 4, 2, 1, 2, 3

It seems reasonable that these numbers should constitute a sample from a Poisson distribution of unknown mean  $\lambda$ . If you had no knowledge of my skill as a typist, you might adopt the reference prior uniform in  $\sqrt{\lambda}$  for which  $\nu = 1$ ,  $S_0 = 0$ . Since  $n = 6$ ,  $T = \sum x_i = 15$ ,

your posterior for  $\lambda$  would then be  $S_1^{-1} \chi_{\nu'}^2$ , that is,  $12^{-1} \chi_{31}^2$ . This distribution has mean and variance  $\nu'/S_1 = 2.6$ ,  $2\nu'/S_1^2 = 0.43$ .

Of course, I have some experience of my own skill as a typist, so if I considered these figures, I would have used a prior with a mean of about 3 and variance about 4. (As a matter of fact, subsequent re-readings have caused me to adjust my prior beliefs about  $\lambda$  in an upwards direction!) If then I seek a prior in the conjugate family, I need  $\nu/S_0 = 3$ ,  $2\nu/S_0^2 = 4$ ,

which implies  $\nu = 4.5$  and  $S_0 = 1.5$ . This means that my posterior has  $\nu' = 34.5$ ,  $S_1 = 13.5$  and so has mean and variance  $\nu'/S_1 = 2.6$ ,  $2\nu'/S_1^2 = 0.38$ .

The difference between the two posteriors is not great and of course would become less and less as more data were included in the analysis. It would be easy enough to give HDRs. According to arguments presented in other cases, it would be appropriate to use HDRs for the chi (rather than the chi-squared distribution), but it really would not make much difference if the regions were based on HDRs for chi-squared or on values of chi-squared corresponding to HDRs for log chi-squared.

### 3.4.4 Predictive distribution

Once we know that  $\lambda$  has a posterior distribution

$$p(\lambda) \propto \lambda^{\nu'/2-1} \exp\left(-\frac{1}{2} S_1 \lambda\right)$$

then since

$$p(x|\lambda) \propto (\lambda^x/x!) \exp(-\lambda)$$

it follows that the predictive distribution

$$\begin{aligned} p(x) &= \int p(x|\lambda)p(\lambda) d\lambda \\ &\propto \int_0^\infty (\lambda^x/x!) \exp(-\lambda) \lambda^{v'/2-1} \exp\left(-\frac{1}{2}S_1\lambda\right) d\lambda \\ &\propto \left\{\left(\frac{1}{2}S_1 + 1\right)^{-\frac{1}{2}v'-x-1} / x!\right\} \int_0^\infty z^{\frac{1}{2}v'+x-1} \exp(-z) dz \\ &\propto \left\{\left(\frac{1}{2}S_1 + 1\right)^{-x} / x!\right\} \Gamma\left(x + \frac{1}{2}v'\right) \end{aligned}$$

(dropping a factor which depends on  $v'$  alone). Setting  $\pi = 1 - (\frac{1}{2}S_1 + 1)^{-1}$ , you can find the constant by reference to Appendix A. In fact, at least when  $\frac{1}{2}v'$  is an integer, the predictive distribution is negative binomial, that is  $x \sim \text{NB}\left(\frac{1}{2}v', \pi\right)$ .

Further, although this point is not very important, it is not difficult to see that the negative binomial distribution can be generalized to the case where  $\frac{1}{2}v'$  is not an integer. All we need to do is to replace some factorials by corresponding gamma functions and note that (using the functional equation for the gamma function)  $\Gamma\left(\frac{1}{2}v' + x\right) / \Gamma\left(\frac{1}{2}v'\right) = \left(\frac{1}{2}v' + x - 1\right) \left(\frac{1}{2}v' + x - 2\right) \cdots \left(\frac{1}{2}v' + 1\right) \frac{1}{2}v'$ ,

so that you can write the general binomial coefficient as

$$\binom{\frac{1}{2}v' + x - 1}{x} = \frac{\Gamma(\frac{1}{2}v' + x)}{x! \Gamma(\frac{1}{2}v')}.$$

The negative binomial distribution is usually defined in terms of a sequence of independent trials each of which results in success or failure with the same probabilities  $\pi$  and  $1 - \pi$  (such trials are often called *Bernoulli trials*) and considering the number  $x$  of failures before the  $n$ th success. We will not have much more use for this distribution in this book, but it is interesting to see it turning up here in a rather different context.

## 3.5 The uniform distribution

### 3.5.1 Preliminary definitions

The *support* of a density  $p(x|\theta)$  is defined as the set of values of  $x$  for which it is nonzero. A simple example of a family of densities in which the support depends on the unknown parameter is the family of uniform distributions (defined later). While problems involving the uniform distribution do not arise all that often in practice, it is worth while seeing what complications can arise in cases where the

support does depend on the unknown parameter.

It is useful to begin with a few definitions. The *indicator function* of any set  $A$  is defined by  $I_A(x) = \begin{cases} 1 & (x \in A) \\ 0 & (x \notin A). \end{cases}$

This is sometimes called the *characteristic function* of the set  $A$  in some other branches of mathematics, but not in probability and statistics (where the term characteristic function is applied to the Fourier–Stieltjes transform of the distribution function).

We say that  $y$  has a *Pareto distribution* with parameters  $\xi$  and  $\gamma$  and write  $y \sim \text{Pa}(\xi, \gamma)$

if it has density

$$p(y|\xi, \gamma) = \begin{cases} \gamma \xi^\gamma y^{-\gamma-1} & (y > \xi) \\ 0 & (y \leq \xi) \end{cases}$$

$$\propto y^{-\gamma-1} I_{(\xi, \infty)}(y).$$

This distribution is often used as a model for distributions of income. A survey of its properties and applications can be found in Arnold (1983).

We say that  $x$  has a *uniform distribution* (or a *rectangular distribution*) on  $(\alpha, \beta)$  and write  $x \sim \text{U}(\alpha, \beta)$

if it has density

$$p(x|\alpha, \beta) \propto \begin{cases} (\beta - \alpha)^{-1} & (\alpha < x < \beta) \\ 0 & (\text{otherwise}) \end{cases}$$

$$\propto (\beta - \alpha)^{-1} I_{(\alpha, \beta)}(x),$$

so that all values in the interval  $(\alpha, \beta)$  are equally likely.

### 3.5.2 Uniform distribution with a fixed lower endpoint

Now suppose we have  $n$  independent observations  $x = (x_1, x_2, \dots, x_n)$  such that  $x_i \sim \text{U}(0, \theta)$

for each  $i$ , where  $\theta$  is a single unknown parameter. Then

$$p(x|\theta) = \begin{cases} \theta^{-n} & (0 < x_i < \theta \text{ for all } i) \\ 0 & (\text{otherwise}). \end{cases}$$

It is now easy to see that we can write the likelihood as

$$l(\theta|x) = \begin{cases} \theta^{-n} & (\theta > x_i \text{ for all } i) \\ 0 & (\text{otherwise}). \end{cases}$$

Defining

$$M = \max\{x_1, x_2, \dots, x_n\}$$

it is clear that

$$l(\theta|x) = \begin{cases} \theta^{-n} & (\theta > M) \\ 0 & (\text{otherwise}), \end{cases}$$

$$\propto \theta^{-n} I_{(M,\infty)}(\theta).$$

Because the likelihood depends on the data through  $M$  alone, it follows that  $M$  is sufficient for  $\theta$  given  $x$ .

It is now possible to see that the Pareto distribution provides the conjugate prior for the above likelihood. For if  $\theta$  has prior  $p(\theta) \propto \theta^{-\gamma-1} I_{(\xi,\infty)}(\theta)$  then the posterior is

$$\begin{aligned} p(\theta|x) &\propto p(\theta) l(\theta|x) \\ &\propto \theta^{-\gamma-1} I_{(\xi,\infty)}(\theta) \theta^{-n} I_{(M,\infty)}(\theta) \\ &\propto \theta^{-(\gamma+n)-1} I_{(\xi,\infty)}(\theta) I_{(M,\infty)}(\theta). \end{aligned}$$

If now we write

$$\gamma' = \gamma + n$$

$$\xi' = \max\{\xi, M\},$$

so that  $\theta > \xi'$  if and only if  $\theta > \xi$  and  $\theta > M$ , and hence  $I_{(\xi',\infty)}(\theta) = I_{(\xi,\infty)}(\theta) I_{(M,\infty)}(\theta)$  we see that

$$p(\theta|x) \propto \theta^{-\gamma'-1} I_{(\xi',\infty)}(\theta).$$

It follows that if the prior is  $\text{Pa}(\xi, \gamma)$  then the posterior is  $\text{Pa}(\xi', \gamma')$  and hence that the Pareto distribution does indeed provide the conjugate family. We should note that neither the uniform nor the Pareto distribution falls into the exponential family, so that we are not here employing the unambiguous definition of conjugacy given in Section 2.11 on ‘The exponential family’. Although this means that the cautionary remarks of Diaconis and Ylvisaker (1979 and 1985) (quoted in Section 2.10 on conjugate prior distributions) apply, there is no doubt of the ‘naturalness’ of the Pareto distribution in this context.

### 3.5.3 The general uniform distribution

The case where both parameters of a uniform distribution are unknown is less important, but it can be dealt with similarly. In this case, it turns out that an appropriate family of conjugate prior distributions is given by the *bilateral bivariate Pareto distribution*. We say that the ordered pair  $(y, z)$  has such a distribution and write  $(y, z) \sim \text{Pabb}(\xi, \eta, \gamma)$ .

if the joint density is

$$p(y, z | \xi, \eta, \gamma) = \begin{cases} \gamma(\gamma + 1)(\xi - \eta)^{\gamma}(z - y)^{-\gamma-2} & (y < \eta < \xi < z) \\ 0 & (\text{otherwise}) \end{cases}$$

$$\propto (z - y)^{-\gamma-2} I_{(\xi, \infty)}(z) I_{(-\infty, \eta)}(y).$$

Now suppose we have  $n$  independent observations  $x = (x_1, x_2, \dots, x_n)$  such that

$$x_i \sim U(\alpha, \beta)$$

where  $\alpha$  and  $\beta$  are unknown. Then

$$p(x | \alpha, \beta) = \begin{cases} (\beta - \alpha)^{-n} & (\alpha < x_i < \beta \text{ for all } i) \\ 0 & (\text{otherwise}). \end{cases}$$

Defining

$$M = \max\{x_1, x_2, \dots, x_n\}$$

$$m = \min\{x_1, x_2, \dots, x_n\}$$

it is clear that the likelihood  $l(\alpha, \beta | x)$  can be written as

$$l(\alpha, \beta | x) = (\beta - \alpha)^{-n} I_{(M, \infty)}(\beta) I_{(-\infty, m)}(\alpha).$$

Because the likelihood depends on the data through  $m$  and  $M$  alone, it follows that  $(m, M)$  is sufficient for  $(\alpha, \beta)$  given  $x$ .

It is now possible to see that the bilateral bivariate Pareto distribution provides the conjugate prior for the aforementioned likelihood. For if  $(\alpha, \beta)$  has prior  $p(\alpha, \beta) \propto (\beta - \alpha)^{-\gamma-2} I_{(\xi, \infty)}(\beta) I_{(-\infty, \eta)}(\alpha)$

then the posterior is

$$\begin{aligned} p(\alpha, \beta | x) &\propto p(\alpha, \beta) l(\alpha, \beta | x) \\ &\propto (\beta - \alpha)^{-(\gamma+n)-2} I_{(\xi, \infty)}(\beta) I_{(M, \infty)}(\beta) I_{(-\infty, \eta)}(\alpha) I_{(-\infty, m)}(\alpha). \end{aligned}$$

If now we write

$$\gamma' = \gamma + n$$

$$\xi' = \max\{\xi, M\}$$

$$\eta' = \min\{\eta, m\},$$

we see that

$$p(\alpha, \beta | x) \propto (\beta - \alpha)^{-\gamma'-2} I_{(\xi', \infty)}(\beta) I_{(-\infty, \eta')}\alpha.$$

It follows that if the prior is  $Pabb(\xi, \eta, \gamma)$  then the posterior is  $Pabb(\xi', \eta', \gamma')$  and hence that the bilateral bivariate Pareto distribution does indeed provide the conjugate prior.

The properties of this and of the ordinary Pareto distribution are, as usual, described in Appendix A.

### 3.5.4 Examples

I realize that the case of the uniform distribution, and in particular the case of a uniform distribution on  $(0, \theta)$ , must be of considerable importance, since it is

considered in virtually all the standard text books on statistics. Strangely, however, none of the standard references seems to be able to find any reasonably plausible practical case in which it arises [with apologies to DeGroot (1970, Section 9.7) if his case really does arise]. In the circumstances, consideration of examples is deferred until Section 3.6, and even then the example considered will be artificial.

## 3.6 Reference prior for the uniform distribution

### 3.6.1 Lower limit of the interval fixed

If  $x = (x_1, x_2, \dots, x_n)$  consists of independent random variables with  $U(0, \theta)$

$$p(x|\theta) = \begin{cases} \theta^{-n} & (\theta > M) \\ 0 & (\text{otherwise}), \end{cases}$$

distributions, then where  $M = \max\{x_1, x_2, \dots, x_n\}$ , so that the likelihood can be written in the form

$$l(\theta|x) = (M/\theta)^n I_{(M,\infty)}(\theta)$$

after multiplying by a constant (as far as  $\theta$  is concerned). Hence,

$$l(\theta|x) = g(\psi(\theta) - t(x))$$

with

$$g(y) = \exp(-ny)I_{(0,\infty)}(y),$$

$$t(x) = \log M,$$

$$\psi(\theta) = \log \theta.$$

It follows that the likelihood is *data translated*, and the general argument about data translated likelihoods in Section 2.5 now suggests that we take a prior which is at least locally uniform in  $\psi = \log \theta$ , that is  $p(\psi) \propto 1$ . In terms of the parameter  $\theta$ , the usual change of variable rule shows that this means  $p(\theta) \propto 1/\theta$ , which is the same prior that is conventionally used for variances. This is not a coincidence, but represents the fact that both are measures of spread (strictly, the standard deviation is more closely analogous to  $\theta$  in this case, but a prior for the variance proportional to the reciprocal of the variance corresponds to a prior for the standard deviation proportional to the reciprocal of the standard deviation).

As the density of  $Pa(\xi, \gamma)$  is proportional to  $y^{-\gamma-1} I_{(M,\infty)}(y)$ , the density proportional to  $1/\theta$  can be regarded as the limit  $Pa(0, 0)$  of  $Pa(\xi, \gamma)$  as  $\xi \rightarrow 0$  and  $\gamma \rightarrow 0$ . Certainly, if the likelihood is  $\theta^{-n} I_{(M,\infty)}(\theta)$ , then the posterior is  $Pa(M, n)$  which is what would be expected when the general rule is applied to the particular case of a  $Pa(0, 0)$  prior.

### 3.6.2 Example

A very artificial example can be obtained by taking groups of random digits from Neave (1978, Table 7.1) ignoring all values greater than some value  $\theta$  [an alternative source of random digits is Lindley and Scott (1995, Table 27)]. A sample of 10 such values is:

0.49487; 0.52802; 0.28667; 0.62058; 0.14704; 0.18519;  
0.17889; 0.14554; 0.29480; 0.46317.

This sample was constructed this sample using the value  $\theta = 0.75$ , but we want to investigate how far this method succeeds in giving information about  $\theta$ , so we note that the posterior is  $\text{Pa}(0.62058, 10)$ . Since the density function of a Pareto distribution  $\text{Pa}(\xi, \gamma)$  decreases monotonically beyond  $\xi$ , an HDR must be of the form  $(\xi, x)$  for some  $x$ , and since the distribution function is (see Appendix A)  $F(x) = [1 - (\xi/x)^\gamma]I_{(\xi, \infty)}(x)$

a 90% HDR for  $\theta$  is  $(0.62058, x)$  where  $x$  is such that  $0.90 = [1 - (0.62058/x)^{10}]$  and so is 0.78126. Thus, a 90% HDR for  $\theta$  is the interval  $(0.62, 0.78)$ . We can see that the true value of  $\theta$  in this artificial example does turn out to lie in the 90% HDR.

### 3.6.3 Both limits unknown

In the two parameter case, when  $x \sim U(\alpha, \beta)$  where  $\alpha < \beta$  are both unknown and  $x$  is any one of the observations, note that it is easily shown that  $\theta = \mathbb{E}x = \frac{1}{2}(\alpha + \beta)$

$$\phi = \mathbb{V}x = (\beta - \alpha)^2/12.$$

Very similar arguments to those used in the case of the normal distribution with mean and variance both unknown in Section 2.12 can now be deployed to suggest independent priors uniform in  $\theta$  and  $\log \phi$ , so that  $p(\theta, \phi) \propto 1/\phi$ .

But

$$\frac{\partial(\theta, \phi)}{\partial(\alpha, \beta)} = \begin{vmatrix} \frac{1}{2} & \frac{1}{2} \\ -(\beta - \alpha)/6 & (\beta - \alpha)/6 \end{vmatrix} \propto (\beta - \alpha),$$

so that this corresponds to

$$p(\alpha, \beta) = p(\theta, \phi) \frac{\partial(\theta, \phi)}{\partial(\alpha, \beta)} \propto \frac{12}{(\beta - \alpha)^2} (\beta - \alpha) \propto 1/(\beta - \alpha).$$

It may be noted that the density of  $\text{Pabb}(\xi, \eta, \gamma)$  is proportional to

$$(z - y)^{-\gamma-2} I_{(\xi, \infty)}(z) I_{(-\infty, \eta)}(y),$$

so that in some sense a density

$$p(\alpha, \beta) \propto (\beta - \alpha)^{-2} I_{(\xi, \infty)}(\beta) I_{(-\infty, \xi)}(\alpha)$$

$$\propto (\beta - \alpha)^{-2} I_{(\alpha, \beta)}(\xi)$$

might be regarded as a limit of  $\text{Pabb}(\xi, \eta, \gamma)$  as  $\eta \rightarrow \xi$  and  $\gamma \rightarrow 0$ . Integrating over a uniform prior for  $\xi$ , which might well seem reasonable, gives

$$p(\alpha, \beta) \propto \int (\beta - \alpha)^{-2} I_{(\alpha, \beta)}(\xi) d\xi = \int_{\alpha}^{\beta} (\beta - \alpha)^{-2} d\xi$$

$$\propto 1/(\beta - \alpha).$$

If the likelihood takes the form

$$l(\alpha, \beta | x) \propto (\beta - \alpha)^{-n} I_{(M, \infty)}(\beta) I_{(-\infty, m)}(\alpha)$$

then the posterior from this prior is  $\text{Pabb}(M, m, n-1)$ . Thus, our reference prior could be regarded as a  $\text{Pabb}(-\infty, \infty, -1)$  distribution, and if we think of it as such, the same formulae as before can be used.

The rule  $p(\alpha, \beta) \propto (\beta - \alpha)^{-2}$  or  $\text{Pabb}(-\infty, \infty, 0)$  corresponds to  $p(\theta, \phi) \propto \phi^{-3/2}$  which is the prior Jeffreys' rule gave us in the normal case with both parameters unknown (see Section 3.3 on Jeffreys' rule).

## 3.7 The tramcar problem

### 3.7.1 The discrete uniform distribution

Occasionally, we encounter problems to do with the discrete uniform distribution. We say that  $x$  has a discrete uniform distribution on  $[\alpha, \beta]$  and write  $x \sim \text{UD}(\alpha, \beta)$

if

$$p(x) = (\beta - \alpha + 1)^{-1} \quad (x = \alpha, \alpha + 1, \alpha + 2, \dots, \beta).$$

One context in which it arises was cited by Jeffreys (1961, Section 4.8). He says,

The following problem was suggested to me several years ago by Professor M. H. A. Newman. A man travelling in a foreign country has to change trains at a junction and goes into the town, of the existence of which he has only just heard. He has no idea of its size. The first thing that he sees is a tramcar numbered 100. What can he infer about the number of tramcars in the town? It may be assumed for the purpose that they are numbered consecutively from 1 upwards.

Clearly, if there are  $\nu$  tramcars in the town and you are equally likely to see any one of the tramcars, then the number  $n$  of the car you observe has a discrete uniform distribution  $\text{UD}(1, \nu)$ . Jeffreys suggests that (assuming  $\nu$  is not too small) we can deal with this problem by analogy with problems involving a

continuous distribution  $U(0, \nu)$ . In the absence of prior information, the arguments of Section 3.7 suggest a reference prior  $p(\nu) \propto 1/\nu$  in the latter case, so his suggestion is that the prior for  $\nu$  in a problem involving a discrete uniform distribution  $UD(1, \nu)$  should be, at least approximately, proportional to  $1/\nu$ . But

$$p(\nu) \propto 1/\nu \quad (\nu = 1, 2, 3, \dots),$$

$$\text{if } p(n|\nu) = 1/\nu \quad (n = 1, 2, \dots, n),$$

then by Bayes' Theorem

$$p(\nu|n) \propto \nu^{-2} \quad (\nu = n, n+1, n+2, \dots).$$

It follows that

$$p(\nu|n) = \nu^{-2} / \left( \sum_{\mu \geq n} \mu^{-2} \right).$$

In particular, the posterior probability that  $\nu \geq \lambda$  is approximately

$$\left( \sum_{\mu \geq \lambda} \mu^{-2} \right) / \left( \sum_{\mu \geq n} \mu^{-2} \right)$$

Approximating the sums by integrals and noting that  $\int \mu^{-2} d\mu = \mu^{-1}$ , this is approximately  $n/\lambda$ . Consequently, the posterior median is  $2n$ , and so 200 if you observed tramcar number 100.<sup>1</sup>

The argument seems rather unconvincing, because it puts quite a lot of weight on the prior as opposed to the likelihood and yet the arguments for the prior are not all that strong, but we may agree with Jeffreys that it may be ‘worth recording’. It is hard to take the reference prior suggested terribly seriously, although if you had a lot more data, then it would not matter what prior you took.

## 3.8 The first digit problem; invariant priors

### 3.8.1 A prior in search of an explanation

The problem we are going to consider in this section is not really one of statistical inference as such. What is introduced here is another argument that can sometimes be taken into account in deriving a prior distribution – that of invariance. To introduce the notion, we consider a population which appears to be invariant in a particular sense.

### 3.8.2 The problem

The problem we are going to consider in this section has a long history going back to Newcomb (1881). Recent references include Knuth (1969, Section 4.2.4B), Raimi (1976) and Turner (1987).

Newcomb's basic observation, in the days where large tables of logarithms were in frequent use, was that the early pages of such tables tended to look dirtier and more worn than the later ones. This appears to suggest that numbers whose logarithms we need to find are more likely to have 1 as their first digit than 9. If you then look up a few tables of physical constants, you can get some idea as to whether this is borne out. For example, *Whitaker's Almanack* (1988, p. 202) quotes the areas of 40 European countries (in square kilometres) as 28 778; 453; 83 849; 30 513; 110 912; 9251; 127 869; 43 069; 1399; 337 032; 547 026; 108 178; 248 577; 6; 131 944; 93 030; 103 000; 70 283; 301 225; 157; 2586; 316; 1; 40 844; 324 219; 312 677; 92 082; 237 500; 61; 504 782; 449 964; 41 293; 23 623; 130 439; 20 768; 78 772; 14 121; 5 571 000; 0.44; 255 804.

The first significant digits of these are distributed as follows:

Digit	1	2	3	4	5	6	7	8	9	Total
Frequency	10	7	6	6	3	2	2	1	3	40
Percentage	25	17.5	15	15	7.5	5	5	2.5	7.5	100

We will see that there are grounds for thinking that the distribution should be approximately as follows:

Digit	1	2	3	4	5	6	7	8	9	Total
Percentage	30	18	12	10	8	7	6	5	5	100

### 3.8.3 A solution

An argument for this distribution runs as follows. The quantities we measure are generally measured in an arbitrary scale, and we would expect that if we measured them in another scale (thus in the case of the aforementioned example, we might measure areas in square miles instead of square kilometres), then the *population* of values (or at least of their first significant figures) would look much the same, although individual values would of course change. This implies that if  $\theta$  is a randomly chosen constant, then for any fixed  $c$  the transformation  $\theta \rightarrow \psi(\theta) = c\theta$

should leave the probability distribution of values of constants alone. This means that if the functional form of the density of values of  $\theta$  is  $p(\theta) = f(\theta)$ , then the corresponding density of values of  $\psi$  will be

$$p(\psi) = f(\psi).$$

Using the usual change-of-variable rule, we know that  $p(\psi) = p(\theta) \cdot |\mathrm{d}\theta/\mathrm{d}\psi|$ , so

that we are entitled to deduce that  $f(c\theta) = f(\psi) = f(\theta)c^{-1}$ .

But if  $f(\theta)$  is any function such that  $f(\theta) = cf(c\theta)$  for all  $c$  and  $\theta$ , then we may take  $c = 1/\theta$  to see that  $f(\theta) = (1/\theta)f(1)$ , so that  $f(\theta) \propto 1/\theta$ . It seems, therefore, that the distribution of constants that are likely to arise in a scientific context should, at least approximately, satisfy  $p(\theta) \propto 1/\theta$ .

Naturally, the reservations expressed in Section 2.5 on locally uniform priors about the use of improper priors as representing genuine prior beliefs over a whole infinite range still apply. But it is possible to regard the prior  $p(\theta) \propto 1/\theta$  for such constants as valid over any interval  $(a, b)$  where  $0 < a < b < \infty$  which is not too large. So consider those constants between  $10^k = a$  and  $10^{k+1} = b$ .

Because

$$\int_a^b d\theta/\theta = \log_e(b/a) = \log_e 10,$$

the prior density for constants  $\theta$  between  $a$  and  $b$  is  $1/(\theta \log_e 10)$

and so the probability that such a constant has first digit  $d$ , that is, that it lies

between  $da$  and  $(d+1)a$ , is  $\int_{da}^{(d+1)a} d\theta/(\theta \log_e 10) = \log_e(1 + d^{-1})/\log_e 10 = \log_{10}(1 + d^{-1})$ .

Since this is true for all values of  $k$ , and any constant lies between  $10^k$  and  $10^{k+1}$  for some  $k$ , it seems reasonable to conclude that the probability that a physical constant has first digit  $d$  is approximately  $\log_{10}(1 + d^{-1})$

which is the density tabulated earlier. This is sometimes known as *Benford's Law* because of the work of Benford (1938) on this problem.

This subsection was headed 'A solution' rather than 'The solution' because a number of other reasons for this density have been adduced. Nevertheless, it is quite an interesting solution. It also leads us into the whole notion of invariant priors.

It has been noted that falsified data is rarely adjusted so as to comply with Benford's Law and this has been proposed as a method of detecting such data in, for example, clinical trials (see Weir and Murray, 2011). Recently Rauch *et al.* (2011) pointed out that deficit data reported to Eurostat by Greece demonstrated that Greek data relevant to the euro deficit criteria showed the greatest deviation from Benford's Law, and that this fact should have given rise to suspicion.

Benford's Law can be related to another empirical law, *Zipf's Law*, originally proposed by Zipf (1935), states that the relative frequency of the  $k$ th most common word in a list of  $n$  words is approximately proportional to  $1/k$ , so that the most frequent word will occur approximately twice as often as the second

most frequent word, three times as often as the third most frequent word, and so on. The relationship between this law and Benford's Law is explored in Pietronero *et al.* (2001).

### 3.8.4 Haar priors

It is sometimes the case that your prior beliefs about a parameter  $\theta$  are in some sense symmetric. Now when a mathematician hears of symmetry, he or she tends immediately to think of groups, and the notions aforementioned generalize very easily to general symmetry groups. If the parameter values  $\theta$  can be thought of as members of an abstract group  $\Theta$ , then the fact that your prior beliefs about  $\theta$  are not altered when the values of  $\theta$  are all multiplied by the same value  $c$  can be expressed by saying that the transformation  $\theta \rightarrow \psi(\theta) = c\theta$

should leave the probability distribution of values of the parameter alone. A density which is unaltered by this operation for arbitrary values of  $c$  is known as a *Haar measure* or, in this context, as a *Haar prior* or an *invariant prior*. Such priors are, in general, unique (at least up to multiplicative constants about which there is an arbitrariness if the priors are improper). This is just the condition used earlier to deduce Benford's Law, except that  $c\theta$  is now to be interpreted in terms of the multiplicative operation of the symmetry group, which will not, in general, be ordinary multiplication.

This gives another argument for a uniform prior for the mean  $\theta$  of a normal distribution  $N(\theta, \phi)$  of known variance, since it might well seem that adding the same constant to all possible values of the mean would leave your prior beliefs unaltered – there seems to be a symmetry under additive operations. If this is so, then the transformation  $\theta \rightarrow \psi(\theta) = c + \theta$

should leave the functional form of the prior density for  $\theta$  unchanged, and it is easy to see that this is the case if and only if  $p(\theta)$  is constant. A similar argument about the multiplicative group might be used about an unknown variance when the mean is known to produce the usual reference prior  $p(\psi) \propto 1/\psi$ . A good discussion of this approach and some references can be found in Berger (1985, Section 3.3.2).

## 3.9 The circular normal distribution

### 3.9.1 Distributions on the circle

In this section, the variable is an angle running from  $0^\circ$  to  $360^\circ$ , that is, from 0 to  $2\pi$  radians. Such variables occur in a number of contexts, for example in connection with the homing ability of birds and in various problems in astronomy and crystallography. Useful references for such problems are Mardia (1972), Mardia and Jupp (2001), and Batschelet (1981). The method used here is a naïve numerical integration technique; for a modern approach using Monte Carlo Markov Chain (MCMC) methods, see Damian and Walker (1999).

The only distribution for such angles which will be considered is the so-called *circular normal* or *von Mises*' distribution. An angle  $\eta$  is said to have such a distribution with mean direction  $\mu$  and concentration parameter  $\kappa$  if

$$p(\eta|\mu, \kappa) = [2\pi I_0(\kappa)]^{-1} \exp\{\kappa \cos(\eta - \mu)\} \quad (0 \leq \eta < 2\pi)$$

and when this is so we write

$$\eta \sim M(\mu, \kappa).$$

The function  $I_0(\kappa)$  is the modified Bessel function of the first kind and order zero, but as far as we are concerned it may as well be regarded as defined by

$$I_0(\kappa) = (2\pi)^{-1} \int_0^{2\pi} \exp\{\kappa \cos(\eta - \mu)\} d\eta.$$

It is tabulated in many standard tables, for example, British Association (1937) or Abramowitz and Stegun (1965, Section 9.7.1). It can be shown that

$$I_0(\kappa) = \sum_{r=0}^{\infty} \left(\frac{1}{2}\kappa\right)^{2r} / (r!)^2.$$

The circular normal distribution was originally introduced by von Mises (1918). It plays a prominent role in statistical inference on the circle and in that context its importance is almost the same as that of the normal distribution on the line. There is a relationship with the normal distribution, since as  $\kappa \rightarrow \infty$  the distribution of  $\sqrt{\kappa}(\eta - \mu)$

approaches the standard normal form  $N(0, 1)$  and hence  $M(\mu, \kappa)$  is approximately  $N(\mu, 1/\kappa)$ . It follows that the concentration parameter is analogous to the precision of a normal distribution. This is related to the fact that asymptotically for large  $\kappa$

$$I_0(\kappa) \sim (2\pi\kappa)^{-\frac{1}{2}} \exp(\kappa).$$

However, the equivalent of the Central Limit Theorem does not result in convergence to the circular normal distribution. Further, the circular normal distribution is not in the exponential family. It should not be confused with the so-called wrapped normal distribution.

The likelihood of  $n$  observations  $\eta = (\eta_1, \eta_2, \dots, \eta_n)$  from an  $M(\mu, \kappa)$

$$l(\mu, \kappa | \eta) \propto \{I_0(\kappa)\}^{-n} \exp \left\{ \sum \kappa \cos(\eta_i - \mu) \right\}$$

distribution is

$$= \{I_0(\kappa)\}^{-n} \exp \left\{ \kappa \cos \mu \sum \cos \eta_i + \kappa \sin \mu \sum \sin \eta_i \right\},$$

so that if we define

$$c = n^{-1} \sum \cos \eta_i$$

$$s = n^{-1} \sum \sin \eta_i$$

then  $(c, s)$  is sufficient for  $(\mu, \kappa)$  given  $\eta$ , and indeed  
 $l(\mu, \kappa | \eta) \propto \{I_0(\kappa)\}^{-n} \exp \{n\kappa c \cos \mu + n\kappa s \sin \mu\}$ .

If we define

$$\rho = \sqrt{(c^2 + s^2)}$$

$$\hat{\mu} = \tan^{-1}(s/c)$$

then we get

$$c = \rho \cos \hat{\mu} \quad \text{and} \quad s = \rho \sin \hat{\mu}$$

and hence

$$l(\mu, \kappa | \eta) \propto \{I_0(\kappa)\}^{-n} \exp \{n\kappa \rho \cos(\mu - \hat{\mu})\}.$$

(It may be worth noting that it can be shown by differentiating  $\rho^2$  with respect to the  $\eta_i$  that  $\rho$  is a maximum when all the observations are equal and that it then equals unity.) It is easy enough now to construct a family of conjugate priors, but for simplicity let us consider a reference prior  $p(\mu, \kappa) \propto 1$ .

It seems reasonable enough to take a uniform prior in  $\mu$  and to take independent priors for  $\mu$  and  $\kappa$ , but it is not so clear that a uniform prior in  $\kappa$  is sensible. Schmitt (1969, Section 10.2) argues that a uniform prior in  $\kappa$  is a sensible compromise and notes that there are difficulties in using a prior proportional to  $1/\kappa$  since, unlike the precision of a normal variable, the concentration parameter of a circular normal distribution can actually equal zero. If this is taken as the prior, then of course  $p(\mu, \kappa | \eta) \propto \{I_0(\kappa)\}^{-n} \exp \{n\kappa \rho \cos(\mu - \hat{\mu})\}$ .

### 3.9.2 Example

Batschelet (1981, Example 4.3.1) quotes data on the time of day of major traffic accidents in a major city. In an obvious sense, the time of day can be regarded as a circular measure, and it is meaningful to ask what is the mean time of day at which accidents occur and how tightly clustered about this time these times are. Writing  $\eta = 360\{h + (m/60)\}/24$

the  $n = 21$  observations are as follows:

hr	min	$\eta$	$\cos \eta$	$\sin \eta$
00	56	14°	0.9703	0.2419
03	08	47°	0.6820	0.7314
04	52	73°	0.2923	0.9563
07	16	109°	-0.3256	0.9455
08	08	122°	-0.5299	0.8480
10	00	150°	-0.8660	0.5000
11	24	171°	-0.9877	0.1564
12	08	182°	-0.9994	-0.0349
13	28	202°	-0.9272	-0.3746
14	16	214°	-0.8290	-0.5592
16	20	245°	-0.4226	-0.9063
16	44	251°	-0.3256	-0.9455
17	04	256°	-0.2419	-0.9703
17	20	260°	-0.1736	-0.9848
17	24	261°	-0.1564	-0.9877
18	08	272°	0.0349	-0.9994
18	16	274°	0.0698	-0.9976
18	56	284°	0.2419	-0.9703
19	32	293°	0.3907	-0.9205
20	52	313°	0.6820	-0.7314
22	08	332°	0.8829	-0.4695
Total			-2.5381	-6.4723
$c = -0.1209$			$s = -0.3082$	

This results in  $\rho = 0.3311$  and  $\tan \hat{\mu} = 2.5492$  and so (allowing for the signs of  $c$  and  $s$ )  $\hat{\mu} = 248^\circ 34'$  (or in terms of a time scale 16h 34m) and so the posterior density takes the form  $p(\mu, \kappa | \eta) \propto \exp\{-n \log\{I_0(\kappa)\} + n\kappa\rho \cos(\mu - \hat{\mu})\}$ ,

where  $\rho$  and  $\mu$  take these values. It is, however, difficult to understand what this means without experience of this distribution, and yet there is no simple way of finding HDRs. This, indeed, is one reason why a consideration of the circular normal distribution has been included, since it serves to emphasize that there are cases where it is difficult if not impossible to avoid numerical integration.

### 3.9.3 Construction of an HDR by numerical integration

By writing  $\lambda = (\frac{1}{2}\kappa)^2$  and taking the first few terms in the power series quoted earlier for  $I_0(\kappa)$ , we see that for  $0 \leq \kappa \leq 2.0$

$$I_0(\kappa) = 1 + \lambda + \lambda^2/4 + \lambda^3/36$$

to within 0.002. We can thus deduce some values for  $I_0(\kappa)$  and  $\kappa\rho$ , namely,

$\kappa$	0.0	0.5	1.0	1.5	2.0
$\lambda = (\frac{1}{2}\kappa)^2$	0.0	0.062	0.250	0.562	1.000
$I_0(\kappa)$	1.0	1.063	1.266	1.647	2.278
$\log\{I_0(\kappa)\}$	0.0	0.061	0.236	0.499	0.823
$\kappa\rho$	0.0	0.166	0.331	0.497	0.662

As  $n = 21$ , this implies that (ignoring the constant) the posterior density for  $\kappa = 0.0, 0.5, 1.0, 1.5$ , and  $2.0$  and for values of  $\mu$  at  $45^\circ$  intervals from  $\hat{\mu}$  is

$\mu \setminus \kappa$	0.0	0.5	1.0	1.5	2.0
158°	1.000	0.278	0.007	0.000	0.000
203°	1.000	3.267	0.960	0.045	0.000
248°	1.000	9.070	7.352	0.959	0.034
293°	1.000	3.267	0.960	0.045	0.000
338°	1.000	0.278	0.007	0.000	0.000

In order to say anything about the marginal density of  $\mu$ , we need to integrate out  $\kappa$ . In order to do this, we can use Simpson's Rule. Using this rule, the integral of a function between  $a$  and  $b$  can be approximated by the sum

$$\int_a^b f(x) dx \propto f(x_0) + 4f(x_1) + 2f(x_2) + 4f(x_3) + f(x_4),$$

where the  $x_i$  are equally spaced with  $x_0=a$  and  $x_4=b$ . Applying it to the aforementioned figures, we can say that very roughly the density of  $\mu$  is

proportional to the following values:  $\mu \quad 158 \quad 203 \quad 248 \quad 293 \quad 338$   
Density    2.13    16.17    55.84    16.17    2.13

Integrating over intervals of values of  $\mu$  using the (even more crude) approximation

$$\int_{a-45}^{a+45} f(x) dx \propto f(a-45) + 4f(a) + f(a+45)$$

(and taking the densities below 158 and above 338 to be negligible) the probabilities that  $\mu$  lies in intervals centred on various values are proportional to

Centre    158    203    248    293    338    Total  
the values stated: Value    25    123    256    123    25    552

It follows that the probability that  $\mu$  lies in the range (203, 293) is about  $256/552=0.46$ , and thus this interval is close to being a 45% HDR.

### 3.9.4 Remarks

The main purpose of this section is to show in some detail, albeit with very crude numerical methods, how a Bayesian approach can deal with a problem which does not lead to a neat posterior distribution values of which are tabulated and readily available. In practice, if you need to approach such a problem, you would have to have recourse to numerical integration techniques on a computer, probably using MCMC as mentioned at the start of this section, but the basic

ideas would be much the same.

## 3.10 Approximations based on the likelihood

### 3.10.1 Maximum likelihood

Suppose, as usual, that we have independent observations  $x = (x_1, x_2, \dots, x_n)$  whose distribution depends on an unknown parameter  $\theta$  about which we want to make inferences. Sometimes it is useful to quote the posterior mode, that is, that value of  $\theta$  at which the posterior density is a maximum, as a single number giving some idea of the location of the posterior distribution of  $\theta$ ; it could be regarded as the ultimate limit of the idea of an HDR. However, some Bayesians are opposed to the use of *any* single number in this way [see Box and Tiao (1992, Section A5.6)].

If the likelihood dominates the prior, the posterior mode will occur very close to the point  $\hat{\theta}$  at which the likelihood is a maximum. Use of  $\hat{\theta}$  is known as the *method of maximum likelihood* and is originally due to Fisher (1922). One notable point about maximum likelihood estimators is that if  $\psi(\theta)$  is any function of  $\theta$  then it is easily seen that  $\hat{\psi} = \psi(\hat{\theta})$

because the point at which  $p(x|\theta)$  is a maximum is not affected by how it is labelled. This invariance is not true of the exact position of the maximum of the posterior, nor indeed of HDRs, because these are affected by the factor  $|d\psi/d\theta|$ .

You should note that the maximum likelihood estimator is often found by the Newton–Raphson method. Suppose that the likelihood is  $l(\theta|x)$  and that its logarithm (in which it is often easier to work) is  $L(\theta|x)$ . In order to simplify the notation, we may sometimes omit explicit reference to the data and write  $L(\theta)$

for  $L(\theta|x)$ . We seek  $\hat{\theta}$  such that  $\frac{\partial l(\theta|x)}{\partial \theta} \Big|_{\theta=\hat{\theta}} = 0$

or equivalently that it satisfies the so-called likelihood equation

$$L'(\hat{\theta}) = 0,$$

so that the score vanishes.

### 3.10.2 Iterative methods

If  $\theta_k$  is an approximation to  $\hat{\theta}$  then using Taylor's Theorem  $0 = L'(\hat{\theta}) = L'(\theta_k) + (\hat{\theta} - \theta_k)L''(\theta^*)$ ,

where  $\theta^*$  is between  $\hat{\theta}$  and  $\theta_k$ . In most cases,  $L''(\theta^*)$  will not differ much from

$L''(\hat{\theta})$  and neither will differ much from its expectation over  $x$ . However,  $\mathbb{E}L''(\theta) = -I(\theta|x)$ ,

where  $I(\theta|x)$  is Fisher's information which was introduced earlier in Section 3.3 in connection with Jeffreys' rule. We note that, although  $L''(\theta)$  does depend on the value  $x$  observed, the information  $I(\theta|x)$  depends on the *distribution* of the random variable  $\tilde{x}$  rather than on the value  $x$  observed on this particular occasion, and to this extent the notation, good though it is for other purposes, is misleading. However, the value of  $I(\hat{\theta}|x)$  does depend on  $x$ , because  $\hat{\theta}$  does.

It follows that as  $k \rightarrow \infty$  the value of  $L''(\theta^*)$  tends to  $-I(\theta|x)$ , so that a better approximation than  $\theta_k$  will usually be provided by either of  $\theta_{k+1} = \theta_k - L'(\theta_k)/L''(\theta_k)$ ,

the Newton–Raphson method, or by

$$\theta_{k+1} = \theta_k + L'(\theta_k)/I(\theta_k|x),$$

the *method of scoring for parameters*. The latter method was first published in a paper by Fisher (1925a).

It has been shown by Kale (1961) that the method of scoring will usually be the quicker process for large  $n$  unless high accuracy is ultimately required. In perverse cases both methods can fail to converge or can converge to a root which does not give the absolute maximum.

### 3.10.3 Approximation to the posterior density

We can also observe that, since  $L'(\hat{\theta}) = 0$ , in the neighbourhood of  $\hat{\theta}$

$$L(\theta) = L(\hat{\theta}) + \frac{1}{2}(\theta - \hat{\theta})^2 L''(\hat{\theta}) + \dots,$$

so that approximately

$$\begin{aligned} l(\theta|x) &= \exp \left\{ L(\hat{\theta}) + \frac{1}{2}(\theta - \hat{\theta})^2 L''(\hat{\theta}) \right\} \\ &\propto \exp \left\{ \frac{1}{2}(\theta - \hat{\theta})^2 L''(\hat{\theta}) \right\}. \end{aligned}$$

Hence, the likelihood is approximately proportional to an  $N(\hat{\theta}, -1/L''(\hat{\theta}))$  density, and so approximately to an  $N(\hat{\theta}, 1/I(\hat{\theta}|x))$  density. We can thus construct approximate HDRs by using this approximation to the likelihood and assuming that the likelihood dominates the prior.

### 3.10.4 Examples

*Normal variance.* For the normal variance (with known mean  $\theta$ )

$$L(\phi) = \log l(\phi|x) = -\frac{1}{2}n \log \phi - \frac{1}{2}S/\phi + \text{constant}$$

where  $S = \sum(X_i - \theta)^2$ , so that

$$L'(\phi) = -\frac{1}{2}n/\phi + \frac{1}{2}S/\phi^2.$$

In this case, the likelihood equation is solved without recourse to iteration to give

$$\hat{\phi} = S/n.$$

Further

$$\begin{aligned} L''(\phi) &= \frac{1}{2}n/\phi^2 - S/\phi^3, \\ -1/L''(\hat{\phi}) &= 2S^2/n^3. \end{aligned}$$

Alternatively

$$I(\phi|x) = -\mathbb{E}L''(\phi) = -\frac{1}{2}n/\phi^2 + \mathbb{E}S/\phi^3$$

and as  $S \sim \phi\chi_n^2$ , so that  $\mathbb{E}S = n\phi$ , we have

$$\begin{aligned} I(\phi|x) &= \frac{1}{2}n/\phi^2, \\ 1/I(\hat{\phi}|x) &= 2S^2/n^3. \end{aligned}$$

Of course, there is no need to use an iterative method to find  $\hat{\phi}$  in this case, but the difference between the formulae for  $L''(\phi)$  and  $I(\phi|x)$  is illustrative of the extent to which the Newton–Raphson method and the method of scoring differ from one another. The results suggest that we approximate the posterior distribution of  $\phi$  [which we found to be  $(S_0 + S)\chi_{\nu+n}^{-2}$  if we took a conjugate prior] by  $\phi \sim N(S/n, 2S^2/n^3)$ .

With the data we considered in Section 2.8 on HDRs for the normal variance, we had  $n = 20$  and  $S = 664$ , so that  $2S^2/n^3 = 110.224$ . The approximation would suggest a 95% HDR between  $664/20 \pm 1.96\sqrt{110.224}$ , that is the interval (13, 54) as opposed to the interval (19, 67) which was found in Section 2.8.

This example is deceptively simple – the method is of greatest use when analytic solutions are difficult or impossible. Further, the accuracy is greater when sample sizes are larger.

*Poisson distribution.* We can get another deceptively simple example by supposing that  $x = (x_1, x_2, \dots, x_n)$  is an  $n$ -sample from  $P(\lambda)$  and that  $T = \sum x_i$ , so

$$\begin{aligned} I(\lambda|x) &\propto (\lambda^T/T!) \exp(-n\lambda) \\ L(\lambda) &= T \log \lambda - n\lambda + \text{constant} \end{aligned}$$

that (as shown in Section 3.4)  $L'(\lambda) = (T/\lambda) - n$

and the likelihood equation is again solved without iteration, this time giving

$$L''(\lambda) = -T/\lambda^2$$

$\hat{\lambda} = T/n = \bar{x}$ . Further  $I(\lambda|x) = n\lambda/\lambda^2 = n/\lambda$

and  $I(\hat{\lambda}|x) = -L''(\hat{\lambda}) = n^2/T$ . This suggests that we can approximate the posterior

of  $\lambda$  (which we found to be  $(S_0 + 2n)\chi_{\nu+2T}^2$  if we took a conjugate prior) by  $\lambda \sim N(T/n, T/n^2)$ .

*Cauchy distribution.* Suppose  $x = (x_1, x_2, \dots, x_n)$  is an  $n$ -sample from  $C(\theta, 1)$ , so

$$p(x|\theta) = \prod \pi^{-1}\{1 + (x_i - \theta)^2\}^{-1} \quad (-\infty < x_i < \infty \text{ for all } i)$$

$$L(\theta) = \text{constant} - \sum \log\{1 + (x_i - \theta)^2\}$$

$$L'(\theta) = 2 \sum (x_i - \theta)/\{1 + (x_i - \theta)^2\}$$

$$L''(\theta) = 2 \sum \{(x_i - \theta)^2 - 1\}/\{1 + (x_i - \theta)^2\}^2.$$

$$\text{that } L'(\theta)/L''(\theta) = \frac{\sum (x_i - \theta)/\{1 + (x_i - \theta)^2\}}{\sum \{(x_i - \theta)^2 - 1\}/\{1 + (x_i - \theta)^2\}^2}.$$

It is easily seen that

$$I(\theta|x) = -E L''(\theta) = (4n/\pi) \int_0^\infty (1-x^2)/(1+x^2)^3 dx.$$

On substituting  $x = \tan \psi$  and using standard reduction formulae, it follows that

$$I(\theta|x) = n/2$$

from which it can be seen that successive approximations to  $\hat{\theta}$  can be found

$$\theta_{k+1} = \theta_k + (4/n) \sum_{i=1}^n (x_i - \theta_k)/\{1 + (x_i - \theta_k)^2\}.$$

using the method of scoring by setting

The iteration could, for example, be started from the sample median, that is, the observation which is in the middle when they are arranged in increasing order. For small  $n$  the iteration may not converge, or may converge to the wrong answer (see Barnett, 1966), but the process usually behaves satisfactorily.

Real life data from a Cauchy distribution are rarely encountered, but the following values are simulated from a  $C(\theta, 1)$  distribution (the value of  $\theta$  being, in fact, 0): -0.774; 0.597; 7.575; 0.397; -0.865; -0.318; -0.125; 0.961; 1.039.

The sample median of the  $n = 9$  values is 0.397. If we take this as our first approximation  $\theta_0$  to  $\hat{\theta}$ , then

$$\theta_1 = 0.107; \quad \theta_2 = 0.201; \quad \theta_3 = 0.173; \quad \theta_4 = 0.181; \quad \theta_5 = 0.179$$

and all subsequent  $\theta_k$  equal 0.179 which is, in fact, the correct value of  $\hat{\theta}$ . Since  $I(\theta|x) = n/2 = 9/2$ , an approximate 95% HDR for  $\theta$  is  $0.179 \pm 1.96\sqrt{(2/9)}$ , that is the interval (-0.74, 1.10). This does include the true value, which we happen to know is 0, but of course the value of  $n$  has been chosen unrealistically small in order to illustrate the method without too much calculation.

It would also be possible in this case to carry out an iteration based on the Newton-Raphson method

$$\theta_{k+1} = \theta_k - L'(\theta_k)/L''(\theta_k)$$

using the above formula for  $L'(\theta)/L''(\theta)$ , but as explained earlier, it is in general

better to use the method of scoring.

### 3.10.5 Extension to more than one parameter

If we have two parameters, say  $\theta$  and  $\phi$ , which are both unknown, a similar argument shows that the maximum likelihood occurs at  $(\hat{\theta}, \hat{\phi})$ , where  $\partial L/\partial\theta = \partial L/\partial\phi = 0$ .

Similarly, if  $(\theta_k, \phi_k)$  is an approximation, a better one is  $(\theta_{k+1}, \phi_{k+1})$ , where

$$\begin{pmatrix} \theta_{k+1} \\ \phi_{k+1} \end{pmatrix} = \begin{pmatrix} \theta_k \\ \phi_k \end{pmatrix} - \begin{pmatrix} \partial^2 L/\partial\theta^2 & \partial^2 L/\partial\theta\partial\phi \\ \partial^2 L/\partial\theta\partial\phi & \partial^2 L/\partial\phi^2 \end{pmatrix}^{-1} \begin{pmatrix} \partial L/\partial\theta \\ \partial L/\partial\phi \end{pmatrix},$$

where the derivatives are evaluated at  $(\theta_k, \phi_k)$  and the matrix of second derivatives can be replaced by its expectation, which is minus the information matrix as defined in Section 3.3 on Jeffreys' rule.

Further, the likelihood and hence the posterior can be approximated by a bivariate normal distribution of mean  $(\hat{\theta}, \hat{\phi})$  and variance–covariance matrix whose inverse is equal to minus the matrix of second derivatives (or the information matrix) evaluated at  $(\hat{\theta}, \hat{\phi})$ .

All of this extends in an obvious way to the case of more than two unknown parameters.

### 3.10.6 Example

We shall consider only one, very simple, case, that of a normal distribution of unknown mean and variance. In this case,

$$L(\theta, \phi) = -\frac{1}{2}n \log \phi - \frac{1}{2}\{S + n(\bar{x} - \theta)^2\}/\phi$$

$$\partial L/\partial\theta = n(\bar{x} - \theta)/\phi$$

$$\partial L/\partial\phi = -\frac{1}{2}n/\phi + \frac{1}{2}\{S + n(\bar{x} - \theta)^2\}/\phi^2,$$

where  $S = \sum(x_i - \bar{x})^2$ , so that

$$\hat{\theta} = \bar{x},$$

$$\hat{\phi} = S/n = (n - 1)s^2/n.$$

Further, it is easily seen that

$$\begin{pmatrix} \partial^2 L/\partial\theta^2 & \partial^2 L/\partial\theta\partial\phi \\ \partial^2 L/\partial\theta\partial\phi & \partial^2 L/\partial\phi^2 \end{pmatrix} = \begin{pmatrix} -n/\phi & -n(\bar{x} - \theta)/\phi^2 \\ -n(\bar{x} - \theta)/\phi^2 & \frac{1}{2}n/\phi^2 - \{S + n(\bar{x} - \theta)^2\}/\phi^3 \end{pmatrix}$$

which at  $(\hat{\theta}, \hat{\phi})$  reduces to

$$\begin{pmatrix} -n/\hat{\phi} & 0 \\ 0 & -n/2\hat{\phi}^2 \end{pmatrix}.$$

Because the off-diagonal elements vanish, the posteriors for  $\theta$  and  $\phi$  are

approximately independent. Further, we see that approximately  $\theta \sim N(\widehat{\theta}, \widehat{\phi}/n)$  and  $\phi \sim N(\widehat{\phi}, 2\widehat{\phi}^2/n)$ .

In fact, we found in Section 2.12 on normal mean and variance both unknown that with standard reference priors, the posterior for  $\theta$  and  $\phi$  is a normal/chi-squared distribution and the marginals are such that  $(\bar{x} - \theta)/(s/\sqrt{n}) \sim t_{n-1}$  and  $\phi \sim S\chi_{n-1}^{-2}$

which implies that the means and variances are

$$\begin{aligned} E\theta &= \bar{x} = \widehat{\theta}, & V\theta &= (n-1)s^2/n(n-3) \cong \widehat{\phi}/n, \\ E\phi &\cong S/(n-3), & V\phi &= 2S^2/(n-3)^2(n-5) \cong 2\widehat{\phi}^2/n. \end{aligned}$$

This shows that for large  $n$  the approximation is indeed valid.

## 3.11 Reference posterior distributions

### 3.11.1 The information provided by an experiment

Bernardo (1979) suggested another way of arriving at a reference standard for Bayesian theory. The starting point for this is that the log-likelihood ratio  $\log\{p(x|\theta_1)/p(x|\theta_2)\}$  can be regarded as the information provided by the observation  $x$  for discrimination in favour of  $\theta_1$  against  $\theta_2$  (cf. Good, 1950, Section 6.1). This led Kullback and Leibler (1951) to define the mean

$$\mathcal{I}(1:2) = \int p(x|\theta_1) \log\{p(x|\theta_1)/p(x|\theta_2)\} dx$$

information in such data to be

(cf. Kullback, 1968, and Barnett, 1999, Section 8.6).<sup>2</sup> Note that although there is a relationship between information as defined here and Fisher's information  $I$  as defined in Section 3.3 earlier (see Kullback, 1968, Chapter 2, Section 6), you are best advised to think of this as a quite separate notion. It has in common with Fisher's information the property that it depends on the distribution of the data rather than on any particular value of it.

Following this, Lindley (1956) defined the expected amount  $\mathcal{I}(x|p)$  of information that the observation  $x$  will provide about an unknown parameter  $\theta$

$$\text{when the prior density for } \theta \text{ is } p(\theta) \text{ to be } \int p(\theta|x) \log \frac{p(\theta|x)}{p(\theta)} d\theta.$$

The observation  $x$  is, of course, random, and hence we can define the expected information that the observation  $x$  will provide to be

$$\mathcal{I} = \int p(x) \int p(\theta|x) \log \frac{p(\theta|x)}{p(\theta)} d\theta dx$$

(a similar expression occurs in Shannon, 1948, Section 24). Two obviously equivalent expressions are

$$\int p(\theta) \int p(x|\theta) \log \frac{p(x|\theta)}{p(x)} dx d\theta \quad \text{and} \quad \iint p(\theta, x) \log \frac{p(\theta, x)}{p(\theta)p(x)} d\theta dx.$$

It is easily seen using the usual change-of-variable rule that the information defined by this expression is invariant under a one-to-one transformation. It can be used as a basis for Bayesian design of experiments. It has various appealing properties, notably that  $\mathcal{I} \geq 0$  with equality if and only if  $p(x|\theta)$  does not depend on  $\theta$  (see Hardy, Littlewood and Pólya, 1952, Theorem 205). Further, it turns out that if an experiment consists of two observations, then the total information it provides is the information provided by one observation plus the mean amount provided by the second given the first (as shown by Lindley, 1956).

We now define  $\mathcal{I}_n$  to be the amount of information about  $\theta$  to be expected from  $n$  independent observations with the same distribution as  $x$ . By making an infinite number of observations one would get to know the precise value of  $\theta$ , and consequently  $\mathcal{I}_\infty$  measures the amount of information about  $\theta$  when the prior is  $p(\theta)$ . It seems natural to define ‘vague initial knowledge’ about  $\theta$  as that described by that density  $p(\theta)$  which maximizes the missing information.

In the continuous case, we usually find that  $\mathcal{I}_n \rightarrow \infty$  for all prior  $p(\theta)$ , and hence we need to use a limiting process. This is to be expected since an infinite amount of information would be required to know a real number exactly. We define  $p_n(\theta|x)$  to be the posterior density corresponding to that prior  $p_n(\theta)$  which maximizes  $\mathcal{I}_n$  (it can be shown that in reasonable cases a unique maximizing function exists). Then the *reference posterior*  $p(\theta|x)$  is defined as the limit  $\lim p_n(\theta|x)$ . Functions can converge in various senses, and so we need to say what we mean by the limit of these densities. In fact we have to take convergence to mean convergence of the distribution functions at all points at which the limiting distribution function is continuous.

We can then define a *reference prior* as any prior  $p(\theta)$  which satisfies  $p(\theta|x) \propto p(\theta)p(x|\theta)$ . This rather indirect definition is necessary because convergence of a set of posteriors does not necessarily imply that the corresponding priors converge in the same sense. To see this, consider a case where the observations consist of a single binomial variable  $x \sim B(k, \theta)$  and the sequence of priors is  $Be(1/n, 1/n)$ . Then the posteriors are  $Be(x+1/n, k-x+1/n)$  which clearly converge to  $Be(x, k-x)$ , which is the posterior corresponding to the Haldane prior  $Be(0, 0)$ . However, the priors themselves have distribution functions which approach a step function with steps of  $\frac{1}{2}$  at 0 and 1, and that corresponds to a discrete prior distribution which gives probability  $\frac{1}{2}$  each to the

values 0 and 1.

To proceed further, we suppose that  $x = (x_1, \dots, x_n)$  is the result of our  $n$  independent observations of  $x$  and we define *entropy* by

$$H\{p(\theta)\} = - \int p(\theta) \log p(\theta) d\theta$$

(this is a function of a distribution for  $\theta$  and is not a function of any particular value of  $\theta$ ). Then using  $p(\theta) = \int p(x)p(\theta|x) dx$  and  $p(x) = \int p(\theta)p(x|\theta) d\theta$  we see

$$\begin{aligned} I_n &= \int p(x) \int p(\theta|x) \log \frac{p(\theta|x)}{p(\theta)} d\theta dx \\ &= H\{p(\theta)\} - \int p(x) H\{p(\theta|x)\} dx \\ &= H\{p(\theta)\} - \int p(\theta) \int p(x|\theta) H\{p(\theta|x)\} dx d\theta \\ &\text{that } = \int p(\theta) \log \left\{ \exp \left( - \int p(x|\theta) H\{p(\theta|x)\} dx \right) \right\} d\theta \end{aligned}$$

(the last equation results from simple manipulations as  $\exp$  and  $\log$  are inverse functions). It follows that we can write  $I_n$  in the form

$$\mathcal{I}_n = \int p(\theta) \log \frac{f(\theta)}{p(\theta)} d\theta.$$

It can be shown using the calculus of variations that the information  $\mathcal{I}_n$  is maximized when  $p(\theta) \propto f(\theta)$  (see Bernardo and Smith, 1994, Section 5.4.2). It follows (provided the functions involved are well behaved) that the sequence of

$$\text{densities } p_n(\theta) \propto \exp \left( - \int p(x|\theta) H\{p(\theta|x)\} dx \right)$$

approaches the reference prior. There is a slight difficulty in that the posterior density  $p(\theta|x)$  which figures in the above expression depends on the prior, but we know that this dependence dies away as  $n \rightarrow \infty$ .

### 3.11.2 Reference priors under asymptotic normality

In cases where the approximations derived in Section 3.10 are valid, the posterior distribution  $p(\theta|x)$  is  $N(\hat{\theta}, 1/I(\hat{\theta}|x))$  which by the additive property of Fisher's information is  $N(\hat{\theta}, 1/nI(\hat{\theta}|x))$ . Now it is easily seen that the entropy of an  $N(\theta, \phi)$  density is

$$-\int \frac{1}{\sqrt{2\pi\phi}} \exp(-\frac{1}{2}(z-\theta)^2/\phi) \left\{ -\log \sqrt{2\pi\phi} - \frac{1}{2}(z-\theta)^2/\phi \right\} dz = \log \sqrt{2\pi e \phi}$$

(writing  $\frac{1}{2}$  as  $\log \sqrt{e}$ ) from which it follows that  $H\{p(\theta|x)\} = \log \sqrt{2\pi e / nI(\hat{\theta}|x)}$  to the extent to which the approximation established in the last section is correct. Thus, we have

$$\begin{aligned}
\int p(x|\theta) H\{p(\theta|x)\} dx &= - \int p(x|\theta) \log \sqrt{2\pi e/nI(\hat{\theta}|x)} dx \\
&= - \int p(\hat{\theta}|\theta) \log \sqrt{2\pi e/nI(\hat{\theta}|x)} d\hat{\theta} \\
&= - \log \sqrt{2\pi e/nI(\theta|x)}.
\end{aligned}$$

since the approximation in the previous section shows that  $p(\hat{\theta}|\theta)$  is negligible except where  $\hat{\theta}$  is close to  $\theta$ . It follows on dropping a constant that  $p_n(\theta) \propto \exp\{\log \sqrt{I(\theta|x)}\} = \sqrt{I(\theta|x)}$

and so we have another justification for Jeffreys' prior which we first introduced in Section 3.3.

If this were all that this method could achieve, it would not be worth the aforementioned discussion. Its importance lies in that it can be used for a wider class of problems and that further it gives sensible answers when we have nuisance parameters.

### 3.11.3 Uniform distribution of unit length

To see the first point, consider the case of a uniform distribution over an interval of unit length with unknown centre, so that we have observations  $x \sim U(\theta - \frac{1}{2}, \theta + \frac{1}{2})$ , and as usual let  $x$  be the result of our  $n$  independent observations of  $x$ . Much as in Section 3.5, we find that if  $m = \min x_i$  and  $M = \max x_i$  then the posterior is  $p(\theta|x) \propto p(\theta) I_{(M-\frac{1}{2}, m+\frac{1}{2})}(\theta)$ .

For a large sample, the interval in which this is nonzero will be small and (assuming suitable regularity)  $p(\theta)$  will not vary much in it, so that asymptotically  $p(\theta|x) \sim U(M - \frac{1}{2}, m + \frac{1}{2})$ . It follows that

$$\begin{aligned}
H\{p(\theta|x)\} &= \int_{M-\frac{1}{2}}^{m+\frac{1}{2}} \frac{1}{1-(M-m)} \log\{1-(M-m)\}^{-1} d\theta = -\log\{1-(M-m)\} \\
&= -E \log\{1-(M-m)\}
\end{aligned}$$

which is asymptotically equal to

$$-\log\{1-(EM-Em)\}.$$

Since  $M - (\theta - \frac{1}{2})$  is the maximum of  $n$  observations uniformly distributed on  $[0, 1]$  we have  $P(M \leq u) = P(X_i \leq u \text{ for } i = 1, \dots, n) = u^n$

from which it follows that the density of  $M - (\theta - \frac{1}{2})$  is proportional to  $u^{n-1}$ , so that  $M - (\theta + \frac{1}{2}) \sim \text{Be}(n, 1)$  and hence  $EM = n/(n+1)$ . Similarly we find that  $Em - (\theta - \frac{1}{2}) = 1/(n+1)$ , so that

$$\begin{aligned}-\log\{1-(EM-Em)\} &= -\log\left\{1-\left(\frac{n}{n+1}+\theta-\frac{1}{2}-\frac{1}{n+1}-\theta+\frac{1}{2}\right)\right\} \\ &= \log\left(\frac{n+1}{2}\right).\end{aligned}$$

Because this does not depend on  $\theta$  it follows that  $p_n(\theta)$  does not depend on  $\theta$  and so is uniform. Taking limits, the reference prior is also uniform.

Note that in this case the posterior is very far from normality, so that the theory cannot be applied as in Subsection 3.11.2, headed ‘Reference priors under asymptotic normality’, but nevertheless a satisfactory reference prior can be devised.

### 3.11.4 Normal mean and variance

When we have two parameters, as in the case of the mean and variance of an  $N(\theta, \phi)$  distribution, we often want to make inferences about the mean  $\theta$ , so that  $\phi$  is a nuisance parameter. In such a case, we have to choose a conditional prior  $p(\phi | \theta)$  for the nuisance parameter which describes personal opinions, previous observations, or else is ‘diffuse’ in the sense of the priors we have been talking about.

When we want  $p(\phi | \theta)$  to describe diffuse opinions about  $\phi$  given  $\theta$ , we would expect, for the aforementioned reasons, to maximize the missing information about  $\phi$  given  $\theta$ . This results in the sequence  $p_n(\phi | \theta) \propto \exp\left(-\int p(x | \theta, \phi) H\{p(\phi | \theta, x)\} dx\right)$ .

Now we found in Section 3.10 that in the case where we have a sample of size  $n$  from a normal distribution, the asymptotic posterior distribution of  $\phi$  is  $N(S/n, 2S^2/n^3)$ , which we may write as  $N(\hat{\phi}, 2\hat{\phi}^2/n)$ , and consequently (using the form derived at the start of the subsection on ‘Reference priors under asymptotic normality’) its entropy is  $H\{p(\phi | \theta, x)\} = \log \sqrt{2\pi e 2\hat{\phi}^2/n} = \log \hat{\phi} + \text{const.}$

It follows that

$$\begin{aligned}p_n(\phi | \theta) &\propto \exp\left(-\int p(x | \theta, \phi) \log \hat{\phi} dx\right) \\ &= \exp\left(-\int p(\hat{\phi} | \theta, \phi) \log \hat{\phi} d\hat{\phi}\right) \\ &= \exp(-\log \hat{\phi}) \\ &= 1/\hat{\phi}.\end{aligned}$$

In the limit we get that

$$p(\phi | \theta) \propto 1/\hat{\phi}.$$

In this case, the posterior for the mean is well approximated by  $N(\hat{\theta}, \hat{\phi}/n)$ , where  $\hat{\theta} = \bar{x}$  and  $\hat{\phi} = S/n$ , so that the entropy is  $H\{p(\theta | x)\} = \log \sqrt{2\pi e \hat{\phi}/n} = \frac{1}{2} \log \hat{\phi} + \text{const.}$ . We thus get

$$\begin{aligned} p_n(\theta) &\propto \exp\left(-\int p(x | \theta, \phi) H\{p(\theta | x)\} dx\right) \\ &\propto \exp\left(-\int p(x | \theta, \phi) \frac{1}{2} \log \hat{\phi} dx\right) \\ &= \exp\left(-\int p(\hat{\phi} | \theta, \phi) \frac{1}{2} \log \hat{\phi} d\hat{\phi}\right) \\ &= \exp\left(-\int p(\phi | \theta) \frac{1}{2} \log \phi d\phi\right) \end{aligned}$$

using the facts that  $p(\hat{\phi} | \theta, \phi)$  is negligible except where  $\hat{\phi}$  is close to  $\phi$  and that, of course,  $p(\phi | \theta, \phi)$  must equal  $p(\phi | \theta)$ . We note that if  $p(\phi | \theta)$  does not depend on  $\theta$ , and so, in particular, in the case where  $p(\phi | \theta) \propto 1/\phi$ , the density  $p_n(\theta)$  is a constant and in the limit the reference prior  $p(\theta)$  is also constant, so giving the usual reference prior  $p(\theta) \propto 1$ . It then follows that the joint reference prior is  $p(\theta, \phi) = p(\theta)p(\phi | \theta) \propto 1/\phi$ .

This, as we noted at the end of Section 3.3 is *not* the same as the prior  $p(\theta, \phi) \propto \phi^{-3/2}$  given by the two-parameter version of Jeffreys' rule. If we want to make inferences about  $\phi$  with  $\theta$  being the nuisance parameter, we obtain the same reference prior.

There is a temptation to think that whatever parameters we adopt we will get the same reference prior, but this is not the case. If we define the standard deviation as  $\sigma = \sqrt{\phi}$  and the coefficient of variation or standardized mean as  $\lambda = \theta/\sigma$ , then we find  $p(\lambda, \sigma) \propto (1 + \frac{1}{2}\lambda^2)^{-1/2} \phi^{-1}$

(see Bernardo and Smith, 1994, Examples 5.17 and 5.26) which corresponds to

$$p(\theta, \phi) \propto \left(1 + \frac{1}{2} \frac{\theta^2}{\phi}\right)^{-1/2} \phi^{-3/2}.$$

### 3.11.5 Technical complications

There are actually some considerable technical complications about the process of obtaining reference posteriors and priors in the presence of nuisance parameters, since some of the integrals may be infinite. It is usually possible to deal with this difficulty by restricting the parameter of interest to a finite range and then increasing this range sequentially, so that in the limit all possible values are included. For details, see Bernardo and Smith (1994, Section 5.4.4).

## 3.12 Exercises on Chapter 3

1. Laplace claimed that the probability that an event which has occurred  $n$  times, and has not hitherto failed, will occur again is  $(n+1)/(n+2)$  [see Laplace (1774)], which is sometimes known as *Laplace's rule of succession*. Suggest grounds for this assertion.
2. Find a suitable interval of 90% posterior probability to quote in a case when your posterior distribution for an unknown parameter  $\pi$  is  $\text{Be}(20, 12)$ , and compare this interval with similar intervals for the cases of  $\text{Be}(20.5, 12.5)$  and  $\text{Be}(21, 13)$  posteriors. Comment on the relevance of the results to the choice of a reference prior for the binomial distribution.
3. Suppose that your prior beliefs about the probability  $\pi$  of success in Bernoulli trials have mean  $1/3$  and variance  $1/32$ . Give a 95% posterior HDR for  $\pi$  given that you have observed 8 successes in 20 trials.
4. Suppose that you have a prior distribution for the probability  $\pi$  of success in a certain kind of gambling game which has mean 0.4, and that you regard your prior information as equivalent to 12 trials. You then play the game 25 times and win 12 times. What is your posterior distribution for  $\pi$ ?
5. Suppose that you are interested in the proportion of females in a certain organization and that as a first step in your investigation you intend to find out the sex of the first 11 members on the membership list. Before doing so, you have prior beliefs which you regard as equivalent to 25% of this data, and your prior beliefs suggest that a third of the membership is female.  
Suggest a suitable prior distribution and find its standard deviation.  
Suppose that 3 of the first 11 members turn out to be female; find your posterior distribution and give a 50% posterior HDR for this distribution.  
Find the mean, median and mode of the posterior distribution.  
Would it surprise you to learn that in fact 86 of the total number of 433 members are female?
6. Show that if  $g(x) = \sinh^{-1} \sqrt{(x/n)}$  then
$$g'(x) = \frac{1}{2} n^{-\frac{1}{2}} [(x/n)\{1 + (x/n)\}]^{-\frac{1}{2}}.$$
Deduce that if  $x \sim \text{NB}(n, \pi)$  has a negative binomial distribution of index  $n$  and parameter  $\pi$  and  $z=g(x)$  then  $Ez \cong \sinh^{-1} \sqrt{(x/n)}$  and  $Vz \cong 1/4n$ . What does this suggest as a reference prior for  $\pi$ ?
7. The following data were collected by von Bortkiewicz (1898) on the

number of men killed by horses in certain Prussian army corps in twenty years, the unit being one army corps for one year:

Number of deaths: 0 1 2 3 4 5 and more

Number of units: 144 91 32 11 2 0.

Give an interval in which the mean number  $\lambda$  of such deaths in a particular army corps in a particular year lies with 95% probability.

**8.** Recalculate the answer to the previous question assuming that you had a prior distribution for  $\lambda$  of mean 0.66 and standard deviation 0.115.

**9.** Find the Jeffreys prior for the parameter  $\alpha$  of the Maxwell distribution

$$p(x|\alpha) = \sqrt{\frac{2}{\pi}} \alpha^{3/2} x^2 \exp(-\frac{1}{2}\alpha x^2)$$

and find a transformation of this parameter in which the corresponding prior is uniform.

**10.** Use the two-dimensional version of Jeffreys' rule to determine a prior for the trinomial distribution

$$p(x, y, z|\pi, \rho) \propto \pi^x \rho^y (1 - \pi - \rho)^z.$$

(cf. Exercise 15 on Chapter 2).

**11.** Suppose that  $x$  has a Pareto distribution  $\text{Pa}(\xi, \gamma)$ , where  $\xi$  is known but  $\gamma$  is unknown, that is,  $p(x|\gamma) = \gamma \xi^\gamma x^{-\gamma-1} I_{(\xi, \infty)}(x)$ .

Use Jeffreys' rule to find a suitable reference prior for  $\gamma$ .

**12.** Consider a uniform distribution on the interval  $(\alpha, \beta)$ , where the values of  $\alpha$  and  $\beta$  are unknown, and suppose that the joint distribution of  $\alpha$  and  $\beta$  is a bilateral bivariate Pareto distribution with  $\gamma = 2$ . How large a random sample must be taken from the uniform distribution in order that the coefficient of variation (that is, the standard deviation divided by the mean) of the length  $\beta - \alpha$  of the interval should be reduced to 0.01 or less?

**13.** Suppose that observations  $x_1, x_2, \dots, x_n$  are available from a density

$$p(x|\theta) = (c+1)\theta^{-(c+1)}x^c \quad (0 < x < \theta).$$

Explain how you would make inferences about the parameter  $\theta$  using a conjugate prior.

**14.** What could you conclude if you observed two tramcars numbered, say, 71 and 100?

**15.** In Section 3.8, we discussed Newcomb's observation that the front pages of a well-used table of logarithms tend to get dirtier than the back pages do. What if we had an *antilogarithm* table, that is, a table giving the value of  $x$  when  $\log_{10}x$  is given? Which pages of such a table would be the dirtiest?

**16.** We sometimes investigate distributions on a circle (e.g. von Mises' distribution which is discussed in Section 3.9 on 'The circular normal distribution'). Find a Haar prior for a location parameter on the circle (such as  $\mu$  in the case of von Mises' distribution).

**17.** Suppose that the prior distribution  $p(\mu, \sigma)$  for the parameters  $\mu$  and  $\sigma$  of a Cauchy distribution  $p(x|\mu, \sigma) = \frac{1}{\pi} \frac{\sigma}{\sigma^2 + (x - \mu)^2}$

is uniform in  $\mu$  and  $\sigma$ , and that two observations  $x_1=2$  and  $x_2=6$  are available from this distribution. Calculate the value of the posterior density  $p(\mu, \sigma|x)$  (ignoring the factor  $1/\pi^2$ ) to two decimal places for  $\mu = 0, 2, 4, 6, 8$  and  $\sigma = 1, 2, 3, 4, 5$ . Use Simpson's rule to approximate the posterior marginal density of  $\mu$ , and hence go on to find an approximation to the posterior probability that  $3 < \mu < 5$ .

**18.** Show that if the log-likelihood  $L(\theta|x)$  is a concave function of  $\theta$  for each scalar  $x$  (that is,  $L''(\theta|x) \leq 0$  for all  $\theta$ ), then the likelihood function  $L(\theta|x)$  for  $\theta$  given an  $n$ -sample  $x = (x_1, x_2, \dots, x_n)$  has a unique maximum. Prove that this is the case if the observations  $x_i$  come from a logistic density  $p(x|\theta) = \exp(\theta - x)/\{1 + \exp(\theta - x)\}^2 \quad (-\infty < x < \infty)$ ,

where  $\theta$  is an unknown real parameter. Fill in the details of the Newton–Raphson method and the method of scoring for finding the position of the maximum, and suggest a suitable starting point for the algorithms.

[In many applications of Gibbs sampling, which we consider later in Section 9.4, all full conditional densities are log-concave (see Gilks *et al.*, 1996, Section 5.3.3), so the study of such densities is of real interest.]

**19.** Show that if an experiment consists of two observations, then the total information it provides is the information provided by one observation plus the mean amount provided by the second given the first.

**20.** Find the entropy  $H\{p(\theta)\}$  of a (negative) exponential distribution with density  $p(\theta) = \beta^{-1} \exp(-\theta/\beta)$ .

<sup>1</sup> This problem reappeared as the German tank problem; see Spencer and Largey (1993).

<sup>2</sup> Sometimes denoted  $D_k L(1\|2)$  or  $KL(1\|2)$ .

# 4

## Hypothesis testing

### 4.1 Hypothesis testing

#### 4.1.1 Introduction

If preferred, the reader may begin with the example at the end of this section, then return to the general theory at the beginning.

#### 4.1.2 Classical hypothesis testing

Most simple problems in which tests of hypotheses arise are of the following general form. There is one unknown parameter  $\theta$  which is known to be from a set  $\Theta$ , and you want to know whether  $\theta \in \Theta_0$  or  $\theta \in \Theta_1$  where  $\Theta_0 \cup \Theta_1 = \Theta$  and  $\Theta_0 \cap \Theta_1 = \emptyset$ .

Usually, you are able to make use of a set of observations  $x_1, x_2, \dots, x_n$  whose density  $p(x|\theta)$  depends on  $\theta$ . It is convenient to denote the set of all possible observations  $x = (x_1, x_2, \dots, x_n)$  by  $\mathcal{X}$ .

In the language of classical statistics, it is usual to refer to

$H_0 : \theta \in \Theta_0$  as the *null hypothesis*

and to

$H_1 : \theta \in \Theta_1$  as the *alternative hypothesis*

and to say that if you decide to reject  $H_0$  when it is true then you have made a *Type I error* while if you decide *not* to reject  $H_0$  when it is false then you have made a *Type II error*.

A test is decided by a *rejection region*  $R$  where

$R = \{x; \text{observing } x \text{ would lead to the rejection of } H_0\}$ .

Classical statisticians then say that decisions between tests should be based on the probabilities of Type I errors, that is,  $P(R|\theta)$  for  $\theta \in \Theta_0$  and of Type II errors, that is,

$1 - P(R|\theta)$  for  $\theta \in \Theta_1$ .

In general, the smaller the probability of Type I error, the larger the probability

of Type II error and vice versa. Consequently, classical statisticians recommend a choice of  $R$  which in some sense represents an optimal balance between the two types of errors. Very often  $R$  is chosen, so that the probability of a Type II error is as small as possible subject to the requirement that the probability of a Type I error is always less than or equal to some fixed value  $\alpha$  known as the *size* of the test. This theory, which is largely due to Neyman and Pearson, is to be found in most books on statistical inference and is to be found in its fullest form in Lehmann (1986).

### 4.1.3 Difficulties with the classical approach

Other points will be made later about the comparison between the classical and the Bayesian approaches, but one thing to note at the outset is that, in the classical approach, we consider the probability (for various values of  $\theta$ ) of a set  $R$  to which the vector  $x$  of observations does, or does not, belong. Consequently, we are concerned not merely with the single vector of observations we actually made but also with others we *might* have made but *did not*. Thus, classically, if we suppose that  $x \sim N(\theta, 1)$  and we wish to test whether  $H_0 : \theta = 0$  or  $H_1 : \theta > 0$  is true (negative values being supposed impossible), then we reject  $H_0$  on the basis of a single observation  $x = 3$  because the probability that an  $N(0, 1)$  random variable is 3 *or greater* is 0.001 350, even though we certainly did not make an observation greater than 3. This aspect of the classical approach led Jeffreys (1961, Section 7.2) to remark: What the use of  $P$  implies, therefore, is that a hypothesis that may be true may be rejected because it has not predicted observable results that have not occurred.

Note, however, that the form of the model, in this case the assumption of normally distributed observations of unit variance, does depend on an assumption about the whole distribution of all possible observations.

### 4.1.4 The Bayesian approach

The Bayesian approach is in many ways more straightforward. All we need to do is to calculate the posterior probabilities

$$p_0 = P(\theta \in \Theta_0 | x), \quad p_1 = P(\theta \in \Theta_1 | x)$$

and decide between  $H_0$  and  $H_1$  accordingly. (We note that  $p_0 + p_1 = 1$  as  $\Theta = \Theta_0 \cup \Theta_1$  and  $\Theta_0 \cap \Theta_1 = \emptyset$ .) Although posterior probabilities of hypotheses are our ultimate goal we also need prior probabilities

$$\pi_0 = P(\theta \in \Theta_0) \quad \text{and} \quad \pi_1 = P(\theta \in \Theta_1)$$

to find them. (We note that  $\pi_0 + \pi_1 = 1$  just as  $p_0 + p_1 = 1$ .) It is also useful to consider the *prior odds* on  $H_0$  against  $H_1$ , namely  $\pi_0/\pi_1$  and the *posterior odds* on  $H_0$  against  $H_1$ , namely  $p_0/p_1$ .

(The notion of odds was originally introduced in the very first section of this book). Observe that if your prior odds are close to 1 then you regard  $H_0$  as more or less as likely as  $H_1$  a priori, while if the ratio is large you regard  $H_0$  as relatively likely and when it is small you regard it as relatively unlikely. Similar remarks apply to the interpretation of the posterior odds.

It is also useful to define the *Bayes factor*  $B$  in favour of  $H_0$  against  $H_1$  as  $B = \frac{(p_0/p_1)}{(\pi_0/\pi_1)} = \frac{p_0\pi_1}{p_1\pi_0}$ .

The interest in the Bayes factor is that it can sometimes be interpreted as the ‘odds in favour of  $H_0$  against  $H_1$  that are *given by the data*’. It is worth noting that because  $p_0/p_1 = B(\pi_0/\pi_1)$  and  $p_1 = 1 - p_0$  we can find the posterior probability  $p_0$  of  $H_0$  from its prior probability and the Bayes factor by

$$p_0 = \frac{1}{[1 + (\pi_1/\pi_0)B^{-1}]} = \frac{1}{[1 + \{(1 - \pi_0)/\pi_0\}B^{-1}]}$$

The aforementioned interpretation is clearly valid when the hypotheses are *simple*, that is,

$$\Theta_0 = \{\theta_0\} \quad \text{and} \quad \Theta_1 = \{\theta_1\}$$

for some  $\theta_0$  and  $\theta_1$ . For if so, then  $p_0 \propto \pi_0 p(x|\theta_0)$  and  $p_1 \propto \pi_1 p(x|\theta_1)$ , so that  $\frac{p_0}{p_1} = \frac{\pi_0}{\pi_1} \frac{p(x|\theta_0)}{p(x|\theta_1)}$ ,

and hence, the Bayes factor is

$$B = \frac{p(x|\theta_0)}{p(x|\theta_1)}.$$

It follows that  $B$  is the *likelihood ratio* of  $H_0$  against  $H_1$  which most statisticians (whether Bayesian or not) view as the odds in favour of  $H_0$  against  $H_1$  that are given by the data.

However, the interpretation is not quite as simple when  $H_0$  and  $H_1$  are *composite*, that is, contain more than one member. In such a case, it is convenient to write  $\rho_0(\theta) = p(\theta)/\pi_0$  for  $\theta \in \Theta_0$

and

$$\rho_1(\theta) = p(\theta)/\pi_1 \quad \text{for} \quad \theta \in \Theta_1,$$

where  $p(\theta)$  is the prior density of  $\theta$ , so that  $\rho_0(\theta)$  is the restriction of  $p(\theta)$  to  $\Theta_0$  renormalized to give a probability density over  $\Theta_0$ , and similarly for  $\rho_1(\theta)$ . We

$$\begin{aligned}
p_0 &= P(\theta \in \Theta_0 | x) \\
&= \int_{\theta \in \Theta_0} p(\theta | x) d\theta \\
&\propto \int_{\theta \in \Theta_0} p(\theta) p(x | \theta) d\theta \\
&= \pi_0 \int_{\theta \in \Theta_0} p(x | \theta) \rho_0(\theta) d\theta,
\end{aligned}$$

then have

the constant of proportionality depending solely on  $x$ . Similarly,

$$p_1 \propto \pi_1 \int_{\theta \in \Theta_1} p(x | \theta) \rho_1(\theta) d\theta,$$

and hence, the Bayes factor is

$$B = \frac{(p_0/p_1)}{(\pi_0/\pi_1)} = \frac{\int_{\theta \in \Theta_0} p(x | \theta) \rho_0(\theta) d\theta}{\int_{\theta \in \Theta_1} p(x | \theta) \rho_1(\theta) d\theta}$$

which is the ratio of ‘weighted’ (by  $\rho_0$  and  $\rho_1$ ) likelihoods of  $\Theta_0$  and  $\Theta_1$ .

Because this expression for the Bayes factor involves  $\rho_0$  and  $\rho_1$  as well as the likelihood function  $p(x | \theta)$  itself, the Bayes factor cannot be regarded as a measure of the relative support for the hypotheses provided *solely* by the data. Sometimes, however,  $B$  will be relatively little affected within reasonable limits by the choice of  $\rho_0$  and  $\rho_1$ , and then we *can* regard  $B$  as a measure of relative support for the hypotheses provided by the data. When this is so, the Bayes factor is reasonably objective and might, for example, be included in a scientific report, so that different users of the data could determine their personal posterior odds by multiplying their personal prior odds by the factor.

It may be noted that the Bayes factor is referred to by a few authors simply as the factor. Jeffreys (1961) denoted it by  $K$ , but did not give it a name. A number of authors, most notably Peirce (1878) and (independently) Good (1950, 1983 and elsewhere), refer to the logarithm of the Bayes factor as the *weight of evidence*. The point of taking the logarithm is, of course, that if you have several experiments about two simple hypotheses, then the Bayes factors multiply, and so the weight of evidence adds.

#### 4.1.5 Example

According to Watkins (1986, Section 13.3), the electroweak theory predicted the existence of a new particle, the W particle, of a mass  $m$  of  $82.4 \pm 1.1$  GeV. Experimental results showed that such a particle existed and had a mass of  $82.1 \pm 1.7$  GeV. If we take the mass to have a normal prior and likelihood and assume that the values after the  $\pm$  signs represent known standard deviations,

and if we are prepared to take both the theory and the experiment into account, then we can conclude that the posterior for the mass is  $N(\theta_1, \phi_1)$  where  $\phi_1 = (1.1^{-2} + 1.7^{-2})^{-1} = 0.853 = 0.92^2$

$$\theta_1 = 0.853(82.4/1.1^2 + 82.1/1.7^2) = 82.3$$

(following the procedure of Section 2.2 on ‘Normal Prior and Likelihood’). Suppose that for some reason it was important to know whether or not this mass was less than 83.0 GeV. Then, since the prior distribution is  $N(82.4, 1.1^2)$ , the prior probability  $\pi_0$  of this hypothesis is given by  $\pi_0 = P(m \leq 83.0) = \Phi((83.0 - 82.4)/1.1) = \Phi(0.55)$ ,

where  $\Phi$  is the distribution function of the standard normal distribution. From tables of the normal distribution, it follows that  $\pi_0 \cong 0.7088$ , so that the prior odds are  $\pi_0/(1 - \pi_0) \cong 2.43$ .

Similarly, the posterior probability of the hypothesis that  $m \leq 83.0$  is  $p_0 = \Phi((83.0 - 82.3)/0.92) = \Phi(0.76) = 0.7764$ , and hence the posterior odds are  $p_0/(1 - p_0) \cong 3.47$ .

Thus, the Bayes factor is

$$B = \frac{(p_0/p_1)}{(\pi_0/\pi_1)} = \frac{p_0\pi_1}{p_1\pi_0} = \frac{3.47}{2.43} = 1.43.$$

In this case, the experiment has not much altered beliefs about the hypothesis under discussion, and this is represented by the nearness of  $B$  to 1.

### 4.1.6 Comment

A point about hypothesis tests well worth making is that they ‘are traditionally used as a method for testing between two terminal acts [but that] in *actual practice* [they] are far more commonly used [when we are] *given* the outcome of a sample [to decide whether] any final or terminal decision [should] be reached or should judgement be suspended until more sample evidence is available’ (Schlaifer, 1961, Section 13.2).

## 4.2 One-sided hypothesis tests

### 4.2.1 Definition

A hypothesis testing situation of the type described in Section 4.1 is said to be *one-sided* if the set  $\Theta$  of possible values of the parameter  $\theta$  is the set of real numbers or a subset of it and either  $\theta_0 < \theta_1$  whenever  $\theta_0 \in \Theta_0$  and  $\theta_1 \in \Theta_1$  or

$$\theta_0 > \theta_1 \quad \text{whenever} \quad \theta_0 \in \Theta_0 \text{ and } \theta_1 \in \Theta_1.$$

From the Bayesian point of view, there is nothing particularly special about this situation. The interesting point is that this is one of the few situations in which classical results, and in particular the use of  $P$ -values, has a Bayesian justification.

### 4.2.2 $P$ -values

This is one of the places where it helps to use the ‘tilde’ notation to emphasize which quantities are random. If  $\tilde{x} \sim N(\theta, \phi)$  where  $\phi$  is known and the reference prior  $p(\theta) \propto 1$  is used, then the posterior distribution of  $\theta$  given  $\tilde{x} = x$  is  $N(x, \phi)$ . Consider now the situation in which we wish to test  $H_0 : \theta \leq \theta_0$  versus  $H_1 : \theta > \theta_0$ .

$$p_0 = P(\tilde{\theta} \leq \theta_0 | \tilde{x} = x)$$

Then, if we observe that  $\tilde{x} = x$  we have a posterior probability  $= \Phi((\theta_0 - x)/\sqrt{\phi})$ .

Now the classical  $P$ -value (sometimes called the *exact significance level*) against  $H_0$  is defined as the probability, when  $\theta = \theta_0$ , of observing an  $\tilde{x}$  ‘at least as

$$\begin{aligned} P\text{-value} &= P(\tilde{x} \geq x | \theta = \theta_0) \\ &= 1 - \Phi((x - \theta_0)/\sqrt{\phi}) \\ &= \Phi((\theta_0 - x)/\sqrt{\phi}) \end{aligned}$$

extreme’ as the actual data  $x$  and so is

$$= p_0.$$

For example, if we observe a value of  $x$  which is 1.5 standard deviations above  $\theta_0$  then a Bayesian using the reference prior would conclude that the posterior probability of the null hypothesis is  $\Phi(-1.5) = 0.0668$ , whereas a classical statistician would report a  $P$ -value of 0.0668. Of course  $p_1 = 1 - p_0 = 1 - P\text{-value}$ , so

$$\text{the posterior odds are } \frac{p_0}{p_1} = \frac{p_0}{1 - p_0} = \frac{P\text{-value}}{1 - P\text{-value}}.$$

In such a case, the prior distribution could perhaps be said to imply prior odds of 1 (but beware! – this comes from taking  $\infty/\infty = 1$ ), and so we get a Bayes factor

$$\text{of } B = \frac{p_0}{1 - p_0} = \frac{P\text{-value}}{1 - P\text{-value}},$$

implying that

$$p_0 = P\text{-value} = B/(1 + B) = (1 + B^{-1})^{-1}$$

$$p_1 = 1/(1 + B).$$

On the other hand, the classical probabilities of Type I and Type II errors do *not* have any close correspondence to the probabilities of hypotheses, and to that extent the increasing tendency of classical statisticians to quote  $P$ -values rather than just the probabilities of Type I and Type II errors is to be welcomed, even though a full Bayesian analysis would be better.

A *partial* interpretation of the traditional use of the probability of a Type I error (sometimes called a *significance level*) is as follows. A result is significant at level  $\alpha$  if and only if the  $P$ -value is less than or equal to  $\alpha$ , and hence if and only if the posterior probability  $p_0 = P(\tilde{\theta} \leq \theta_0 | \tilde{x} = x) \leq \alpha$

or equivalently

$$p_1 = P(\tilde{\theta} > \theta_0 | \tilde{x} = x) \geq 1 - \alpha.$$

## 4.3 Lindley's method

### 4.3.1 A compromise with classical statistics

The following method appears first to have been suggested by Lindley (1965, Section 5.6), and has since been advocated by a few other authors, for example, Zellner (1971, Section 10.2; 1974, Section 3.7).

Suppose, as is common in classical statistics, that you wish to conduct a test of a point (or sharp) null hypothesis

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta \neq \theta_0.$$

Suppose further that your prior knowledge is *vague* or *diffuse*, so that you have no particular reason to believe that  $\theta = \theta_0$  rather than that  $\theta = \theta_1$  where  $\theta_1$  is any value in the neighbourhood of  $\theta_0$ .

The suggested procedure depends on finding the posterior distribution of  $\theta$  using a reference prior. To conduct a significance test at level  $\alpha$ , it is then suggested that you find a  $100(1 - \alpha)\%$  highest density region (HDR) from the posterior distribution and reject  $H_0 : \theta = \theta_0$  if and only if  $\theta_0$  is *outside* this HDR.

### 4.3.2 Example

With the data on the uterine weight of rats which you met in Section 2.8 on ‘HDRs for the normal variance’, we found the posterior distribution of the variance  $\phi$  to be  $\phi \sim 664\chi_{20}^{-2}$ ,

so that an interval corresponding to a 95% HDR for  $\log \chi^2$  is (19, 67). Consequently, on the basis of the data, there you should reject a null hypothesis  $H_0 : \phi = 16$  at the 5% level, but on the other hand, you should not reject a null hypothesis  $H_0 : \phi = 20$  at that level.

### 4.3.3 Discussion

This procedure is appropriate only when prior information is vague or diffuse and even then it is not often the best way of summarizing posterior beliefs; clearly the significance level is a very incomplete expression of these beliefs. For many problems, including the one considered in the above example, I think that this method is to be seen as mainly of historical interest in that it gave a way of arriving at results related to those in classical statistics and thus helped to wean statisticians brought up on these methods towards the Bayesian approach as one which can get results like these as special cases, as well as having its own distinctive conclusions. However, it can have a use in situations where there are several unknown parameters and the complete posterior is difficult to describe or take in. Thus, when we come to consider the analysis of variance in Sections 6.5 and 6.6, we shall use the significance level as described in this section to give some idea of the size of the treatment effect.

## 4.4 Point (or sharp) null hypotheses with prior information

### 4.4.1 When are point null hypotheses reasonable?

As was mentioned in Section 4.3, it is very common in classical statistics to conduct a test of a point (or sharp) null hypothesis  $H_0 : \theta = \theta_0$  versus  $H_1 : \theta \neq \theta_0$ .

In such a case, the full-scale Bayesian approach (as opposed to the compromise described in the previous section) gives rise to conclusions which differ *radically* from the classical answers.

Before getting on to the answers, a few basic comments about the whole problem are in order. First, tests of point null hypotheses are often performed in inappropriate circumstances. It will virtually never be the case that one seriously entertains the hypothesis that  $\theta = \theta_0$  exactly, a point which classical statisticians fully admit (cf. Lehmann, 1986, Sections 4.5, 5.2). More reasonable would be the null hypothesis  $H_0 : \theta \in \Theta_0 = (\theta_0 - \varepsilon, \theta_0 + \varepsilon)$ ,

where  $\varepsilon > 0$  is so chosen that all  $\theta \in \Theta_0$  can be considered ‘indistinguishable’ from  $\theta_0$ . An example in which this might arise would be an attempt to analyze a chemical by observing some aspect, described by a parameter  $\theta$ , of its reaction with a known chemical. If it were desired to test whether or not the unknown chemical was a specific compound, with a reaction strength  $\theta_0$  known to an

accuracy of  $\varepsilon$ , it would be reasonable to test  $H_0 : \theta \in \Theta_0 = (\theta_0 - \varepsilon, \theta_0 + \varepsilon)$  versus  $H_1 : \theta \notin \Theta_0$ .

An example where  $\varepsilon$  might be extremely close to zero is a test for extra-sensory perception (ESP) with  $\theta_0$  representing the hypothesis of no ESP. (The only reason that  $\varepsilon$  would probably not be zero here is that an experiment designed to test for ESP probably would not lead to a perfectly well-defined  $\theta_0$ .) Of course, there are also many decision problems that would lead to a null hypothesis of the aforementioned form with a large  $\varepsilon$ , but such problems will rarely be well approximated by testing a point null hypothesis.

The question arises, if a realistic null hypothesis is  $H_0 : \theta \in \Theta_0 = (\theta_0 - \varepsilon, \theta_0 + \varepsilon)$ , when is it reasonable to approximate it by  $H_0 : \theta = \theta_0$ ? From a Bayesian viewpoint, it will be reasonable if and only when we spread the quantity  $p_0$  of prior probability over  $\Theta_0$ , the posterior probability  $P(\theta \in \Theta_0)$  is close to that of  $\theta_0$  when a lump of prior probability  $p_0$  is concentrated on the single value  $\theta_0$ . This will certainly happen if the likelihood function is approximately constant on  $\Theta_0$ , but this is a very strong condition, and one can often get away with less.

#### 4.4.2 A case of nearly constant likelihood

Suppose that  $x_1, x_2, \dots, x_n$  are independently  $N(\theta, \phi)$  where  $\phi$  is known. Then we know from Section 2.3 on ‘Several normal observations with a normal prior’ that the likelihood is proportional to an  $N(\bar{x}, \phi/n)$  density for  $\theta$ . Now over the interval  $H_0 : \theta \in \Theta_0 = (\theta_0 - \varepsilon, \theta_0 + \varepsilon)$  this likelihood varies by a factor  $\frac{(2\pi\phi/n)^{-\frac{1}{2}} \exp[-\frac{1}{2}(\bar{x} - (\theta_0 + \varepsilon))^2/(\phi/n)]}{(2\pi\phi/n)^{-\frac{1}{2}} \exp[-\frac{1}{2}(\bar{x} - (\theta_0 - \varepsilon))^2/(\phi/n)]} = \exp\left(\frac{2\varepsilon n}{\phi}(\bar{x} - \theta_0)\right)$ .

It follows that if we define  $z$  to be the statistic

$$z = \frac{|\bar{x} - \theta_0|}{\sqrt{(\phi/n)}}$$

used in classical tests of significance, and

$$k \geq \frac{\varepsilon}{\sqrt{(\phi/n)}} z,$$

then the likelihood varies over  $\Theta_0$  by a factor which is at most  $\exp(2k)$ . Hence, provided that  $\varepsilon$  is reasonably small, there is a useful bound on the variation of the likelihood.

For example, if  $\varepsilon$  can be taken to be 0.0025 and

$$k = \frac{0.0025}{\sqrt{(\phi/n)}} z$$

then the likelihood varies by at most  $\exp(2k)$  over  $\Theta_0 = (\theta_0 - \varepsilon, \theta_0 + \varepsilon)$ . More

specifically, if  $z = 2$ ,  $\phi = 1$  and  $n = 25$ , then  $k$  becomes  $k = (0.0025 \times 5)(2) = 0.025$  and  $\exp(2k) = 1.05 = 1/0.95$ . In summary, if all values within  $\pm 0.0025$  of  $\theta_0$  are regarded as indistinguishable from  $\theta_0$ , then we can feel reassured that the likelihood function does not vary by more than 5% over this range of indistinguishable values, and if the interval can be made even smaller then the likelihood is even nearer to being constant.

Note that the bound depends on  $|\bar{x} - \theta_0|$  as well as on  $\phi/n$ .

#### 4.4.3 The Bayesian method for point null hypotheses

We shall now develop a theory for testing point null hypotheses, which can then be compared with the classical theory. If there is doubt as to the adequacy of the point null hypothesis as a representation of the real null hypothesis, it is always possible to test an interval null hypothesis directly by Bayesian methods and compare the results (and this will generally be easier than checking the constancy of the likelihood function).

You cannot use a continuous prior density to conduct a test of  $H_0 : \theta = \theta_0$  because that would of necessity give  $\theta = \theta_0$  a prior probability of zero and hence a posterior probability of zero. A reasonable way of proceeding is to give  $\theta = \theta_0$  a prior probability of  $\pi_0 > 0$  and to assign a probability density  $\pi_1 \rho_1(\theta)$  to values  $\theta \neq \theta_0$  where  $\pi_1 = 1 - \pi_0$  and  $\rho_1(\theta)$  integrates to unity. If you are thinking of the hypothesis  $\theta = \theta_0$  as an approximation to a hypothesis  $\theta \in (\theta_0 - \varepsilon, \theta_0 + \varepsilon)$  then  $\pi_0$  is really your prior probability for the whole interval  $\theta \in (\theta_0 - \varepsilon, \theta_0 + \varepsilon)$ .

You can then derive the predictive density  $p(x)$  of a vector  $x = (x_1, x_2, \dots, x_n)$  of observations in the form

Writing

$$p_1(x) = \int \rho_1(\theta) p(x|\theta) d\theta$$

for what might be called the predictive distribution under the alternative hypothesis we see that

$$p(x) = \pi_0 p(x|\theta_0) + \pi_1 p_1(x).$$

It follows that the posterior probabilities are

$$p_0 = \frac{\pi_0 p(x|\theta_0)}{\pi_0 p(x|\theta_0) + \pi_1 p_1(x)} = \frac{\pi_0 p(x|\theta_0)}{p(x)}$$

$$p_1 = \frac{\pi_1 p_1(x)}{\pi_0 p(x|\theta_0) + \pi_1 p_1(x)} = \frac{\pi_1 p_1(x)}{p(x)},$$

and so, the Bayes factor is

$$B = \frac{(p_0/p_1)}{(\pi_0/\pi_1)} = \frac{p(x|\theta_0)}{p_1(x)}.$$

Of course, it is possible to find the posterior probabilities  $p_0$  and  $p_1$  in terms of the Bayes factor  $B$  and the prior probability  $\pi_0$  as noted in Section 4.1 when hypothesis testing in general was discussed.

#### 4.4.4 Sufficient statistics

Sometimes, we have a sufficient statistic  $t = t(x)$  for  $x$  given  $\theta$ , so that  $p(x|\theta) = p(t|\theta)p(x|t)$ ,

where  $p(x|t)$  is not a function of  $\theta$ . Clearly in such a case,

$$\begin{aligned} p_1(x) &= \int \rho_1(\theta)p(t|\theta)p(x|t)d\theta \\ &= \left( \int \rho_1(\theta)p(t|\theta)d\theta \right) p(x|t) \\ &= p_1(t)p(x|t), \end{aligned}$$

so that we can cancel a common factor  $p(x|t)$  to get

$$p_0 = \frac{\pi_0 p(t|\theta_0)}{\pi_0 p(t|\theta_0) + \pi_1 p_1(t)} = \frac{\pi_0 p(t|\theta_0)}{p(t)},$$

$$p_1 = \frac{\pi_1 p_1(t)}{\pi_0 p(t|\theta_0) + \pi_1 p_1(t)} = \frac{\pi_1 p_1(t)}{p(t)},$$

and the Bayes factor is

$$B = \frac{(p_0/p_1)}{(\pi_0/\pi_1)} = \frac{p(t|\theta_0)}{p_1(t)}.$$

In short,  $x$  can be replaced by  $t$  in the formulas for  $p_0$ ,  $p_1$  and the Bayes factor  $B$ .

Many of the ideas in this section should become clearer when you come to look at Section 4.5, in which the particular case of the normal mean is explored in detail.

### 4.5 Point null hypotheses for the normal distribution

#### 4.5.1 Calculation of the Bayes' factor

Suppose  $x = (x_1, x_2, \dots, x_n)$  is a vector of independently  $N(\theta, \phi)$  random variables, and that  $\phi$  is known. Because of the remarks at the end of the last section, we can work entirely in terms of the sufficient statistic  $\bar{x} \sim N(\theta, \phi/n)$ .

We have to make some assumption about the density  $\rho_1(\theta)$  of  $\theta$  under the

alternative hypothesis, and clearly one of the most natural things to do is to suppose that this density is normal, say  $N(\mu, \psi)$ . Strictly, this should be regarded as a density on values of  $\theta$  other than  $\theta_0$ , but when probabilities are found by integration of this density, the odd point will make no difference. It will usually seem sensible to take  $\mu = \theta_0$  as, presumably, values near to  $\theta_0$  are more likely than those far away, and this assumption will accordingly be made from now on. We note that the standard deviation  $\sqrt{\psi}$  of the density of  $\theta$  under the alternative hypothesis is supposed to be considerably greater than the width  $2\varepsilon$  of the interval of values of  $\theta$  considered ‘indistinguishable’ from  $\theta_0$ .

It is quite easy to find the predictive distribution  $p_1(\bar{x})$  of  $\bar{x}$  under the alternative, namely,

$$p_1(\bar{x}) = \int p_1(\theta) p(\bar{x}|\theta) d\theta,$$

by writing

$$\bar{x} = (\bar{x} - \theta) + \theta$$

as in Section 2.2 on ‘Normal prior and likelihood’. Then because, independently of one another,  $\bar{x} - \theta \sim N(0, \phi/n)$  and  $\theta \sim N(\theta_0, \psi)$ , the required density of  $\bar{x}$  is  $N(\theta_0, \psi + \phi/n)$ .

It follows that the Bayes factor  $B$  is

$$\begin{aligned} B &= \frac{p(\bar{x}|\theta_0)}{p_1(\bar{x})} \\ &= \frac{(2\pi\phi/n)^{-\frac{1}{2}} \exp[-\frac{1}{2}(\bar{x} - \theta_0)^2/(\phi/n)]}{(2\pi(\psi + \phi/n))^{-\frac{1}{2}} \exp[-\frac{1}{2}(\bar{x} - \theta_0)^2/(\psi + \phi/n)]} \\ &= [1 + n\psi/\phi]^{\frac{1}{2}} \exp[-\frac{1}{2}(\bar{x} - \theta_0)^2\phi^{-1}n\{1 + \phi/n\psi\}^{-1}]. \end{aligned}$$

It is now useful to write

$$z = |\bar{x} - \theta_0|/\sqrt{(\phi/n)}$$

for the statistic used in classical tests of significance. With this definition

$$B = [1 + n\psi/\phi]^{\frac{1}{2}} \exp[-\frac{1}{2}z^2\{1 + \phi/n\psi\}^{-1}].$$

The posterior probability  $p_0$  can now be found in terms of the prior probability  $\pi_0$  and the Bayes factor  $B$  by the usual formula

$$p_0 = \frac{1}{[1 + (\pi_1/\pi_0)B^{-1}]} = \frac{1}{[1 + \{(1 - \pi_0)/\pi_0\}B^{-1}]}$$

derived in Section 4.1 when we first met hypothesis tests.

## 4.5.2 Numerical examples

For example, if  $\pi_0 = \frac{1}{2}$  and  $\psi = \phi$ , then the values  $z = 1.96$ ,  $n = 15$  give rise to a Bayes factor

$$B = \{1 + 15\}^{\frac{1}{2}} \exp[-\frac{1}{2}(1.96)^2 \{1 + \frac{1}{15}\}^{-1}] \\ = 0.66$$

and hence to a posterior probability

$$p_0 = [1 + 0.66^{-1}]^{-1} \\ = 0.4.$$

This result is quite extraordinarily different from the conclusion that a classical statistician would arrive at with the same data. Such a person would say that, since  $z$  has a sampling distribution that is  $N(0, 1)$ , a value of  $z$  that is, in modulus, 1.96 or greater would arise with probability only 5% (i.e. the two-tailed  $P$ -value of  $z=1.96$  is 0.05), and consequently would reject the null hypothesis that  $\theta = \theta_0$  at the 5% level. With the above assumptions about prior beliefs, we have, on the contrary, arrived at a posterior probability of 40% that the null hypothesis is true! Some further sample values are as follows (cf.

Berger, 1985, Section 4.3):

$P$ -value (2-tailed)	$z \setminus n$	1985,						Section 4.3):
		1	5	10	20	50	100	
0.1	1.645	0.418	0.442	0.492	0.558	0.655	0.725	0.891
0.05	1.960	0.351	0.331	0.367	0.424	0.521	0.600	0.823
0.01	2.576	0.212	0.134	0.140	0.163	0.216	0.273	0.535
0.001	3.291	0.086	0.026	0.024	0.026	0.034	0.045	0.124

The results of classical and Bayesian analyses differ more and more as the sample size  $n \rightarrow \infty$ . For fixed  $z$ , it is easy to see that  $B$  is asymptotically  $B \approx (n\psi/\phi)^{\frac{1}{2}} \exp[-\frac{1}{2}z^2]$

and hence  $B \rightarrow \infty$ . Consequently,  $1-p_0$  is of order  $1/\sqrt{n}$  and thus  $p_0 \rightarrow 1$ . So, with the specified prior, the result that  $z=1.96$ , which a classical statistician would regard as just sufficient to result in rejection of the null hypothesis at the 5% level irrespective of the value of  $n$ , can result in an arbitrarily high posterior probability  $p_0$  of the null hypothesis. Despite this, beginning users of statistical techniques often get the impression that if some data are significant at the 5% level then in some sense the null hypothesis has a probability after the event of at most 5%.

A specific example of a problem with a large sample size arises in connection with Weldon's dice data, quoted by Fisher (1925b, Section 18 and Section 23). It transpired that when 12 dice were thrown 26 306 times, the mean and variance of the number of dice showing more than 4 were 4.0524 and 2.6983, as compared with a theoretical mean of  $12 \times \frac{1}{3} = 4$  for fair dice. Approximating the binomial distribution by a normal distribution leads to a  $z$  statistic of  $z = (4.0524 - 4)/\sqrt{(2.6983/26306)} = 5.17$ .

The corresponding two-tailed  $P$ -value is approximately  $2\phi(z)/z$  where  $\phi$  is the

density function of the standard normal distribution (cf. Abramowitz and Stegun, 1965, equation 26.2.12), so about 1 in 4 000 000. However, a Bayesian analysis (assuming  $\psi = \phi$  and  $\theta_0 = \frac{1}{2}$  as usual) depends on a Bayes factor

$$B = (1 + 26306)^{\frac{1}{2}} \exp[-\frac{1}{2}(5.17)^2 \{1 + (26306)^{-1}\}^{-1}] \\ = 0.00025$$

and so to a posterior probability of 1 in 4000 that the dice were fair. This is small, but nevertheless the conclusion is not as startling as that which the classical analysis leads to.

### 4.5.3 Lindley's paradox

This result is sometimes known as *Lindley's paradox* (cf. Bartlett, 1957; Lindley, 1957; Shafer, 1982) and sometimes as *Jeffreys' paradox*, because it was in essence known to Jeffreys (see Jeffreys, 1961, Section 5.2), although he did not refer to it as a paradox. A useful recent reference is Berger and Delempady (1987).

It does relate to something which has been noted by users of statistics. Lindley once pointed out (see Zellner, 1974, Section 3.7) that experienced workers often lament that for large sample sizes, say 5000, as encountered in survey data, use of the usual t-statistic and the 5% significance level shows that the values of parameters are usually different from zero and that many of them sense that with such a large sample the 5% level is not the right thing to use, but do not know what else to use (see also Jeffreys, 1961, Appendix B). On the other hand, in many scientific contexts, it is unrealistic to use a very large sample because systematic bias may vitiate it or the observer may tire [see Wilson (1952, Section 9.6) or Baird (1962, Section 2.8)].

Since the result is so different from that found by so many statisticians, it is important to check that it does not depend very precisely on the nature of the prior distribution which led to it.

We assumed that the prior probability  $\pi_0$  of the null hypothesis was  $\frac{1}{2}$ , and this assumption does seem ‘natural’ and could be said to be ‘objective’; in any case a slight change in the value of  $\pi_0$  would not make much difference to the qualitative feel of the results.

We also assumed that the prior density of  $\theta_0$  under the alternative hypothesis was normal of mean  $\theta_0$  with some variance  $\psi$ . In fact, the precise choice of  $\rho_1(\theta)$  does not make a great deal of difference unless  $|\bar{x} - \theta_0|$  is *large*. Lindley (1957) took  $\rho_1(\theta)$  to be a uniform distribution  $U(\theta_0 - \frac{1}{2}\sqrt{\psi}, \theta_0 + \frac{1}{2}\sqrt{\psi})$

over an interval centred on  $\theta_0$ , while Jeffreys (1961, Section 5.2) argues that it should be a Cauchy distribution, that is,  $\rho_1(\theta) = \frac{1}{\pi} \frac{\sqrt{\psi}}{\psi + (\theta - \theta_0)^2}$  although his arguments are far from overwhelming and do not seem to have convinced anyone else. An examination of their work will show that in general terms they arrive at similar conclusions to those derived earlier.

There is also a scale parameter  $\psi$  in the distribution  $\rho_1(\theta)$  to be decided on (and this is true whether this distribution is normal, uniform or Cauchy). Although it seems reasonable that  $\psi$  should be chosen proportional to  $\phi$ , there does not seem to be any convincing argument for choosing this to have any particular value (although Jeffreys tries to give a rational argument for the Cauchy form in general, he seems to have no argument for the choice of  $\psi$  beyond saying that it should be proportional to  $\phi$ ). But it is easily seen that the effect of taking  $\psi = k\phi$

on  $B$  and  $p_0$  is just the same as taking  $\psi = \phi$  if  $n$  is multiplied by a factor  $k$ . It should be noted that it will not do to let  $\psi \rightarrow \infty$  and thus to take  $\rho_1(\theta)$  as a uniform distribution on the whole real line, because this is equivalent to multiplying  $n$  by a factor which tends to  $\infty$  and so leads to  $B \rightarrow \infty$  and  $p_0 \rightarrow 1$ . It would clearly not be sensible to use a procedure which always gave the null hypothesis a posterior value of unity. In any case, as Jeffreys points out (1961, Section 5.0), ‘the mere fact that it has been suggested that  $[\theta]$  is zero corresponds to some presumption that it is fairly small’.

#### 4.5.4 A bound which does not depend on the prior distribution

In fact, it is possible to give a bound on  $B$  which does not depend on any

$$p_1(\bar{x}) = \int \rho_1(\theta) p(\bar{x}|\theta) d\theta$$

assumptions about  $\rho_1(\theta)$ . We know that

$$\leq p(\bar{x}|\hat{\theta}),$$

where  $\hat{\theta}$  is the *maximum likelihood* estimator of  $\theta$ , that is,

$$p(\bar{x}|\hat{\theta}) = \sup_{\theta} p(\bar{x}|\theta).$$

In the case being considered,  $\bar{x}$  has a normal distribution of mean  $\theta$  and hence  $\hat{\theta} = \bar{x}$ , so that  $p_1(\bar{x}) \leq p(\bar{x}|\bar{x}) = (2\pi\phi/n)^{-\frac{1}{2}}$ .

It follows that the Bayes factor satisfies

$$B = \frac{p(\bar{x}|\theta_0)}{p_1(\bar{x})} \geq \frac{\{2\pi\phi/n\}^{-\frac{1}{2}} \exp[-\frac{1}{2}(\bar{x} - \theta_0)^2/(\phi/n)]}{\{2\pi\phi/n\}^{-\frac{1}{2}}}$$

so writing  $z = |\bar{x} - \theta_0|/\sqrt{(\phi/n)}$  as before, we see that

$$B \geq \exp(-\frac{1}{2}z^2)$$

implying a corresponding lower bound on  $p_0$ . Some sample values (assuming

P-value (2-tailed)	$z$	Bound on $B$	Bound on $p_0$
0.1	1.645	0.258	0.205
0.05	1.960	0.146	0.128
0.01	2.576	0.036	0.035
that $\pi_0 = \frac{1}{2}$ ) are as follows:	0.001	3.291	0.004

[cf. Berger, 1985, Section 4.3; Berger further claims that if  $\pi_0 = \frac{1}{2}$  and  $z > 1.68$  then  $p_0 \geq (P\text{-value}) \times (1.25 z)$ ]. Note that this bound does not depend on the sample size  $n$  and so does not demonstrate Lindley's paradox.

As an example, if  $z=1.96$  then the Bayes factor  $B$  is *at least* 0.146 and hence the posterior probability of the null hypothesis is *at least* 0.128. Unlike the results derived earlier assuming a more precise form for  $\rho_1(\theta)$ , the bounds no longer depend on the sample size, but it should be noted that the conclusion still does not accord at all well with the classical result of significance at the 5% level.

#### 4.5.5 The case of an unknown variance

In the case where  $\phi$  is unknown, similar conclusions follow, although there are a few more complications. It will do now harm if the rest of this section is ignored at a first reading (or even at a second).

We need first to find the density  $p(x|\theta_0)$ . If  $\phi$  is unknown, then as was shown in Section 2.12 on 'Normal mean and variance both unknown'

$$p(x|\theta_0, \phi) \propto \phi^{-n/2} \exp\left[-\frac{1}{2}\{S + n(\bar{x} - \theta_0)^2\}/\phi\right],$$

where  $S = \sum(x_i - \bar{x})^2$ . Using a reference prior  $p(\phi) \propto 1/\phi$  for  $\phi$ , it is easy to integrate  $\phi$  out much as was done there to get

$$p(x|\theta_0) = \int p(x, \phi|\theta_0) d\phi = \int p(\phi)p(x|\theta_0, \phi) d\phi$$

$$\propto \int \phi^{-n/2} \exp\left[-\frac{1}{2}\{S + n(\bar{x} - \theta_0)^2\}/\phi\right] d\phi$$

$$\propto \{1 + t^2/v\}^{-(v+1)/2},$$

where  $v = n - 1$  and

$$t = \frac{\theta - \bar{x}}{s/\sqrt{n}},$$

$$s^2 = S/v.$$

It is now necessary to find the predictive density  $p_1(x)$  under the alternative

hypothesis. To do this, first return to  $p(x|\theta, \phi) \propto \phi^{-n/2} \exp[-\frac{1}{2}\{S + n(\bar{x} - \theta)^2\}/\phi]$ .

Assuming a prior  $\theta \sim N(\theta_0, \psi)$ , we can integrate  $\theta$  out; thus,

$$\begin{aligned}
p_1(x|\phi) &= \int p(x, \theta|\phi) d\theta = \int p(\theta)p(x|\theta, \phi) d\theta \\
&\propto \int \phi^{-n/2} \psi^{-\frac{1}{2}} \exp \left[ -\frac{1}{2} \left( \frac{S}{\phi} + \frac{n}{\phi} (\bar{x} - \theta)^2 + \frac{1}{\psi} (\theta - \theta_0)^2 \right) \right] d\theta \\
&\propto \phi^{-n/2} \psi^{-\frac{1}{2}} \exp \left[ -\frac{1}{2} \left( \frac{S}{\phi} + \frac{n\bar{x}^2}{\phi} + \frac{\theta_0^2}{\psi} - \frac{(n\bar{x}/\phi + \theta_0/\psi)^2}{n/\phi + 1/\psi} \right) \right] \\
&\times \int \exp \left[ -\frac{1}{2} \left( \frac{n}{\phi} + \frac{1}{\psi} \right) \left( \theta - \frac{n\bar{x}/\phi + \theta_0/\psi}{n/\phi + 1/\psi} \right)^2 \right] d\theta.
\end{aligned}$$

The last integral is of course proportional to  $(n/\phi + 1/\psi)^{-\frac{1}{2}}$ , so to  $(1 + n\psi/\phi)^{-\frac{1}{2}}$ ,

$$\begin{aligned} \frac{n\bar{x}^2}{\phi} + \frac{\theta_0^2}{\psi} - \frac{(n\bar{x}/\phi + \theta_0/\psi)^2}{n/\phi + 1/\psi} &= \frac{(n/\phi)(1/\psi)}{n/\phi + 1/\psi} (\bar{x} - \theta_0)^2 \\ &= \frac{1}{1 + n\psi/\phi} \frac{n(\bar{x} - \theta_0)^2}{\phi}. \end{aligned}$$

while a little manipulation shows that

It follows that

$$p_1(x|\phi) \propto \phi^{-n/2} \left(1 + \frac{n\psi}{\phi}\right)^{-\frac{1}{2}} \exp \left[ -\frac{1}{2} \left( S + \frac{n(\bar{x} - \theta_0)^2}{1+n\psi/\phi} \right) \phi \right].$$

To go any further, it is necessary to make some assumption about the relationship between  $\phi$  and  $\psi$ . If it is assumed that  $\psi = k\phi$

and a reference prior  $p(\phi) \propto 1/\phi$  is used, then the predictive distribution under the alternative hypothesis becomes

$$\begin{aligned}
p_1(x) &= \int p(x, \phi) d\phi = \int p(\phi)p(x|\phi) d\phi \\
&\propto (1+nk)^{-\frac{1}{2}} \int \phi^{-(v+1)/2-1} \exp\left[-\frac{1}{2}\left(S + \frac{n(\bar{x} - \theta_0)^2}{1+nk}\right)\phi\right] d\phi \\
&\propto (1+nk)^{-\frac{1}{2}} \left\{1 + \frac{1}{v} \frac{t^2}{1+nk}\right\}^{-(v+1)/2},
\end{aligned}$$

where  $t$  is the same statistic encountered in the case of the null hypothesis. It

follows that the Bayes factor is  $\frac{(1+t^2/v)^{-(v+1)/2}}{(1+nk)^{-\frac{1}{2}}(1+t^2(1+nk)^{-1}/v)^{-(v+1)/2}}$ , and hence it is possible to find  $p_1$  and  $p_2$

It should be noted that as  $n \rightarrow \infty$  the exponential limit shows that the Bayes factor is asymptotically  $\frac{\exp[-\frac{1}{2}t^2]}{(1+nk)^{-\frac{1}{2}}\exp[-\frac{1}{2}t^2(1+nk)^{-1}]} = (1+nk)^{\frac{1}{2}}\exp[-\frac{1}{2}t^2nk/(1+nk)]$  which as  $t \cong z$  is the same as in the known variance case.

## 4.6 The Doogian philosophy

### 4.6.1 Description of the method

Good (1983, Chapter 4 and elsewhere) has argued in favour of a compromise between Bayesian and non-Bayesian approaches to hypothesis testing. His technique can be summarized as follows (in his own words): The Bayes/non-Bayes synthesis is the following technique for synthesizing subjective and objective techniques in statistics. (i) We use the neo-Bayes/Laplace philosophy (i.e. the techniques described in Section 4.4 on point null hypotheses with prior information) in order to arrive at a factor  $F$  (which is  $1/B$  in the notation used here) in favour of the non-null hypothesis. For the particular case of discrimination between two simple statistical hypotheses, the factor in favour is equal to the likelihood ratio [as was shown in Section 4.1 when hypothesis testing was first considered], but not in general. The neo-Bayes/Laplace philosophy usually works with inequalities between probabilities, but for definiteness we here assume that the initial distributions are taken as precise, though not necessarily uniform. (ii) We then use  $F$  as a statistic and try to obtain its distribution on the null hypothesis, and work out its tail area,  $P$ . (iii) Finally, we look to see if  $F$  lies in the range  $(1/30P, 3/10P)$ . If it does not lie in this range we think again. (Note that  $F$  is here the factor *against*  $H_0$ .) This is certainly not unworkable. For example, in Section 4.5 we found that

$$1/F = B = \{1 + n\psi/\phi\}^{\frac{1}{2}} \exp\left[-\frac{1}{2}z^2(1 + \phi/n\psi)^{-1}\right],$$

so that  $B$  is a monotonic function of  $z^2$ , and hence the probability  $B \geq b$  equals the (two-tailed)  $P$ -value corresponding to the value of  $z$ , which is easily found as  $z$  has a standard normal distribution.

### 4.6.2 Numerical example

Thus if, as in an example discussed in Section 4.5,  $\pi_0 = \frac{1}{2}$  and the density under the alternative hypothesis has  $\psi = \phi$  (and so is  $N(\theta_0, \phi)$ ), then for  $z = 1.96$  and  $n = 15$

the  $P$ -value is  $P=0.05$  and the Bayes factor is  $B=0.66=1/1.5$ . Good's method then asks us to check whether  $F=1.5$  lies in the range  $(1/30P, 3/10P)$ , that is,  $(0.67, 6.0)$ . Consequently, we do not in this case need to 'think again'.

Good attributes this approach to 'the Tibetan lama K. Caj Doog', but it does not appear that the lama has made many converts apart from Good himself.

## 4.7 Exercises on Chapter 4

1. Show that if the prior probability  $\pi_0$  of a hypothesis is close to unity, then the posterior probability  $p_0$  satisfies  $1 - p_0 \cong (1 - \pi_0)B^{-1}$  and more exactly  $1 - p_0 \cong (1 - \pi_0)B^{-1} + (1 - \pi_0)^2(B^{-1} - B^{-2})$ .
2. Watkins (1986, Section 13.3) reports that theory predicted the existence of a Z particle of mass  $93.3 \pm 0.9$  GeV, while first experimental results showed its mass to be  $93.0 \pm 1.8$  GeV. Find the prior and posterior odds and the Bayes ratio for the hypothesis that its mass is less than 93.0 GeV.
3. An experimental station wishes to test whether a growth hormone will increase the yield of wheat above the average value of 100 units per plot produced under currently standard conditions. Twelve plots treated with the hormone give the yields:  
140, 103, 73, 171, 137, 91, 81, 157, 146, 69, 121, 134.  
Find the P-value for the hypothesis under consideration.
4. In a genetic experiment, theory predicts that if two genes are on different chromosomes, then the probability of a certain event will be 3/16. In an actual trial, the event occurs 56 times in 300. Use Lindley's method to decide whether there is enough evidence to reject the hypothesis that the genes are on the same chromosome.
5. With the data in the example in Section 3.4 on 'The Poisson distribution', would it be appropriate to reject the hypothesis that the true mean equalled the prior mean (i.e. that  $\lambda = 3$ )? [Use Lindley's method.]
6. Suppose that the standard test statistic  $z = (\bar{x} - \theta_0)/\sqrt{(\phi/n)}$  takes the value  $z=2.5$  and that the sample size is  $n = 100$ . How close to  $\theta_0$  does a value of  $\theta$  have to be for the value of the normal likelihood function at  $\bar{x}$  to be within 10% of its value at  $\theta = \theta_0$ ?
7. Show that the Bayes factor for a test of a point null hypothesis for the normal distribution (where the prior under the alternative hypothesis is also normal) can be expanded in a power series in  $\lambda = \phi/n\psi$  as  $B = \lambda^{-\frac{1}{2}} \exp(-\frac{1}{2}z^2)\{1 + \frac{1}{2}\lambda(z^2 + 1) + \dots\}$ .
8. Suppose that  $x_1, x_2, \dots, x_n \sim N(0, \phi)$ . Show over the interval  $(\phi - \varepsilon, \phi + \varepsilon)$  the likelihood varies by a factor of approximately  $\exp\left\{\frac{\varepsilon}{\phi} \left(\frac{\sum x_i^2/n}{\phi} - 1\right)\right\}$ .
9. At the beginning of Section 4.5, we saw that under the alternative

hypothesis that  $\theta \sim N(\theta_0, \psi)$  the predictive density for  $\bar{x}$  was  $N(\theta_0, \psi + \phi/n)$ , so that  $p_1(\bar{x}) = \{2\pi(\psi + \phi/n)\}^{-\frac{1}{2}} \exp[-\frac{1}{2}(\bar{x} - \theta_0)^2/(\psi + \phi/n)]$ .

Show that a maximum of this density considered as a function of  $\psi$  occurs when  $\psi = (z^2 - 1)\phi/n$ , which gives a possible value for  $\psi$  if  $z \geq 1$ . Hence, show that if  $z \geq 1$  then for any such alternative hypothesis, the Bayes factor satisfies  $B \geq \sqrt{e} z \exp(-\frac{1}{2}z^2)$

and deduce a bound for  $p_0$  (depending on the value of  $\pi_0$ ).

**10.** In the situation discussed in Section 4.5, for a given  $P$ -value (so equivalently for a given  $z$ ) and assuming that  $\phi = \psi$ , at what value of  $n$  is the posterior probability of the null hypothesis a minimum.

**11.** Mendel (1865) reported finding 1850 angular wrinkled seeds to 5474 round or roundish in an experiment in which his theory predicted a ratio of 1:3. Use the method employed for Weldon's dice data in Section 4.5 to test whether his theory is confirmed by the data. [However, Fisher (1936) cast some doubt on the genuineness of the data.]

**12.** A window is broken in forcing entry to a house. The refractive index of a piece of glass found at the scene of the crime is  $x$ , which is supposed  $N(\theta_1, \phi)$ . The refractive index of a piece of glass found on a suspect is  $y$ , which is supposed  $N(\theta_2, \phi)$ . In the process of establishing the guilt or innocence of the suspect, we are interested in investigating whether  $H_0 : \theta_1 = \theta_2$  is true or not. The prior distributions of  $\theta_1$  and  $\theta_2$  are both  $N(\mu, \psi)$  where  $\psi \gg \phi$ . Write  $u = x - y$ ,  $z = \frac{1}{2}(x + y)$ .

Show that, if  $H_0$  is true and  $\theta_1 = \theta_2 = \theta$ , then  $\theta, x - \theta$  and  $y - \theta$  are independent and  $\theta \sim N(\mu, \psi)$ ,  $x - \theta \sim N(0, \phi)$ ,  $y - \theta \sim N(0, \phi)$ .

By writing  $u = (x - \theta) - (y - \theta)$  and  $z = \theta + \frac{1}{2}(x - \theta) + \frac{1}{2}(y - \theta)$ , go on to show that  $u$  has an  $N(0, 2\phi)$  distribution and that  $z$  has an  $N(\mu, \frac{1}{2}\phi + \psi)$ , so approximately an  $N(\mu, \psi)$ , distribution. Conversely, show that if  $H_0$  is false and  $\theta_1$  and  $\theta_2$  are assumed independent, then  $\theta_1, \theta_2, x - \theta_1$  and  $y - \theta_2$  are all independent and  $\theta_1 \sim N(\mu, \psi)$ ,  $\theta_2 \sim N(\mu, \psi)$ ,  $x - \theta_1 \sim N(0, \phi)$ ,  $y - \theta_2 \sim N(0, \phi)$ .

By writing

$$u = \theta_1 - \theta_2 + (x - \theta_1) - (y - \theta_2)$$

and

$$z = \frac{1}{2}[\theta_1 + \theta_2 + (x - \theta_1) + (y - \theta_2)],$$

show that in this case  $u$  has an  $N(0, 2(\phi + \psi))$ , so approximately an  $N(0, 2\psi)$ , distribution, while  $z$  has an  $N(\mu, \frac{1}{2}(\phi + \psi))$ , so approximately an  $N(\mu, \frac{1}{2}\psi)$ ,

distribution. Conclude that the Bayes factor is approximately  $B = \sqrt{(\psi/2\phi)} \exp[-\frac{1}{2}u^2/2\phi + \frac{1}{2}(z-\mu)^2/\psi]$ .

Suppose that the ratio  $\sqrt{(\psi/\phi)}$  of the standard deviations is 100 and that  $u = 2 \times \sqrt{(2\phi)}$ , so that the difference between  $x$  and  $y$  represents two standard deviations, and that  $z = \mu$ , so that both specimens are of commonly occurring glass. Show that a classical test would reject  $H_0$  at the 5% level, but that  $B=9.57$ , so that the odds in favour of  $H_0$  are multiplied by a factor just below 10.

[This problem is due to Lindley (1977); see also Shafer (1982). Lindley comments that, ‘What the [classical] test fails to take into account is the extraordinary coincidence of  $x$  and  $y$  being so close together were the two pieces of glass truly different’.]

**13.** Lindley (1957) originally discussed his paradox under slightly different assumptions from those made in this book. Follow through the reasoning used in Section 4.5 with  $\rho_1(\theta)$  representing a uniform distribution on the interval  $(\theta_0 - \frac{1}{2}\tau, \theta_0 + \frac{1}{2}\tau)$  to find the corresponding Bayes factor assuming that  $\tau^2 \gg \phi/n$ , so that an  $N(\mu, \phi/n)$  variable lies in this interval with very high probability. Check that your answers are unlikely to disagree with those found in Section 4.5 under the assumption that  $\rho_1(\theta)$  represents a normal density.

**14.** Express in your own words the arguments given by Jeffreys (1961, Section 5.2) in favour of a Cauchy distribution  $\rho_1(\theta) = \frac{1}{\pi} \frac{\sqrt{\psi}}{\psi + (\theta - \theta_0)^2}$  in the problem discussed in the previous question.

**15.** Suppose that  $x$  has a binomial distribution  $B(n, \theta)$  of index  $n$  and parameter  $\theta$ , and that it is desired to test  $H_0 : \theta = \theta_0$  against the alternative hypothesis  $H_1 : \theta \neq \theta_0$ : **a.** Find lower bounds on the posterior probability of  $H_0$  and on the Bayes factor for  $H_0$  versus  $H_1$ , bounds which are valid for any  $\rho_1(\theta)$ .

**b.** If  $n = 20$ ,  $\theta_0 = \frac{1}{2}$  and  $x = 15$  is observed, calculate the (two-tailed)  $P$ -value and the lower bound on the posterior probability when the prior probability  $\pi_0$  of the null hypothesis is  $\frac{1}{2}$ .

**16.** Twelve observations from a normal distribution of mean  $\theta$  and variance  $\phi$  are available, of which the sample mean is 1.2 and the sample variance is 1.1. Compare the Bayes factors in favour of the null hypothesis that  $\theta = \theta_0$ ,

assuming that (a)  $\phi$  is unknown and (b) it is known that  $\phi = 1$ .

**17.** Suppose that in testing a point null hypothesis you find a value of the usual Student's statistic of 2.4 on 8 degrees of freedom. Would the methodology of Section 4.6 require you to 'think again'?

**18.** Which entries in the table in Section 4.5 on 'Point null hypotheses for the normal distribution' would, according to the methodology of Section 4.6, cause you to 'think again'?

# 5

## Two-sample problems

### 5.1 Two-sample problems – both variances unknown

#### 5.1.1 The problem of two normal samples

We now want to consider the situation in which we have independent samples from two normal distributions, namely,

$$x_1, x_2, \dots, x_m \sim N(\lambda, \phi)$$

$$y_1, y_2, \dots, y_n \sim N(\mu, \psi)$$

which are independent of each other, and the quantity really of interest is the posterior distribution of

$$\delta = \lambda - \mu.$$

This problem arises in comparative situations, for example, in comparing the achievement in geometry tests of boy and girl pupils.

#### 5.1.2 Paired comparisons

Before proceeding further, you should be warned against a possible misapplication of the model. If  $m = n$  and each of the  $x$ s is in some sense paired with one of the  $y$ s, say  $x_i$  with  $y_i$ , you should define  $w_i = x_i - y_i$  and then investigate the  $w$ s as a sample

$$w_1, w_2, \dots, w_n \sim N(\delta, \omega)$$

for some  $\omega$ . This is known as the method of *paired comparisons*. It might arise if, for example, the comparison of performance of boys and girls were restricted to pairs of twins of opposite sexes. The reason that such a situation is not to be treated as a two sample problem in the sense described at the start is that there will be an effect common to any pair of twins, so that the observations on the boys and on the girls will not be fully independent. It is a very valuable technique which can often give a *more precise* measurement of an effect, but it

is important to distinguish it from a case where the two samples are independent. There is no particular difficulty in analyzing the results of a paired comparison experiment by the methods described in Chapter I for samples from a single normal distribution.

### 5.1.3 Example of a paired comparison problem

‘Student’ (1908) quotes data due to A. R. Cushny and A. R. Peebles on the extra hours of sleep gained by ten patients using laevo (L) and dextro (D) hyoscyamine hydrobromide, as follows:

Patient $i$	1	2	3	4	5	6	7	8	9	10
Gain $x_i$ with L	+1.9	+0.8	+1.1	+0.1	-0.1	+4.4	+5.5	+1.6	+4.6	+3.4
Gain $y_i$ with D	+0.7	-1.6	-0.2	-1.2	-0.1	+3.4	+3.7	+0.8	0	+2.0
Difference $w_i$	+1.2	+2.4	+1.3	+1.3	0	+1.0	+1.8	+0.8	+4.6	+1.4

[In fact he misidentifies the substances involved – see E. S. Pearson (1990, p. 54), but the method is still well illustrated]. If we are interested in the difference between the effects of the two forms of the drug, we should find the mean  $\bar{w} = 1.58$  and the sample sum of squares  $S = 13.616$ , and hence the sample standard deviation  $s = 1.23$ . Assuming a standard reference prior for  $\delta$  and a variance known to equal  $1.23^2$ , the posterior distribution of the effect  $\delta$  of using the L rather than the D form is  $N(1.58, 1.23^2)$ . We can then use this distribution, for example, to give an HDR for  $\delta$  or to test a hypothesis about  $\delta$  (such as  $H_0 : \delta = 0$  versus  $H_1 : \delta \neq 0$ ) in the ways discussed in previous sections. On the other hand, if we are interested simply in the effect of the L form, then the data about the D form are irrelevant and we can use the same methods on the  $x_i$ . It is straightforward to extend the analysis to allow for a non-trivial prior for  $\delta$  or an unknown variance or both.

### 5.1.4 The case where both variances are known

In the case of the two-sample problem proper, there are three cases that can arise:

- (i)  $\phi$  and  $\psi$  are known;
- (ii) It is known that  $\phi = \psi$  but their common value is unknown;
- (iii)  $\phi$  and  $\psi$  are unknown.

For the rest of this section, we shall restrict ourselves to case (i). It should, however, be noted that it is not really likely that you would know the variances exactly (although you might have some idea from past experience). The main

reason for discussing this case, as in the problem of a single sample from a normal distribution, is that it involves fewer complexities than the case where the variances are unknown.

If  $\lambda$  and  $\mu$  have *independent* reference priors  $p(\lambda) = p(\mu) \propto 1$  then it follows from Section 2.3 on ‘Several normal observations with a normal prior’ that the posterior for  $\lambda$  is  $N(\bar{x}, \phi/m)$ , and similarly the posterior for  $\mu$  is  $N(\bar{y}, \psi/n)$  independently of  $\lambda$ . It follows that  $\delta = \lambda - \mu \sim N(\bar{x} - \bar{y}, \phi/m + \psi/n)$ .

### 5.1.5 Example

The weight gains (in grammes) between the 28th and 84th days of age of  $m = 12$  rats receiving diets with high-protein diets were as follows:

Rat $i$	1	2	3	4	5	6	7	8	9	10	11	12
Weight gain $x_i$	134	146	104	119	124	161	107	83	113	129	97	123

while the weight gains for  $n = 7$  rates on a low-protein diet were

Rat $i$	1	2	3	4	5	6	7
Weight gain $y_i$	70	118	101	85	107	132	94

(cf. Armitage *et al.* 2001, Section 4.4). The sample mean and sum of squared deviations about the mean for the high-protein group are  $\bar{x} = 120$  and 5032, implying a sample variance of  $5032/11 = 457$ . For the low-protein group the mean and sum of squared deviations about the mean are  $\bar{y} = 101$  and 2552, implying a sample variance of  $2552/6 = 425$ . Although the values for the variances are derived from the samples, the method will be illustrated by proceeding *as if* they were known (perhaps from past experience). Then  $m = 12$ ,  $n = 7$ ,  $\bar{x} = 120$ ,  $\bar{y} = 101$ ,  $\phi = 457$ ,  $\psi = 425$

from which it follows that the posterior distribution of the parameter  $\delta$  that measures the effect of using a high-protein rather than a low-protein diet is  $N(120 - 101, 457/12 + 425/7)$ , that is  $N(19, 99)$ .

It is now possible to deduce, for example, that a 90% HDR for  $\delta$  is  $19 \pm 1.6449\sqrt{99}$ , that is, (3, 35). Also, the posterior probability that  $\delta > 0$  is  $\Phi(19/\sqrt{99}) = \Phi(1.91) = 0.9719$  or about 97%. Furthermore, it is possible to conduct a test of the point null hypothesis that  $\delta = 0$ . If the variance of  $\delta$  under the alternative hypothesis is denoted  $\omega$  (rather than  $\psi$  as in Section 4.4 on ‘Point null hypotheses for the normal distribution’ since  $\psi$  now has another meaning), then the Bayes factor is  $B = [1 + \omega/(\phi/m + \psi/n)]^{1/2} \exp[-\frac{1}{2}z^2\{1 + (\phi/m + \psi/n)/\omega\}^{-1}]$ ,

where  $z$  is the standardized normal variable (under the null hypothesis), namely,  $z = (19 - 0)/\sqrt{99} = 1.91$ . It is not wholly clear what value should be used for  $\omega$ . One possibility might be to take  $\omega = \phi + \psi = 457 + 425 = 882$ , and if this is done

$$B = \{1 + 882/99\}^{\frac{1}{2}} \exp \left[ -\frac{1}{2}(1.91)^2 \{1 + 99/882\}^{-1} \right]$$

then  $= 0.91^{\frac{1}{2}} \exp(-1.64) = 0.61.$

If the prior probability of the null hypothesis is taken as  $\pi_0 = \frac{1}{2}$ , then this gives a posterior probability of  $p_0 = (1+0.61^{-1})^{-1} = 0.38$ , so that it has dropped, but not dropped very much.

### 5.1.6 Non-trivial prior information

The method is easily generalized to the case where substantial prior information is available. If the prior for  $\lambda$  is  $N(\lambda_0, \phi_0)$  then the posterior is  $\lambda \sim N(\lambda_1, \phi_1)$ , where (as was shown in Section 2.3 on ‘Several normal observations with a

$$\phi_1 = \{\phi_0^{-1} + (\phi/m)^{-1}\}^{-1}$$

normal prior’)  $\lambda_1 = \phi_1 \{\lambda_0/\phi_0 + \bar{x}/(\phi/m)\}$ .

Similarly, if the prior for  $\mu$  is  $N(\mu_0, \psi_0)$  then the posterior for  $\mu$  is  $N(\mu_1, \psi_1)$ , where  $\psi_1$  and  $\mu_1$  are similarly defined. It follows that  $\delta = \lambda - \mu \sim N(\lambda_1 - \mu_1, \phi_1 + \psi_1)$

and inferences can proceed much as before.

## 5.2 Variances unknown but equal

### 5.2.1 Solution using reference priors

We shall now consider the case where we are interested in  $\delta = \lambda - \mu$  and we have independent vectors  $x = (x_1, x_2, \dots, x_m)$  and  $y = (y_1, y_2, \dots, y_n)$  such that  $x_i \sim N(\lambda, \phi)$  and  $y_i \sim N(\mu, \phi)$ , so that the two samples have a common variance  $\phi$ .

We can proceed much as we did in Section 2.12 on ‘Normal mean and variance both unknown’. Begin by defining

$$S_x = \sum (x_i - \bar{x})^2, \quad S_y = \sum (y_i - \bar{y})^2, \quad S = S_x + S_y, \\ v_x = m - 1, \quad v_y = n - 1, \quad v = v_x + v_y.$$

For the moment, take independent priors uniform in  $\lambda$ ,  $\mu$  and  $\log \phi$ , that is,  $p(\lambda, \mu, \phi) \propto 1/\phi$ .

With this prior, the posterior is

$$\begin{aligned}
p(\lambda, \mu, \phi | x, y) &\propto p(\lambda, \mu, \phi)p(x|\lambda, \phi)p(y|\mu, \phi) \\
&\propto (1/\phi)(2\pi\phi)^{-(m+n)/2} \\
&\quad \times \exp\left[-\frac{1}{2}\left\{\sum(x_i - \lambda)^2 + \sum(y_i - \mu)^2\right\}/\phi\right] \\
&\propto \phi^{-(m+n)/2-1} \exp\left[-\frac{1}{2}\{S_x + m(\bar{x} - \lambda)^2 + S_y\right. \\
&\quad \left.+ n(\bar{y} - \mu)^2\}/\phi\}] \\
&\propto \phi^{-v/2-1} \exp\left[-\frac{1}{2}S/\phi\right] (2\pi\phi/m)^{-\frac{1}{2}} \exp\left[-\frac{1}{2}m(\lambda - \bar{x})^2/\phi\right] \\
&\quad \times (2\pi\phi/n)^{-\frac{1}{2}} \exp\left[-\frac{1}{2}n(\mu - \bar{y})^2/\phi\right] \\
&\propto p(\phi|S)p(\lambda|\phi, \bar{x})p(\mu|\phi, \bar{y}),
\end{aligned}$$

where

$$\begin{aligned}
p(\phi|S) &\text{ is an } S\chi_v^{-2} \text{ density,} \\
p(\lambda|\phi, \bar{x}) &\text{ is an } N(\bar{x}, \phi/m) \text{ density,} \\
p(\mu|\phi, \bar{y}) &\text{ is an } N(\bar{y}, \phi/n) \text{ density.}
\end{aligned}$$

It follows that, for given  $\phi$ , the parameters  $\lambda$  and  $\mu$  have independent normal distributions, and hence that the joint density of  $\delta = \lambda - \mu$  and  $\delta$  is  $p(\delta, \phi | x, y) = p(\phi|S) p(\delta | \bar{x} - \bar{y}, \phi)$ ,

where  $p(\delta | \bar{x} - \bar{y}, \phi)$  is an  $N(\bar{x} - \bar{y}, \phi(m^{-1} + n^{-1}))$  density. The variance can now be integrated out just as in Section 2.12 when we considered a single sample from a normal distribution of unknown variance, giving a very similar conclusion, that

is, that if  $t = \frac{\delta - (\bar{x} - \bar{y})}{s(m^{-1} + n^{-1})^{1/2}}$ ,

where  $s^2 = S/\nu$ , then  $t \sim t_\nu$ . Note that the variance estimator  $s^2$  is found by adding the sums of squares  $S_x$  and  $S_y$  about the observed means and dividing by the sum of the corresponding numbers of degrees of freedom,  $\nu_x$  and  $\nu_y$ , and that this latter sum gives the number of degrees of freedom of the resulting Student's t variable. Another way of looking at it is that  $s^2$  is a weighted mean of the variance estimators  $s_x^2$  and  $s_y^2$  given by the two samples with weights proportional to the corresponding degrees of freedom.

## 5.2.2 Example

This section can be illustrated by using the data considered in the last section on the weight growth of rats, this time supposing (more realistically) that the variances are equal but unknown. We found that  $S_x=5032$ ,  $S_y=2552$ ,  $\nu_x = 11$  and  $\nu_y = 6$ , so that  $S = 7584$ ,  $\nu = 17$ ,  $s^2=7584/17=446$  and  $s(m^{-1} + n^{-1})^{1/2} = \{446(12^{-1} + 7^{-1})\}^{1/2} = 10$ .

Since  $\bar{x} = 120$  and  $\bar{y} = 101$ , the posterior distribution of  $\delta$  is given by  $(\delta - 19)/10 \sim t_{17}$ .

From tables of the t distribution it follows, for example, that a 90% HDR for  $\delta$  is

$19 \pm 1.740 \times 10$ , that is (2, 36). This is not very different from the result in Section 5.2, and indeed it will not usually make a great deal of difference to assume that variances are known unless the samples are very small.

It would also be possible to do other things with this posterior distribution, for example, to find the probability that  $\delta > 0$  or to test the point null hypothesis that  $\delta = 0$ , but this should be enough to give the idea.

### 5.2.3 Non-trivial prior information

A simple analysis is possible if we have prior information which, at least approximately, is such that the prior for  $\phi$  is  $S_0 \chi_{\nu_0}^{-2}$  and, conditional on  $\phi$ , the priors for  $\lambda$  and  $\mu$  are such that  $\lambda \sim N(\lambda_0, \phi/m_0)$

priors for  $\lambda$  and  $\mu$  are such that  $\mu \sim N(\mu_0, \phi/n_0)$

independently of one another. This means that

$$p(\lambda, \mu, \phi) \propto \phi^{-(v_0+2)/2-1} \exp \left[ -\frac{1}{2} \{S_0 + m_0(\lambda - \lambda_0)^2 + n_0(\mu - \mu_0)^2\}/\phi \right].$$

Of course, as in any case where conjugate priors provide a nice mathematical theory, it is a question that has to be faced up to in any particular case whether or not a prior of this form is a reasonable approximation to your prior beliefs, and if it is not then a more untidy analysis involving numerical integration will be necessary. The reference prior used earlier is of this form, though it results from the slightly strange choice of values  $v_0 = -2$ ,  $S_0 = m_0 = n_0 = 0$ . With such a prior, the

$$p(\lambda, \mu, \phi | x, y) \propto \phi^{-v_1/2-1} \exp \left[ -\frac{1}{2} S_1/\phi \right] (2\pi\phi/m_1)^{-\frac{1}{2}} \exp \left[ -\frac{1}{2} m_1(\lambda - \lambda_1)^2/\phi \right]$$

posterior is

$$\times (2\pi\phi/n_1)^{-\frac{1}{2}} \exp \left[ -\frac{1}{2} n_1(\mu - \mu_1)^2/\phi \right],$$

where

$$v_1 = v_0 + m + n, \quad m_1 = m_0 + m, \quad n_1 = n_0 + n,$$

$$\lambda_1 = (m_0\lambda_0 + m\bar{x})/m_1, \quad \mu_1 = (n_0\mu_0 + n\bar{y})/n_1,$$

$$S_1 = S_0 + S_x + S_y + (m_0^{-1} + m^{-1})^{-1} (\bar{x} - \lambda_0)^2 + (n_0^{-1} + n^{-1})^{-1} (\bar{y} - \mu_0)^2.$$

(The formula for  $S_1$  takes a little manipulation.) It is now possible to proceed as in the reference prior case, and so, for given  $\phi$ , the parameters  $\lambda$  and  $\mu$  have independent normal distributions, so that the joint density of  $\delta = \lambda - \mu$  and  $\phi$  can be written as  $p(\delta, \phi | x, y) = p(\phi | S_1) p(\delta | \bar{x} - \bar{y}, \phi)$ ,

where  $p(\delta | \bar{x} - \bar{y}, \phi)$  is an  $N(\bar{x} - \bar{y}, \phi(m^{-1} + n^{-1}))$  density. The variance can now be integrated out as before, giving a very similar result, namely, that if

$$t = \frac{\delta - (\bar{x} - \bar{y})}{s_1 (m_1^{-1} + n_1^{-1})^{\frac{1}{2}}},$$

where  $s_1^2 = S_1/v_1$ , then  $t \sim t_{v_1}$ .

The methodology is sufficiently similar to the case where a reference prior is used that it does not seem necessary to give a numerical example. Of course, the difficulty in using it, in practice, lies in finding appropriate values of the parameters of the prior distribution  $p(\lambda, \mu, \phi)$ .

## 5.3 Variances unknown and unequal (Behrens–Fisher problem)

### 5.3.1 Formulation of the problem

In this section, we are concerned with the most general case of the problem of two normal samples, where neither the means nor the variances are assumed equal. Consequently we have independent vectors  $x = (x_1, x_2, \dots, x_m)$  and  $y = (y_1, y_2, \dots, y_n)$  such that  $x_i \sim N(\lambda, \phi)$  and  $y_i \sim N(\mu, \psi)$  and  $\delta = \lambda - \mu$ . This is known as the Behrens–Fisher problem (or sometimes as the Behrens problem or the Fisher–Behrens problem).

It is convenient to use the notation of the previous section, except that sometimes we write  $v(x)$  and  $v(y)$  instead of  $v_x$  and  $v_y$  to avoid using sub-subscripts. In addition, it is useful to define  $s_x^2 = S_x/v_x$ ,  $s_y^2 = S_y/v_y$ .

For the moment, we shall assume independent reference priors uniform in  $\lambda$ ,  $\mu$ ,  $\phi$  and  $\psi$ . Then, just as in Section 2.12 on ‘Normal mean and variance both unknown’, it follows that the posterior distributions of  $\lambda$  and  $\mu$  are independent

$$T_x = \frac{\lambda - \bar{x}}{s_x/\sqrt{m}} \sim t_{v(x)} \quad \text{and} \quad T_y = \frac{\mu - \bar{y}}{s_y/\sqrt{n}} \sim t_{v(y)}$$

and are such that

It is now useful to define  $T$  and  $\theta$  by

$$T = \frac{\delta - (\bar{x} - \bar{y})}{\sqrt{(s_x^2/m + s_y^2/n)}} \quad \text{and} \quad \tan \theta = \frac{s_x/\sqrt{m}}{s_y/\sqrt{n}}$$

( $\theta$  can be taken in the first quadrant). It is then easy to check that

$$T = T_x \sin \theta - T_y \cos \theta.$$

Since  $\theta$  is known (from the data) and the distributions of  $T_x$  and  $T_y$  are known, it follows that the distribution of  $T$  can be evaluated. This distribution is tabulated and is called Behrens’ (or the Behrens–Fisher or Fisher–Behrens) distribution, and it will be denoted  $T \sim BF(v_x, v_y, \theta)$ .

It was first referred to in Behrens (1929).

### 5.3.2 Patil's approximation

Behrens' distribution turns out to have a rather nasty form, so that the density at any one point can only be found by a complicated integral, although a reasonable approximation was given by Patil (1965). To use this approximation, you need to

$$f_1 = \left( \frac{m-1}{m-3} \right) \sin^2 \theta + \left( \frac{n-1}{n-3} \right) \cos^2 \theta$$

$$f_2 = \frac{(m-1)^2}{(m-3)^2(m-5)} \sin^4 \theta + \frac{(n-1)^2}{(n-3)^2(n-5)} \cos^4 \theta$$

$$\text{find } b = 4 + (f_1^2/f_2) \quad \text{and} \quad a = \sqrt{f_1(b-2)/b}.$$

Then approximately

$$T/a \sim t_b.$$

Because  $b$  is not necessarily an integer, use of this approximation may necessitate interpolation in tables of the t distribution.

A rather limited table of percentage points of the Behrens distribution based on this approximation is to be found in the tables at the end of the book, but this will often be enough to give some idea as to what is going on. If more percentage points are required or the tables are not available, Patil's approximation or something like the program in Appendix C has to be used.

### 5.3.3 Example

Yet again we shall consider the data on the weight growth of rats as in Sections 5.1 and 5.2. Recall that  $m = 12$ ,  $n = 7$  (so  $\nu_x = 11$ ,  $\nu_y = 6$ ),  $\bar{x} = 120$ ,  $\bar{y} = 101$ ,  $S_x = 5032$ ,  $S_y = 2552$ , and hence  $s_x^2 = 457$ ,  $s_y^2 = 425$ . Therefore,

$$\tan \theta = \frac{s_x/\sqrt{m}}{s_y/\sqrt{n}} = \left( \frac{457/12}{425/7} \right)^{\frac{1}{2}} = 0.8; \quad \text{and} \quad \sqrt{(s_x^2/m + s_y^2/n)} = 9.9,$$

so that  $\theta = 39^\circ = 0.67$  radians using rounded values, and thus  $T \sim \text{BF}(11, 6, 39^\circ)$ . From the tables in the Appendix the 95% point of  $\text{BF}(12, 6, 30^\circ)$  is 1.91 and that of  $\text{BF}(12, 6, 45^\circ)$  is 1.88, so the 95% point of  $\text{BF}(11, 6, 39^\circ)$  must be about 1.89. [The program in Appendix C gives  $\text{hbehrens}(0.9, 11, 6, 39)$  as the interval  $(-1.882742, 1.882742)$ .] Consequently a 90% HDR for  $\delta$  is given by  $|T| \leq 1.89$  and so is  $(120 - 101) \pm 1.89 \times 9.9$ , that is,  $(0, 38)$ . This is slightly wider than was obtained in the previous section, as is reasonable, because we have made fewer assumptions and can only expect to get less precise conclusions.

The same result can be obtained directly from Patil's approximation. The required numbers turn out to be  $f_1 = 1.39$ ,  $f_2 = 0.44$ ,  $b = 8.39$ ,  $a = 1.03$ , so that  $T/1.03 \sim t_{8.39}$ . Interpolating between the 95% percentage points for  $t_8$  and  $t_9$

(which are 1.860 and 1.833, respectively), the required percentage point for  $t_{8,39}$  is 1.849, and hence a 90% HDR for  $\delta$  is  $(120 - 101) \pm 1.03 \times 1.849 \times 9.9$ , giving a very similar answer to that obtained from the tables. [The program in Appendix C gives this probability as  $pbehrens(19/9.9, 11, 6, 39) = 0.9512438$ .]

Of course, it would need more extensive tables to find, for example, the posterior probability that  $\delta > 0$ , but there is no difficulty in principle in doing so. On the other hand, it would be quite complicated to find the Bayes factor for a test of a point null hypothesis such as  $\delta = 0$ , and since such tests are only to be used with caution in special cases, it would not be likely to be worthwhile.

### 5.3.4 Substantial prior information

If we do happen to have substantial prior information about the parameters which can reasonably well be approximated by independent normal/chi-squared distributions for  $(\lambda, \mu)$  and  $(\phi, \psi)$ , then the method of this section can usually be extended to include it. All that will happen is that  $T_x$  and  $T_y$  will be replaced by slightly different quantities with independent t distributions, derived as in Section 2.12 on ‘Normal mean and variance both unknown’. It should be fairly clear how to carry out the details, so no more will be said about this case.

## 5.4 The Behrens–Fisher controversy

### 5.4.1 The Behrens–Fisher problem from a classical standpoint

As pointed out in Section 2.6 on ‘Highest Density Regions’, in the case of a single normal observation of known variance there is a close relationship between classical results and Bayesian results using a reference prior, which can be summarized in terms of the ‘tilde’ notation by saying that, in classical statistics, results depend on saying that  $(\theta - \tilde{x})/\sqrt{\phi} \sim N(0, 1)$

while Bayesian results depend on saying that

$$(\tilde{\theta} - x)/\sqrt{\phi} \sim N(0, 1).$$

As a result of this, if  $\phi = 1$  then the observation  $x = 5$ , say, leads to the same interval,  $5 \pm 1.96$ , which is regarded as a 95% confidence interval for  $\theta$  by classical statisticians and as a 95% HDR for  $\theta$  by Bayesians (at least if they are using a reference prior). It is not hard to see that very similar relationships exist

if we have a sample of size  $n$  and replace  $x$  by  $\bar{x}$ , and also when the variance is unknown (provided that the normal distribution is replaced by the t distribution).

There is also no great difficulty in dealing with the case of a two sample problem in which the variances are known. If they are unknown but equal (i.e.

$$\phi = \psi), it was shown that if \quad t = \frac{\delta - (\bar{x} - \bar{y})}{s(m^{-1} + n^{-1})^{\frac{1}{2}}}$$

then the posterior distribution of  $t$  is Student's on  $\nu = \nu_x + \nu_y$  degrees of freedom. A classical statistician would say that this 'pivotal quantity' has the same distribution whatever  $(\lambda, \mu, \phi)$  are, and so would be able to give confidence intervals for  $\delta$  which were exactly the same as HDRs derived by a Bayesian statistician (always assuming that the latter used a reference prior).

This seems to suggest that there is always likely to be a way of interpreting classical results in Bayesian terms and vice versa, provided that a suitable prior distribution is used. One of the interesting aspects of the Behrens–Fisher problem is that no such correspondence exists in this case. To see why, recall that the Bayesian analysis led us to conclude that  $T \sim \text{BF}(\nu_x, \nu_y, \theta)$

where

$$T = \frac{\delta - (\bar{x} - \bar{y})}{\sqrt{(s_x^2/m + s_y^2/n)}} \quad \text{and} \quad \tan \theta = \frac{s_x/\sqrt{n}}{s_y/\sqrt{m}}.$$

Moreover, changing the prior inside the conjugate family would only alter the parameters slightly, but would still give results of the same general character. So if there is to be a classical analogue to the Bayesian result, then if  $T$  is regarded as a function of the data  $x = (x_1, x_2, \dots, x_m)$  and  $y = (y_1, y_2, \dots, y_n)$  for fixed values of the parameters  $(\lambda, \mu, \phi, \psi)$ , it must have Behrens' distribution over repeated samples  $x$  and  $y$ . There is an obvious difficulty in this, in that the parameter  $\theta$  depends on the samples, whereas there is no such parameter in the normal or t distributions. However, it is still possible to investigate whether the sampling distribution of  $T$  depends on the parameters  $(\lambda, \mu, \phi, \psi)$ .

It turns out that its distribution over-repeated sampling does not just depend on the sample sizes  $m$  and  $n$  – it also depends on the ratio  $\phi/\psi$  (which is not, in general, known to the statistician). It is easiest to see this when  $m = n$  and so  $\nu_x = \nu_y = \nu/2$  (say). We first suppose that (unknown to the statistician) it is in fact the case that  $\phi/\psi = 1$ . Then the sampling distribution found in Section 5.2, for the case where the statistician did happen to know that  $\phi = \psi$  must still hold (his or her ignorance can scarcely affect what happens in repeated sampling).

Because if  $m = n$  then  $\sqrt{(s_x^2/m + s_y^2/n)} = \sqrt{\{2(S_x + S_y)/vm\}} = s\sqrt{(2/m)} = s(m^{-1} + n^{-1})^{\frac{1}{2}}$

in the notation of Section 5.2, it follows that

$$T = \frac{\delta - (\bar{x} - \bar{y})}{s(m^{-1} + n^{-1})^{\frac{1}{2}}} \sim t_v.$$

On the other hand if  $\psi = 0$ , then necessarily  $S_y = 0$  and so  $s_y^2 = 0$ , and hence  $\sqrt{(s_x^2/m + s_y^2/n)} = \sqrt{(s_x^2/m)}$ . Since it must also be the case that  $\bar{y} = \mu$  and so  $\delta - (\bar{x} - \bar{y}) = \lambda - x$ , the distribution of  $T$  is given by  $T = \frac{\lambda - \bar{x}}{s_x/\sqrt{m}} \sim t_{v(x)}$  that is,  $T$  has a t distribution on  $v_x = v/2$  degrees of freedom. For intermediate values of  $\phi/\psi$  the distribution over repeated samples is intermediate between these forms (but is not, in general, a t distribution).

### 5.4.2 Example

Bartlett (1936) quotes an experiment in which the yields  $x_i$  (in pounds per acre) on  $m$  plots for early hay were compared with the yields  $y_i$  for ordinary hay on another  $n$  plots. It turned out that  $m=n=7$  (so  $v_x = v_y = 6$ ),  $\bar{x} = 448.6$ ,  $\bar{y} = 408.8$ ,  $s_x^2 = 308.6$ ,  $s_y^2 = 1251.3$ . It follows that  $\bar{x} - \bar{y} = 39.8$ ,  $\sqrt{(s_x^2/m + s_y^2/n)} = 14.9$  and  $\tan \theta = \sqrt{(308.6/1251.3)}$ , so that  $\theta = 26^\circ = 0.46$  radians. The Bayesian analysis now proceeds by saying that  $(\delta - 39.8)/14.9 \sim BF(6, 6, 26^\circ)$ . By interpolation in tables of Behrens' distribution a 50% HDR for  $\delta$  is  $39.8 \pm 0.75 \times 14.9$ , that is, (28.6, 51.0). [Using the program in Appendix C we find that  $hbehrens(0.5, 6, 6, 26)$  is the interval (-0.7452094, 0.7452094).]

A classical statistician who was willing to assume that  $\phi = \psi$  would use tables of  $t_{12}$  to conclude that a 50% confidence interval was  $39.8 \pm 0.695 \times 14.9$ , that is, (29.4, 50.2). This interval is different, although not very much so, from the Bayesian's HDR. Without some assumption such as  $\phi = \psi$  he or she would not be able to give any exact answer.

### 5.4.3 The controversy

The Bayesian solution was championed by Fisher (1935, 1937, 1939). Fisher had his own theory of *fiducial* inference which does not have many adherents nowadays, and did not in fact support the Bayesian arguments put forward here. In an introduction to Fisher (1939) reprinted in his *Collected Papers*, Fisher said that Pearson and Neyman have laid it down axiomatically that the level of significance of a test must be equal to the frequency of a wrong decision 'in

repeated samples from the same population'. The idea was foreign to the development of tests of significance given by the author in [*Statistical Methods for Research Workers*], for the experimenter's experience does not consist in repeated samples from the same population, although in simple cases the numerical values are often the same; and it was, I believe, this coincidence which misled Pearson and Neyman, who were not very familiar with the ideas of 'Student' and the author.

Although Fisher was not a Bayesian, the above quotation does put one of the objections which any Bayesian must have to classical tests of significance.

In practice, classical statisticians can at least give intervals which, while they may not have an exact significance level, have a significance level between two reasonably close bounds. A recent review of the problem is given by Robinson (1976, 1982).

## 5.5 Inferences concerning a variance ratio

### 5.5.1 Statement of the problem

In this section, we are concerned with the data of the same form as we met in the Behrens–Fisher problem. Thus say, we have independent vectors  $x = (x_1, x_2, \dots, x_m)$  and  $y = (y_1, y_2, \dots, y_n)$  such that  $x_i \sim N(\lambda, \phi)$  and  $y_i \sim N(\mu, \psi)$

where all of  $(\lambda, \mu, \phi, \psi)$  are unknown. The difference is that in this case the quantity of interest is the ratio  $\kappa = \phi/\psi$

of the two unknown variances, so that the intention is to discover how much more (or less) variable the one population is than the other. We shall use the same notation as before, and in addition we will find it useful to define

$$\kappa = \frac{s_x^2}{s_y^2} = \frac{S_x/v_x}{S_y/v_y} = \frac{v_y}{v_x} \frac{S_x}{S_y},$$

$$\eta = \kappa/k.$$

Again, we shall begin by assuming a reference prior

$$p(\lambda, \mu, \phi, \psi) \propto 1/\phi\psi.$$

As was shown in Section 2.12 on 'Normal mean and variance both unknown', the posterior distributions of  $\phi$  and  $\psi$  are independent and such that  $\phi \sim S_x \chi_{v(x)}^{-2}$  and  $\psi \sim S_y \chi_{v(y)}^{-2}$ , so that  $p(\phi, \psi|x, y) \propto \phi^{-v(x)/2-1} \psi^{-v(y)/2-1} \exp(-\frac{1}{2}S_x/\phi - \frac{1}{2}S_y/\psi)$ .

It turns out that  $\eta$  has (Snedecor's) F distribution on  $v_y$  and  $v_x$  degrees of

freedom (or equivalently that its reciprocal has an F distribution on  $\nu_x$  and  $\nu_y$  degrees of freedom). The proof of this fact, which is not of great importance and can be omitted if you are prepared to take it for granted, is in Section 5.6.

The result is of the same type, although naturally the parameters are slightly different, if the priors for  $\phi$  and  $\psi$  are from the conjugate family. Even if, by a fluke, we happened to know the means but not the variances, the only change would be an increase of 1 in each of the degrees of freedom.

### 5.5.2 Derivation of the F distribution

In order to find the distribution of  $\kappa$ , we need first to change variables to  $(\kappa, \phi)$ , noting that  $\frac{\partial(\kappa, \phi)}{\partial(\phi, \psi)} = \begin{vmatrix} 1/\psi & -\phi/\psi^2 \\ 1 & 0 \end{vmatrix} = \phi/\psi^2 = \kappa^2/\phi$ .

It follows that

$$\begin{aligned} p(\kappa, \phi | x, y) &= p(\phi, \psi | x, y) \left| \frac{\partial(\kappa, \phi)}{\partial(\phi, \psi)} \right|^{-1} = p(\phi, \psi | x, y) \kappa^{-2} \phi \\ &= \kappa^{\nu(y)/2-1} \phi^{-(\nu(x)+\nu(y))/2-1} \exp \left\{ -\frac{1}{2}(S_x + \kappa S_y)/\phi \right\}. \end{aligned}$$

It is now easy enough to integrate  $\phi$  out by substituting  $x = \frac{1}{2}A/\phi$  where  $A = S_x + \kappa S_y$  and thus reducing the integral to a standard gamma function integral (cf. Section 2.12 on ‘Normal mean and variance both unknown’).

$$p(\kappa | x, y) \propto \kappa^{\nu(y)/2-1} \{S_x + \kappa S_y\}^{-(\nu(x)+\nu(y))/2}$$

Hence,

$$\propto \kappa^{\nu(y)/2-1} \{(S_y/S_x) + \kappa\}^{-(\nu(x)+\nu(y))/2}.$$

Defining  $k$  and  $\eta$  as above, and noting that  $d\eta/d\kappa$  is constant, this density can be transformed to give  $p(\eta | x, y) \propto \eta^{\nu(y)/2-1} \{v_x + v_y \eta\}^{-(\nu(x)+\nu(y))/2}$ .

From Appendix A, it can be seen that this is an F distribution on  $\nu_y$  and  $\nu_x$  degrees of freedom, so that  $\eta \sim F_{\nu(y), \nu(x)}$ .

Note that by symmetry

$$\eta^{-1} \sim F_{\nu(x), \nu(y)}.$$

For most purposes, it suffices to think of an F distribution as being, by definition, the distribution of the ratio of two chi-squared (or inverse chi-squared) variables divided by their respective degrees of freedom.

### 5.5.3 Example

Jeffreys (1961, Section 5.4) quotes the following data (due to Lord Rayleigh) on the masses  $x_i$  in grammes of  $m = 12$  samples of nitrogen obtained from air (A) and the masses  $y_i$  of  $n = 8$  samples obtained by chemical method (C) within a

given container at standard temperature and pressure.

A	2.31035	2.31026	2.31024	2.31012	2.31027	2.31017
	2.30986	2.31010	2.31001	2.31024	2.31010	2.31028
C	2.30143	2.29890	2.29816	2.30182	2.29869	2.29940
	2.29849	2.29889				

It turns out that  $\bar{x} = 2.31017$ ,  $\bar{y} = 2.29947$ ,  $s_x^2 = 18.75 \times 10^{-9}$ ,  $s_y^2 = 1902 \times 10^{-9}$ , so that  $k = 19/1902 = 0.010$ . Hence, the posterior of  $\kappa$  is such that  $\eta = \kappa/k = 100\kappa \sim F_{7,11}$  or equivalently

$$\eta^{-1} = k/\kappa = 0.01\kappa \sim F_{11,7}.$$

This makes it possible to give an interval in which we can be reasonably sure that the ratio  $\kappa$  of the variances lies. For reasons similar to those for which we chose to use intervals corresponding to HDRs for  $\log \chi^2$  in Section 2.8 on ‘HDRs for the normal variance’, it seems sensible to use intervals corresponding to HDRs for  $\log F$ . From the tables in the Appendix, such an interval of probability 90% for  $F_{11,7}$  is (0.32, 3.46), so that  $\kappa$  lies in the interval from 0.01/3.46 to 0.01/0.32, that is (0.003, 0.031), with a posterior probability of 90%. Because the distribution is markedly asymmetric, it may also be worth finding the mode of  $\kappa$ , which (from the mode of  $F_{\nu(y),\nu(x)}$  as given in Appendix A) is  $k \frac{\nu_x}{\nu_x + 1} \frac{\nu_y - 2}{\nu_y} = 0.010 \frac{11.5}{12.7} = 0.0065$ .

## 5.6 Comparison of two proportions; the $2 \times 2$ table

### 5.6.1 Methods based on the log-odds ratio

In this section, we are concerned with another two sample problem, but this time one arising from the binomial rather than the normal distribution. Suppose  $x \sim B(m, \pi)$  and  $y \sim B(n, \rho)$

and that we are interested in the relationship between  $\pi$  and  $\rho$ . Another way of describing this situation is in terms of a  $2 \times 2$  table (sometimes called a  $2 \times 2$

	Population I	Population II
Successes	$a = x$	$c = y$
Failures	$b = m - x$	$d = n - y$
Total	$m$	$n$

contingency table)

We shall suppose that the priors for  $\pi$  and  $\rho$  are such that  $\pi \sim Be(\alpha_0, \beta_0)$  and  $\rho \sim Be(\gamma_0, \delta_0)$ , independently of one another. It follows that the posteriors are

also beta distributions, and more precisely if  
 $\alpha = \alpha_0 + x$ ,  $\beta = \beta_0 + m - x$ ,  $\gamma = \gamma_0 + y$ ,  $\delta = \delta_0 + n - y$   
then

$$\pi \sim \text{Be}(\alpha, \beta) \quad \text{and} \quad \rho \sim \text{Be}(\gamma, \delta).$$

We recall from Section 3.1 on the binomial distribution that if

$$\Lambda = \log \lambda = \log\{\pi/(1-\pi)\}, \quad \Lambda' = \log \lambda' = \log\{\rho/(1-\rho)\}$$

then  $\frac{1}{2}\Lambda + \frac{1}{2}\log(\beta/\alpha) \sim z_{2\alpha, 2\beta}$ , so that

$$\mathbb{E}\Lambda \cong \log\left\{\left(\alpha - \frac{1}{2}\right) / \left(\beta - \frac{1}{2}\right)\right\},$$

$$\mathcal{V}\Lambda \cong \alpha^{-1} + \beta^{-1}$$

and similarly for  $\Lambda'$ . Now the  $z$  distribution is approximately normal (this is the reason that Fisher preferred to use the  $z$  distribution rather than the  $F$  distribution, which is not so near to normality), and so  $\Lambda$  and  $\Lambda'$  are approximately normal with these means and variances. Hence the *log-odds ratio*  $\Lambda - \Lambda' = \log(\lambda/\lambda')$

is also approximately normal, that is,

$$\Lambda - \Lambda' \sim N\left(\log\left(\left(\alpha - \frac{1}{2}\right)\left(\delta - \frac{1}{2}\right) / \left(\beta - \frac{1}{2}\right)\left(\gamma - \frac{1}{2}\right)\right), \alpha^{-1} + \beta^{-1} + \gamma^{-1} + \delta^{-1}\right)$$

or more approximately

$$\Lambda - \Lambda' \sim N(\log\{\alpha\delta/\beta\gamma\}, \alpha^{-1} + \beta^{-1} + \gamma^{-1} + \delta^{-1}).$$

If the Haldane reference priors are used, so that  $\alpha_0 = \beta_0 = \gamma_0 = \delta_0 = 0$ , then  $\alpha = a$ ,  $\beta = b$ ,  $\gamma = c$  and  $\delta = d$ , and so  $\Lambda - \Lambda' \sim N(\log\{ad/bc\}, a^{-1} + b^{-1} + c^{-1} + d^{-1})$ .

The quantity  $ad/bc$  is sometimes called the *cross-ratio*, and there are good grounds for saying that any measure of association in the  $2 \times 2$  table should be a function of the cross-ratio (cf. Edwards, 1963).

The log-odds ratio is a sensible measure of the degree to which the two populations are identical, and in particular  $\pi > \rho$  if and only if  $\Lambda - \Lambda' > 0$ . On the other hand, knowledge of the posterior distribution of the log-odds ratio does not in itself imply knowledge of the posterior distribution of the difference  $\pi - \rho$  or the ratio  $\pi/\rho$ . The approximation is likely to be reasonable provided that all of the entries in the  $2 \times 2$  table are at least 5.

## 5.6.2 Example

The table mentioned later [quoted from Di Raimondo (1951)] relates to the effect on mice of bacterial inoculum (*Staphylococcus aureus*). Two different types of injection were tried, a standard one and one with 0.15 U of penicillin per millilitre.

	Standard	Penicillin
Alive	8	48
Dead	12	62
Total	20	110

The cross-ratio is  $ad/bc = (8 \times 62)/(12 \times 48) = 0.861$  so its logarithm is  $-0.150$  and  $a^{-1}+b^{-1}+c^{-1}+d^{-1}=0.245$ , and so the posterior distribution of the log odds-ratio is  $\Lambda - \Lambda' \sim N(-0.150, 0.245)$ . Allowing for the  $\frac{1}{2}$ s in the more exact form for the mean does not make much difference; in fact  $-0.150$  becomes  $-0.169$ . The posterior probability that  $\pi > \rho$ , that is, that the log odds ratio is positive, is  $\Phi(-0.169/\sqrt{0.245}) = \Phi(-0.341) = 0.3665$ .

The data thus shows no great difference between the injections with and without the penicillin.

### 5.6.3 The inverse root-sine transformation

In Section 1.5 on ‘Means and variances’, we saw that if  $x \sim B(m, \pi)$ , then the transformation  $z = \sin^{-1} \sqrt{(x/m)}$  resulted in  $Ez = \sin^{-1} \sqrt{\pi} = \psi$ , say, and  $Vz = 1/4m$ , and in fact it is approximately true that  $z \sim N(\psi, 1/4m)$ . This transformation was also mentioned in Section 3.2 on ‘Reference prior for the binomial likelihood’, and pointed out there that one of the possible reference priors for  $\pi$  was  $Be(\frac{1}{2}, \frac{1}{2})$ , and that this prior was equivalent to a uniform prior in  $\sin^{-1} \sqrt{\pi} = \psi$ . Now if we use such a prior, then clearly the posterior for  $\psi$  is approximately  $N(z, 1/4m)$ , that is,  $\psi = \sin^{-1} \sqrt{\pi} \sim N(\sin^{-1} \sqrt{(x/m)}, 1/4m)$ .

This is of no great use if there is only a single binomial variable, but when there are two it can be used to conclude that approximately  $\sin^{-1} \sqrt{\pi} - \sin^{-1} \sqrt{\rho} \sim N(\sin^{-1} \sqrt{(x/m)} - \sin^{-1} \sqrt{(y/n)}, 1/4m + 1/4n)$

and so to give another approximation to the probability that  $\pi > \rho$ . Thus with the same data as the above,  $\sin^{-1} \sqrt{(x/m)} = \sin^{-1} \sqrt{(8/20)} = 0.685$  radians,  $\sin^{-1} \sqrt{(48/110)} = 0.721$  radians, and  $1/4m + 1/4n = 0.0148$ , so that the posterior probability that  $\pi > \rho$  is about  $\Phi(-0.036/\sqrt{0.0148}) = \Phi(-0.296) = 0.3936$ . The two methods do not give precisely the same answer, but it should be borne in mind that the numbers are not very large, so the approximations involved are not very good, and also that we have assumed slightly different reference priors in deriving the two answers.

If there is non-trivial prior information, it can be incorporated in this method as well as in the previous method. The approximations involved are reasonably accurate provided that  $x(1-x/m)$  and  $y(1-y/n)$  are both at least 5.

## 5.6.4 Other methods

If all the entries in the  $2 \times 2$  table are at least 10, then the posterior beta distributions are reasonably well approximated by normal distributions of the same means and variances. This is quite useful in that it gives rise to an approximation to the distribution of  $\pi - \rho$  which is much more likely to be of interest than some function of  $\pi$  minus the same function of  $\rho$ . It will therefore allow us to give an approximate HDR for  $\pi - \rho$  or to approximate the probability that  $\pi - \rho$  lies in a particular interval.

In quite a different case, where the values of  $\pi$  and  $\rho$  are small, which will be reflected in small values of  $x/m$  and  $y/n$ , then the binomial distributions can be reasonably well approximated by Poisson distributions, which means that the posteriors of  $\pi$  and  $\rho$  are multiples of chi-squared distributions (cf. Section 3.4 on ‘The Poisson distribution’). It follows from this that the posterior of  $\pi/\rho$  is a multiple of an F distribution (cf. Section 5.5). Again, this is quite useful because  $\pi/\rho$  is a quantity of interest in itself. The Poisson approximation to the binomial is likely to be reasonable if  $n > 10$  and either  $x/n < 0.05$  or  $x/n > 0.95$  (in the latter case,  $\pi$  has to be replaced by  $1 - \pi$ ).

The exact probability that  $\pi < \rho$  can be worked out in terms of hypergeometric probabilities (cf. Altham, 1969), although the resulting expression is not usually useful for hand computation. It is even possible to give an expression for the posterior probability that  $\pi/\rho \leq c$  (cf. Weisberg, 1972), but this is even more unwieldy.

## 5.7 Exercises on Chapter 5

1. Two analysts measure the percentage of ammonia in a chemical process over 9 days and find the following discrepancies between their results:

Day	1	2	3	4	5	6	7	8	9
Analyst A	12.04	12.37	12.35	12.43	12.34	12.36	12.48	12.33	12.33
Analyst B	12.18	12.37	12.38	12.36	12.47	12.48	12.57	12.28	12.42

Investigate the mean discrepancy  $\theta$  between their results and in particular give an interval in which you are 90% sure that it lies.

2. With the same data as in the previous question, test the hypothesis that there is no discrepancy between the two analysts.

3. Suppose that you have grounds for believing that observations  $x_i, y_i$  for  $i = 1, 2 \dots, n$  are such that  $x_i \sim N(\theta, \phi_i)$  and also  $y_i \sim N(\theta, \phi_i)$ , but that you are not prepared to assume that the  $\phi_i$  are equal. What statistic would you expect to base inferences about  $\theta$  on?

4. How much difference would it make to the analysis of the data in Section 5.1 on rat diet if we took  $\omega = \frac{1}{2}(\phi + \psi)$  instead of  $\omega = \phi + \psi$ .

5. Two analysts in the same laboratory made repeated determinations of the percentage of fibre in soya cotton cake, the results being as shown:

Analyst A	12.38	12.53	12.25	12.37	12.48	12.58	12.43	12.43	12.30
Analyst B	12.25	12.45	12.31	12.31	12.30	12.20	12.25	12.25	12.26
	12.42	12.17	12.09						

Investigate the mean discrepancy  $\theta$  between their mean determinations and in particular give an interval in which you are 90% sure that it lies (a) assuming that it is known from past experience that the standard deviation of both sets of observations is 0.1, and

(b) assuming simply that it is known that the standard deviations of the two sets of observations are equal.

6. A random sample  $x = (x_1, x_2, \dots, x_m)$  is available from an  $N(\lambda, \phi)$  distribution and a second independent random sample  $y = (y_1, y_2, \dots, y_n)$  is available from an  $N(\mu, 2\phi)$  distribution. Obtain, under the usual assumptions, the posterior distributions of  $\lambda - \mu$  and of  $\phi$ .

7. Verify the formula for  $S_1$  given towards the end of Section 5.2.

8. The following data consists of the lengths in mm of cuckoo's eggs found in nests belonging to the dunnock and to the reed warbler:

Dunnock	22.0	23.9	20.9	23.8	25.0	24.0	21.7	23.8	22.8	23.1
Reed warbler	23.2	22.0	22.2	21.2	21.6	21.9	22.0	22.9	22.8	

Investigate the difference  $\theta$  between these lengths without making any particular assumptions about the variances of the two populations, and in particular give an interval in which you are 90% sure that it lies.

- 9.** Show that if  $m=n$  then the expression  $f_1/f_2$  in Patil's approximation reduces to  $\frac{4(m-5)}{3+\cos 4\theta}$ .

- 10.** Suppose that  $T_x$ ,  $T_y$  and  $\theta$  are defined as in Section 5.3 and that  $T = T_x \sin \theta - T_y \cos \theta$ ,  $U = T_x \cos \theta + T_y \sin \theta$

Show that the transformation from  $(T_x, T_y)$  to  $(T, U)$  has unit Jacobian and hence show that the density of  $T$  satisfies

$$p(T|x, y) \propto \int_0^\infty [1 + (T \sin \theta + U \cos \theta)^2/v_x]^{-(v(x)+1)/2} \times [1 + (-T \cos \theta + U \sin \theta)^2/v_y]^{-(v(y)+1)/2} dU.$$

- 11.** Show that if  $x \sim F_{\nu_1, \nu_2}$  then

$$\frac{\nu_1 x}{\nu_2 + \nu_1 x} \sim \text{Be}\left(\frac{1}{2}\nu_1, \frac{1}{2}\nu_2\right).$$

- 12.** Two different microscopic methods,  $A$  and  $B$ , are available for the measurement of very small dimensions in microns. As a result of several such measurements on the same object, estimates of variance are available

Method	$A$	$B$
No. of observations	$m = 15$	$n = 25$
Estimated variance	$s_1^2 = 7.533$	$s_2^2 = 1.112$

Give an interval in which you are 95% sure that the ratio of the variances lies.

- 13.** Measurement errors when using two different instruments are more or less symmetrically distributed and are believed to be reasonably well approximated by a normal distribution. Ten measurements with each show a sample standard deviation three times as large with one instrument as with the other. Give an interval in which you are 99% sure that the ratio of the true standard deviations lies.

- 14.** Repeat the analysis of Di Raimondo's data in Section 5.6 on the effects of penicillin of mice, this time assuming that you have prior knowledge worth about six observations in each case suggesting that the mean chance of survival is about a half with the standard injection but about two-thirds with the penicillin injection.

- 15.** The undermentioned table [quoted from Jeffreys (1961, Section 5.1)] gives the relationship between grammatical gender in Welsh and

Psycho.\Gram.	<i>M</i>	<i>F</i>
<i>M</i>	45	30
<i>F</i>	28	29
Total	73	59

psychoanalytical symbolism according to Freud:

Find the posterior probability that the log odds-ratio is positive and compare it with the comparable probability found by using the inverse root-sine transformation.

**16.** Show that if  $\pi \cong \rho$  then the log odds-ratio is such that

$$\Lambda - \Lambda' \cong (\pi - \rho)/\{\pi(1 - \pi)\}.$$

**17.** A report issued in 1966 about the effect of radiation on patients with inoperable lung cancer compared the effect of radiation treatment with placebos. The numbers surviving after a year were:

	Radiation	Placebos
No. of cases	308	246
No. surviving	56	34

What are the approximate posterior odds that the one-year survival rate of irradiated patients is at least 0.01 greater than that of those who were not irradiated?

**18.** Suppose that  $x \sim P(8.5)$ , that is  $x$  is Poisson of mean 8.5, and  $y \sim P(11.0)$ . What is the approximate distribution of  $x - y$ ?

# 6

## Correlation, regression and the analysis of variance

### 6.1 Theory of the correlation coefficient

#### 6.1.1 Definitions

The standard measure of association between two random variables, which was first mentioned in Section 1.5 on ‘Means and Variances’, is the *correlation*

$$\text{coefficient } \rho(x, y) = \frac{\mathcal{C}(x, y)}{\sqrt{(\mathcal{V}x)(\mathcal{V}y)}}.$$

It is used to measure the strength of linear association between two variables, most commonly in the case where it might be expected that both have, at least approximately, a normal distribution. It is most important in cases where it is not thought that either variable is dependent on the other. One example of its use would be an investigation of the relationship between the height and the weight of individuals in a population, and another would be in finding how closely related barometric gradients and wind velocities were. You should, however, be warned that it is very easy to conclude that measurements are closely related because they have a high correlation, when, in fact, the relationship is due to their having a common time trend or a common cause and there is no close relationship between the two (see the relationship between the growth of money supply and Scottish dysentery as pointed out in a letter to *The Times* dated 6 April 1977). You should also be aware that two closely related variables can have a low correlation if the relationship between them is highly non-linear.

We suppose, then, that we have a set of  $n$  ordered pairs of observations, the pairs being independent of one another but members of the same pair being, in general, not independent. We shall denote these observations  $(x_i, y_i)$  and, as usual, we shall write  $x = (x_1, x_2, \dots, x_n)$  and  $y = (y_1, y_2, \dots, y_n)$ . Further, suppose that these pairs have a *bivariate normal distribution* with  $\mathbf{E}x_i = \lambda$ ,  $\mathbf{E}y_i = \mu$ ,  $\mathcal{V}x_i = \phi$ ,  $\mathcal{V}y_i = \psi$ ,  $\rho(x_i, y_i) = \rho$ ,

and we shall use the notation

$$\bar{x} = \sum x_i/n, \quad \bar{y} = \sum y_i/n, \quad S_{xx} = \sum (x_i - \bar{x})^2, \quad S_{yy} = \sum (y_i - \bar{y})^2$$

( $S_{xx}$  and  $S_{yy}$  have previously been denoted  $S_x$  and  $S_y$ ), and  $S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y})$ .

It is also useful to define the *sample correlation coefficient*  $r$  by

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{[\sum (x_i - \bar{x})^2]\sum (y_i - \bar{y})^2]} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}},$$

so that  $S_{xy} = r\sqrt{(S_{xx}S_{yy})}$ .

We shall show that, with standard reference priors for  $\lambda$ ,  $\mu$ ,  $\phi$  and  $\psi$ , a reasonable approximation to the posterior density of  $\rho$  is given by

$$p(\rho|x, y) \propto p(\rho) \frac{(1 - \rho^2)^{(n-1)/2}}{(1 - \rho r)^{n-(3/2)}}$$

where  $p(\rho)$  is its prior density. Making the substitution

$$\rho = \tanh \zeta, \quad r = \tanh z$$

we will go on to show that after another approximation

$$\zeta \sim N(z, 1/n)$$

These results will be derived after quite a complicated series of substitutions [due to Fisher (1915, 1921)]. Readers who are prepared to take these results for granted can omit the rest of this section.

### 6.1.2 Approximate posterior distribution of the correlation coefficient

As before, we shall have use for the formulae

$$\sum (x_i - \lambda)^2 = S_{xx} + n(\bar{x} - \lambda)^2, \quad \sum (y_i - \mu)^2 = S_{yy} + n(\bar{y} - \mu)^2,$$

and also for a similar one not used before

$$\sum (x_i - \lambda)(y_i - \mu) = S_{xy} + n(\bar{x} - \lambda)(\bar{y} - \mu).$$

Now the (joint) density function of a single pair  $(x, y)$  of observations from a bivariate normal distribution is

$$p(x, y|\lambda, \mu, \phi, \psi, \rho) = \frac{1}{2\pi\sqrt{\{\phi\psi(1 - \rho^2)\}}} \exp\left(-\frac{1}{2(1 - \rho^2)}Q_0\right)$$

where

$$Q_0 = \phi^{-1}(x - \lambda)^2 - 2\rho(\phi\psi)^{-\frac{1}{2}}(x - \lambda)(y - \mu) + \psi^{-1}(y - \mu)^2$$

and hence the joint density of the vector  $(x, y)$  is

$$p(x, y|\lambda, \mu, \phi, \psi, \rho) \propto \{\phi\psi(1 - \rho^2)\}^{-(n-1)/2} \exp\left\{-\frac{1}{2}\sum Q_i/(1 - \rho^2)\right\}$$

$$\times \left[2\pi n^{-1}\sqrt{\{\phi\psi(1 - \rho^2)\}}\right]^{-1} \exp\left\{-\frac{1}{2}n\sum Q'_i/(1 - \rho^2)\right\},$$

where

$$Q = \phi^{-1}S_{xx} - 2\rho(\phi\psi)^{-\frac{1}{2}}S_{xy} + \psi^{-1}S_{yy},$$

$$Q' = \phi^{-1}(\bar{x} - \lambda)^2 - 2\rho(\phi\psi)^{-\frac{1}{2}}(\bar{x} - \lambda)(\bar{y} - \mu) + \psi^{-1}(\bar{y} - \mu)^2.$$

It follows that the vector  $(\bar{x}, \bar{y}, S_{xx}, S_{yy}, r)$  is sufficient for  $(\lambda, \mu, \phi, \psi, \rho)$ . For the moment, we shall use independent priors of a simple form. For  $\lambda, \mu, \phi$  and  $\psi$ , we shall take the standard reference priors, and for the moment we shall use a perfectly general prior for  $\rho$ , so that  $p(\lambda, \mu, \phi, \psi, \rho) \propto p(\rho)/\phi\psi$  and hence

$$\begin{aligned} p(\lambda, \mu, \phi, \psi, \rho | x, y) &\propto p(\rho)(1 - \rho^2)^{-(n-1)/2}\{\phi\psi\}^{-(n+1)/2} \\ &\quad \times \exp\{-\frac{1}{2}Q/(1 - \rho^2)\} \\ &\quad \times \left[2\pi n^{-1}\sqrt{\{\phi\psi(1 - \rho^2)\}}\right]^{-1} \exp\{-\frac{1}{2}nQ'/(1 - \rho^2)\}. \end{aligned}$$

The last factor is evidently the (joint) density of  $\lambda$  and  $\mu$  considered as bivariate normal with means  $\bar{x}$  and  $\bar{y}$ , variances  $\phi/n$  and  $\psi/n$  and correlation  $\rho$ . Consequently it integrates to unity, and so as the first factor does not depend on  $\lambda$  or  $\mu$

$$p(\phi, \psi, \rho | x, y) \propto p(\rho)(1 - \rho^2)^{-(n-1)/2}\{\phi\psi\}^{-(n+1)/2} \exp\{-\frac{1}{2}Q/(1 - \rho^2)\}.$$

To integrate  $\phi$  and  $\psi$  out, it is convenient to define  $\xi = \sqrt{\{(\phi\psi)/(S_{xx}S_{yy})\}}$ ,  $\omega = \sqrt{\{(\phi S_{yy})/(\psi S_{xx})\}}$ , so that  $\phi = \xi\omega S_{xx}$  and  $\psi = \xi\omega^{-1}S_{yy}$ . The Jacobian is

$$\frac{\partial(\phi, \psi)}{\partial(\xi, \omega)} = \begin{vmatrix} \omega S_{xx} & \xi S_{xx} \\ \omega^{-1}S_{yy} & -\omega^{-2}\xi S_{yy} \end{vmatrix} = 2\omega^{-1}\xi S_{xx}S_{yy}$$

and hence

$$\begin{aligned} p(\xi, \omega, \rho | x, y) &\propto p(\phi, \psi, \rho | x, y) \left| \frac{\partial(\phi, \psi)}{\partial(\xi, \omega)} \right| \\ &\propto p(\rho)(1 - \rho^2)^{-(n-1)/2}\omega^{-1}\xi^{-n} \exp(-\frac{1}{2}R/\xi), \end{aligned}$$

where

$$R = (\omega + \omega^{-1} - 2\rho r)/(1 - \rho^2).$$

The substitution  $x = \frac{1}{2}R/\xi$  (so that  $\xi = \frac{1}{2}R/x$ ) reduces the integral over  $\xi$  to a standard gamma function integral, and hence we can deduce that

$$\begin{aligned} p(\omega, \rho | x, y) &\propto p(\rho)(1 - \rho^2)^{-(n-1)/2}\omega^{-1}R^{-(n-1)} \\ &\propto p(\rho)(1 - \rho^2)^{+(n-1)/2}\omega^{-1}(\omega + \omega^{-1} - 2\rho r)^{-(n-1)}. \end{aligned}$$

Finally, integrating over  $\omega$

$$p(\rho | x, y) \propto p(\rho)(1 - \rho^2)^{(n-1)/2} \int_0^\infty \omega^{-1}(\omega + \omega^{-1} - 2\rho r)^{-(n-1)} d\omega.$$

By substituting  $1/\omega$  for  $\omega$  it is easily checked that the integral from 0 to 1 is equal to that from 1 to  $\infty$ , so that as constant multiples are irrelevant, the lower limit of the integral can be taken to be 1 rather than 0.

By substituting  $\omega = \exp(t)$ , the integral can be put in the alternative form  
 $p(\rho|x, y) \propto p(\rho)(1 - \rho^2)^{(n-1)/2} \int_0^\infty (\cosh t - \rho r)^{-(n-1)} dt.$

The exact distribution corresponding to  $p(\rho) \propto 1$  has been tabulated in David (1954), but for most purposes it suffices to use an approximation. The usual way to proceed is by yet a further substitution, in terms of  $u$  where  $\cosh t - \rho r = (1 - \rho r)/(1 - u)$ , but this is rather messy and gives more than is necessary for a first-order approximation. Instead, note that for small  $t$   $\cosh t \cong 1 + \frac{1}{2}t^2$

while the contribution to the integral from values where  $t$  is large will, at least for large  $n$ , be negligible. Using this approximation

$$\begin{aligned} \int_0^\infty (\cosh t - \rho r)^{-(n-1)} dt &= \int_0^\infty \left(1 + \frac{1}{2}t^2 - \rho r\right)^{-(n-1)} dt \\ &= (1 - \rho r)^{-(n-1)} \int_0^\infty \left\{1 + \frac{1}{2}t^2/(1 - \rho r)\right\}^{-(n-1)} dt. \end{aligned}$$

On substituting

$$t = \sqrt{2(1 - \rho r)} \tan \theta$$

the integral is seen to be proportional to

$$(1 - \rho r)^{-n+(3/2)} \int_0^{\frac{1}{2}\pi} \cos^{n-3} \theta d\theta.$$

Since the integral in this last expression does not depend on  $\rho$ , we can conclude that  $p(\rho|x, y) \propto p(\rho) \frac{(1 - \rho^2)^{(n-1)/2}}{(1 - \rho r)^{n-(3/2)}}.$

Although evaluation of the constant of proportionality would still require the use of numerical methods, it is much simpler to calculate the distribution of  $\rho$  using this expression than to have to evaluate an integral for every value of  $\rho$ . In fact, the approximation is quite good [some numerical comparisons can be found in Box and Tiao (1992, Section 8.4.8)].

### 6.1.3 The hyperbolic tangent substitution

Although the exact mode does not usually occur at  $\rho = r$ , it is easily seen that for plausible choices of the prior  $p(\rho)$ , the approximate density derived earlier is greatest when  $\rho$  is near  $r$ . However, except when  $r = 0$ , this distribution is asymmetrical. Its asymmetry can be reduced by writing  $\rho = \tanh \zeta$ ,  $r = \tanh z$ , so that  $d\rho/d\zeta = \operatorname{sech}^2 \zeta$  and

$$1 - \rho^2 = \operatorname{sech}^2 \zeta, \quad 1 - \rho r = \frac{\cosh(\zeta - z)}{\cosh \zeta \cosh z}.$$

It follows that

$$p(\zeta|x, y) \propto p(\zeta) \cosh^{-5/2} \zeta \cosh^{(3/2)-n}(\zeta - z).$$

If  $n$  is large, since the factor  $p(\zeta) \cosh^{-5/2} \zeta \cosh^{3/2}(\zeta - z)$  does not depend on  $n$ , it may be regarded as approximately constant over the range over which  $\cosh^{-n}(\zeta - z)$  is appreciably different from zero, so that  $p(\zeta|x, y) \propto \cosh^{-n}(\zeta - z)$ .

Finally put

$$\xi = (\zeta - z)\sqrt{n}$$

and note that if  $\zeta$  is close to  $z$  then  $\cosh(\zeta - z) \cong 1 + \frac{1}{2}\xi^2/n + \dots$ . Putting this into the expression for  $p(\zeta|x, y)$  and using the exponential limit  $p(\xi|x, y) \propto \exp(-\frac{1}{2}\xi^2)$ , so that approximately  $\xi \sim N(0, 1)$ , or equivalently

$$\zeta \sim N(z, 1/n).$$

A slightly better approximation to the mean and variance can be found by using approximations based on the likelihood as in Section 3.10. If we take a uniform prior for  $\rho$  or at least assume that the prior does not vary appreciably

$$L(\zeta) = -\frac{5}{2} \log \cosh \zeta + \{\frac{3}{2} - n\} \log \cosh(\zeta - z)$$

$$L'(\zeta) = -\frac{5}{2} \tanh \zeta + \{\frac{3}{2} - n\} \tanh(\zeta - z)$$

over the range of values of interest, we get  $= -\frac{5}{2}\rho + \{\frac{3}{2} - n\}\{(\zeta - z) + \dots\}$ .

We can now approximate  $\rho$  by  $r$  (we could write  $\tanh \zeta = \tanh\{z + (\zeta - z)\}$  and so get a better approximation, but it is not worth it). We can also approximate  $n - \frac{3}{2}$  by  $n$ , so getting the root of the likelihood equation as  $\zeta = z - 5r/2n$ .

Further

$$\begin{aligned} L''(\zeta) &= -\frac{5}{2} \operatorname{sech}^2 + \{\frac{3}{2} - n\} \operatorname{sech}^2(\zeta - z) \\ &= -\frac{5}{2}(1 - \tanh^2 \zeta) + \{\frac{3}{2} - n\}\{1 + \dots\}, \end{aligned}$$

so that again approximating  $\rho$  by  $r$ , we have at  $\zeta = \hat{\zeta}$

$$L''(\hat{\zeta}) = -\frac{5}{2}(1 - r^2) + \frac{3}{2} - n.$$

It follows that the distribution of  $\zeta$  is given slightly more accurately by

$$\zeta \sim N\left(z - 5r/2n, \{n - \frac{3}{2} + \frac{5}{2}(1 - r^2)\}^{-1}\right).$$

This approximation differs a little from that usually given by classical statisticians, who usually quote the variance as  $(n-3)^{-1}$ , but the difference is not of great importance.

### 6.1.4 Reference prior

Clearly, the results will be simplest if the prior used has the form

$$p(\rho) \propto (1 - \rho^2)^c$$

for some  $c$ . The simplest choice is to take  $c = 0$ , that is, a uniform prior with  $p(\rho) \propto 1$ , and it seems quite a reasonable choice. It is possible to use the multi-parameter version of Jeffreys' rule to find a prior for  $(\phi, \psi, \rho)$ , though it is not wholly simple. The easiest way is to write  $\kappa = \rho\phi\psi$  for the covariance and to work in terms of the inverse of the variance–covariance matrix, that is, in terms

$$\text{of } (\alpha, \beta, \gamma), \text{ where } \begin{pmatrix} \alpha & \gamma \\ \gamma & \beta \end{pmatrix} = \begin{pmatrix} \phi & \kappa \\ \kappa & \psi \end{pmatrix}^{-1}.$$

It turns out that  $p(\alpha, \beta, \gamma) \propto \Delta^{3/2}$ , where  $\Delta$  is the determinant  $\phi\psi - \kappa^2$ , and that

$$\text{the Jacobian determinant } \frac{\partial(\alpha, \beta, \gamma)}{\partial(\phi, \psi, \kappa)} = -\Delta^{-3},$$

so that  $p(\phi, \psi, \kappa) \propto \Delta^{-3/2}$ . Finally, transforming to the parameters  $(\phi, \psi, \rho)$  that are really of interest, it transpires that  $p(\phi, \psi, \rho) \propto (\phi\psi)^{-1}(1 - \rho^2)^{-3/2}$

which corresponds to the choice  $c = -\frac{3}{2}$  and the standard reference priors for  $\phi$  and  $\psi$ .

### 6.1.5 Incorporation of prior information

It is not difficult to adapt the aforementioned analysis to the case where prior information from the conjugate family [i.e. inverse chi-squared for  $\phi$  and  $\psi$  and of the form  $(1 - \rho^2)^c$  for  $\rho$ ] is available. In practice, this information will usually be available in the form of previous measurements of a similar type and in this case it is best dealt with by transforming all the information about  $\rho$  into statements about  $\zeta = \tanh^{-1} \rho$ , so that the theory we have built up for the normal distribution can be used.

## 6.2 Examples on the use of the correlation coefficient

### 6.2.1 Use of the hyperbolic tangent transformation

The following data is a small subset of a much larger quantity of data on the length and breadth (in mm) of the eggs of cuckoos (*C. canorus*).

Egg no. ( $i$ ):	1	2	3	4	5	6	7	8	9
Length ( $x_i$ ):	22.5	20.1	23.3	22.9	23.1	22.0	22.3	23.6	24.7
Breadth ( $y_i$ ):	17.0	14.9	16.0	17.4	17.4	16.5	17.2	17.2	18.0

Here  $n = 9$ ,  $\bar{x} = 22.72$ ,  $\bar{y} = 16.84$ ,  $S_{xx} = 12.816$ ,  $S_{yy} = 6.842$ ,  $S_{xy} = 7.581$ ,  $r = 0.810$  and so

$z = \tanh^{-1} 0.810 = 1.127$  and  $1/n = 1/9$ . We can conclude that with 95% posterior probability  $\zeta$  is in the interval  $1.127 \pm 1.960 \times \sqrt{(1/9)}$ , that is,  $(0.474, 1.780)$ , giving rise to  $(0.441, 0.945)$  as a corresponding interval for  $\rho$ , using Lindley and Scott (1995, Table 17) or Neave (1978, Table 6.3).

## 6.2.2 Combination of several correlation coefficients

One of the important uses of the hyperbolic tangent transformation lies in the way in which it makes it possible to combine different observations of the correlation coefficient. Suppose, for example, that on one occasion we observe that  $r=0.7$  on the basis of 19 observations and on another we observe that  $r=0.9$  on the basis of 25 observations. Then after the first set of observations, our posterior for  $\zeta$  is  $N(\tanh^{-1} 0.7, 1/19)$ . The second set of observations now puts us into the situation of a normal prior and likelihood, so the posterior after all the observations is still normal, with variance  $(19 + 25)^{-1} = 0.0227$  and mean

$$0.0227(19 \times \tanh^{-1} 0.7 + 25 \tanh^{-1} 0.9) = 1.210$$

(using rounded values) that is,  $N(1.210, 0.0227)$ , suggesting a point estimate of  $\tanh 1.210 = 0.8367$  for  $\rho$ .

The transformation also allows one to investigate whether or not the correlations on the two occasions really were from the same population or at least from reasonably similar populations.

## 6.2.3 The squared correlation coefficient

There is a temptation to take  $r$  as such too seriously and to think that if it is very close to 1 then the two variables are closely related, but we will see shortly when we come to consider regression that  $r^2$ , which measures the proportion of the variance of one variable that can be accounted for by the other variable, is in many ways at least as useful a quantity to consider.

# 6.3 Regression and the bivariate normal model

## 6.3.1 The model

The problem we will consider in this section is that of using the values of one variable to explain or predict values of another. We shall refer to an *explanatory*

and a *dependent* variable, although it is conventional to refer to an independent and a dependent variable. An important reason for preferring the phrase explanatory variable is that the word ‘independent’ if used in this context has nothing to do with the use of the word in the phrase ‘independent random variable’. Some authors, for example, Novick and Jackson (1974, Section 9.1), refer to the dependent variable as the criterion variable. The theory can be applied, for example, to finding a way of predicting the weight (the dependent variable) of typical individuals in terms of their height (the explanatory variable). It should be noted that the relationship which best predicts weight in terms of height will not necessarily be the best relationship for predicting height in terms of weight.

The basic situation and notation are the same as in the last two sections, although in this case there is not the symmetry between the two variables that there was there. We shall suppose that the  $x$ s represent the explanatory variable and the  $y$ s the dependent variables.

There are two slightly different situations. In the first, the experimenters are free to set the values of  $x$ , whereas in the second both values are random, although one is thought of as having a causal or explanatory relationship with the other. The analysis, however, turns out to be the same in both cases.

The most general model is

$$p(x, y | \theta_1, \theta_2) \propto p(x|\theta_1)p(y|x, \theta_2),$$

where in the first situation described above  $\theta_1$  is a null vector and the distribution of  $x$  is degenerate. If it is assumed that  $\theta_1$  and  $\theta_2$  have independent priors, so that  $p(\theta_1, \theta_2) = p(\theta_1)p(\theta_2)$ , then  $p(\theta_1, \theta_2 | x, y) \propto p(\theta_1)p(\theta_2)p(x|\theta_1)p(y|x, \theta_2)$ .

It is now obvious that we can integrate over  $\theta_1$  to get

$$p(\theta_2 | x, y) \propto p(\theta_2)p(y|x, \theta_2).$$

Technically, given  $\theta_2$ , the vector  $x$  is *sufficient* for  $\theta_1$  and, given  $\theta_1$ , the vector  $x$  is *ancillary* for  $\theta_2$ . It follows that insofar as we wish to make inferences about  $\theta_2$ , we may act as if  $x$  were constant.

### 6.3.2 Bivariate linear regression

We will now move on to a very important particular case. Suppose that conditional on  $x$  we have  $y_i \sim N(\eta_0 + \eta_1 x_i, \phi)$ .

Thus,

$$\theta_2 = (\eta_0, \eta_1, \phi)$$

unless one or more of  $\eta_0$ ,  $\eta_1$  and  $\phi$  are known, in which case the ones that are

known can be dropped from  $\theta_2$ . Thus, we are supposing that, on average, the dependence of the ys on the xs is linear. It would be necessary to use rather different methods if there were grounds for thinking, for example, that  $E(y_i|x_i) = \eta_0 + \eta_1 x_i + \eta_2 x_i^2$  or that  $E(y_i|x_i) = \gamma_0 \cos(x_i + \gamma_1)$ . It is also important to suppose that the ys are *homoscedastic*, that is, that the variance  $V(y_i|x_i)$  has the same constant value  $\phi$  whatever the value of  $x_i$ ; modifications to the analysis would be necessary if it were thought that, for example,  $y_i \sim N(\eta_0 + \eta_1 x_i, x_i)$ , so that the variance increased with  $x_i$ .

It simplifies some expressions to write  $\eta_0 + \eta_1 x_i$  as  $\alpha + \beta(x_i - \bar{x})$  where, of course,  $\bar{x} = \sum x_i/n$ , so that  $\eta_0 = \alpha - \beta\bar{x}$  and  $\eta_1 = \beta$ , hence  $\alpha = \eta_0 + \eta_1 \bar{x}$ . The model can now be written as  $y_i \sim N(\alpha + \beta(x_i - \bar{x}), \phi)$ .

Because a key feature of the model is the *regression line*  $y = \alpha + \beta(x - \bar{x})$  on which the expected values lie, the parameter  $\beta$  is usually referred to as the *slope* and  $\alpha$  is sometimes called the *intercept*, although this term is also sometimes applied to  $\eta_0$ . For the rest of this section, we shall take a reference prior that is independently uniform in  $\alpha$ ,  $\beta$  and  $\log \phi$ , so that  $p(\alpha, \beta, \phi) \propto 1/\phi$ .

In addition to the notation used in Sections 6.1 and 6.2, it is helpful to define

$$S_{ee} = S_{yy} - S_{xy}^2/S_{xx} = S_{yy}(1 - r^2), \\ a = \bar{y}, \quad b = S_{xy}/S_{xx}, \quad e_0 = \bar{y} - b\bar{x}.$$

It then turns out that

$$p(\alpha, \beta, \phi | x, y) \propto p(\alpha, \beta, \phi) p(y | x, \alpha, \beta, \phi).$$

Now since  $\bar{y} - \alpha$  is a constant and  $\sum(x_i - \bar{x}) = \sum(y_i - \bar{y}) = 0$ , the sum of squares

$$\begin{aligned} \sum(y_i - \alpha - \beta(x_i - \bar{x}))^2 &= \sum((y_i - \bar{y}) + (\bar{y} - \alpha) - \beta(x_i - \bar{x}))^2 \\ &= S_{yy} + n(\bar{y} - \alpha)^2 + \beta^2 S_{xx} - 2\beta S_{xy} \\ &= S_{yy} - S_{xy}^2/S_{xx} + n(\alpha - \bar{y})^2 + S_{xx}(\beta - S_{xy}/S_{xx})^2 \\ &= S_{ee} + n(\alpha - a)^2 + S_{xx}(\beta - b)^2. \end{aligned}$$

can be written as

Thus, the joint posterior is

$$p(\alpha, \beta, \phi | x, y) \propto \phi^{-(n+2)/2} \exp[-\frac{1}{2}\{S_{ee} + n(\alpha - a)^2 + S_{xx}(\beta - b)^2\}/\phi].$$

It is now clear that for given  $b$  and  $\phi$  the posterior for  $\beta$  is  $N(b, \phi/S_{xx})$ , and so we can integrate  $\beta$  out to get  $p(\alpha, \phi | x, y) \propto \phi^{-(n+1)/2} \exp[-\frac{1}{2}\{S_{ee} + n(\alpha - a)^2\}/\phi]$

(note the change in the exponent of  $\phi$ ).

In Section 2.12 on ‘Normal mean and variance both unknown’, we showed that if

$$p(\theta, \phi) \propto \phi^{-(v+1)/2-1} \exp[-\frac{1}{2}\{S + n(\theta - \bar{x})^2\}/\phi]$$

and  $s^2 = S/v$  then

$$t = \frac{(\theta - \bar{x})}{s/\sqrt{n}} \sim t_v \quad \text{and} \quad \psi \sim S \chi_v^{-2}.$$

It follows from just the same argument that in this case the posterior for  $\alpha$  given

$$x \text{ and } y \text{ is such that if } s^2 = S_{ee}/(n-2) \text{ then } \frac{(\alpha - a)}{s/\sqrt{n}} \sim t_{n-2}.$$

Similarly the posterior of  $\beta$  can be found by integrating  $\alpha$  out to show that  
 $\frac{(\beta - b)}{s/\sqrt{S_{xx}}} \sim t_{n-2}$ .

Finally, note that

$$\phi \sim S_{ee} \chi_{n-2}^{-2}.$$

It should, however, be noted that the posteriors for  $\alpha$  and  $\beta$  are not independent, although they are independent for given  $\phi$ .

It may be noted that the posterior means of  $\alpha$  and  $\beta$  are  $a$  and  $b$  and that these are the values that minimize the sum of squares  $\sum \{y_i - \alpha - \beta(x_i - \bar{x})\}^2$

and that  $S_{ee}$  is the minimum sum of squares. This fact is clear because the sum is  $S_{ee} + n(\alpha - a)^2 + S_{xx}(\beta - b)^2$

and it constitutes the *principle of least squares*, for which reason  $a$  and  $b$  are referred to as the *least squares estimates* of  $\alpha$  and  $\beta$ . The regression line  $y = a + b(x - \bar{x})$ ,

which can be plotted for all  $x$  as opposed to just those  $x$ , observed, is called the *line of best fit* for  $y$  on  $x$ . The principle is very old; it was probably first published by Legendre but first discovered by Gauss; for its history see Harter (1974, 1975, 1976). It should be noted that the *line of best fit* for  $y$  on  $x$  is not, in general, the same as the line of best fit for  $x$  on  $y$ .

### 6.3.3 Example

This example goes to show that what I naively thought to be true of York's weather is, in fact, false. I guessed that if November was wet, the same thing would be true in December, and so I thought I would try and see how far this December's rainfall could be predicted in terms of November's. It turns out that the two are in fact *negatively* correlated, so that if November is very wet there is a slight indication that December will be on the dry side. However, the data (given in mm) serves quite as well to indicate the method.

Year( $i$ )	1971	1972	1973	1974	1975	1976	1977	1978	1979	1980
Nov( $x_i$ )	23.9	43.3	36.3	40.6	57.0	52.5	46.1	142.0	112.6	23.7
Dec( $y_i$ )	41.0	52.0	18.7	55.0	40.0	29.2	51.0	17.6	46.6	57.0

It turns out that  $\bar{x} = 57.8$ ,  $\bar{y} = 40.8$ ,  $S_{xx} = 13, 539$ ,  $S_{yy} = 1889$  and  $S_{xy} = -2178$ , so that

$$a = \bar{y} = 40.8, \quad b = S_{xy}/S_{xx} = -0.161, \quad r = S_{xy}/\sqrt{(S_{xx}S_{yy})} = -0.431$$

$$S_{ee} = S_{yy} - S_{xy}^2/S_{xx} = S_{yy}(1 - r^2) = 1538.$$

It follows that

$$s^2 = S_{ee}/(n - 2) = 192, \quad s/\sqrt{n} = 4.38, \quad s/\sqrt{S_{xx}} = 0.119.$$

Since the 75th percentile of  $t_8$  is 0.706, it follows that a 50% HDR for the intercept  $\alpha$  is  $40.8 \pm 0.706 \times 4.38$ , that is, (37.7, 43.9). Similarly, a 50% HDR for the slope  $\beta$  is  $-0.161 \pm 0.706 \times 0.119$ , that is, (-0.245, -0.077). Further, from tables of values of  $\chi^2$  corresponding to HDRs for  $\log \chi^2_8$ , an interval of posterior probability 50% for the variance  $\phi$  is from 1538/11.079 to 1538/5.552, that is, (139, 277).

Very often the slope  $\beta$  is of more importance than the intercept  $\alpha$ . Thus, in the above example, the fact that the slope is negative with high probability corresponds to the conclusion that high rainfall in November indicates that there is less likely to be high rainfall in December, as was mentioned earlier.

### 6.3.4 Case of known variance

If, which is not very likely, you should happen to know the variance  $\phi$ , the problem is even simpler. In this case, it is easy to deduce that (with the same

$$p(\alpha, \beta | x, y) \propto \exp \left[ -\frac{1}{2} \sum \{y_i - \alpha - \beta(x_i - \bar{x})\}^2 / \phi \right]$$

priors for  $\alpha$  and  $\beta$ )

$$\propto \exp \left[ -\frac{1}{2} \{n(\alpha - a)^2 + S_{xx}(\beta - b)^2\} / \phi \right].$$

It is clear that in this case the posteriors for  $\alpha$  and  $\beta$  are independent and such that  $\alpha \sim N(a, \phi/n)$  and  $\beta \sim N(b, \phi/S_{xx})$ .

### 6.3.5 The mean value at a given value of the explanatory variable

Sometimes there are other quantities of interest than  $\alpha$ ,  $\beta$  and  $\phi$ . For example, you might want to know what the expected value of  $y$  is at a given value of  $x$ . A particular case would arise if you wanted estimate the average weight of women of a certain height on the basis of data on the heights and weights of  $n$  individuals. Similarly, you might want to know about the value of the parameter  $\eta_0$  in the original formulation (which corresponds to the particular value  $x = 0$ ). Suppose that the parameter of interest is  $\gamma = \alpha + \beta(x_0 - \bar{x})$ .

Now we know that for given  $x$ ,  $y$ ,  $x_0$  and  $\phi$

$$\alpha \sim N(a, \phi/n) \quad \text{and} \quad \beta \sim N(b, \phi/S_{xx})$$

independently of one another. It follows that, given the same values,

$$\gamma \sim N(a + b(x_0 - \bar{x}), \phi\{n^{-1} + (x_0 - \bar{x})^2/S_{xx}\}).$$

It is now easy to deduce  $p(\gamma, \phi|x, y, x_0)$  from the fact that  $\phi$  has a (multiple of an) inverse chi-squared distribution. The same arguments used in Section 2.12 on ‘Normal mean and variance both unknown’ can be used to deduce that

$$\frac{\gamma - a - b(x_0 - \bar{x})}{s\sqrt{n^{-1} + (x_0 - \bar{x})^2/S_{xx}}} \sim t_{n-2}.$$

In particular, setting  $x_0=0$  and writing  $e_0 = a - b\bar{x}$  we get  $\frac{\eta_0 - e_0}{s\sqrt{n^{-1} + \bar{x}^2/S_{xx}}} \sim t_{n-2}$ .

### 6.3.6 Prediction of observations at a given value of the explanatory variable

It should be noted that if you are interested in the distribution of a potential observation at a value  $x=x_0$ , that is, the *predictive distribution*, then the result is slightly different. The mean of such observations conditional on  $x$ ,  $y$  and  $x_0$  is still  $a + b(x_0 - \bar{x})$ , but since  $y_0 - \alpha - \beta(x_0 - \bar{x}) \sim N(0, \phi)$

in addition to the above distribution for  $\gamma = \alpha + \beta(x_0 - \bar{x})$ , it follows that  $y_0 \sim N(a + b(x_0 - \bar{x}), \phi\{1 + n^{-1} + (x_0 - \bar{x})^2\}/S_{xx})$

and so on integrating  $\phi$  out

$$\frac{\gamma_0 - a - b(x_0 - \bar{x})}{s\sqrt{1 + n^{-1} + (x_0 - \bar{x})^2/S_{xx}}} \sim t_{n-2}.$$

### 6.3.7 Continuation of the example

To find the mean rainfall to be expected in December in a year when there are  $x_0=46.1$  mm in November, we first find  $a + b(x_0 - \bar{x}) = 42.7$  and  $n^{-1} + (x_0 - \bar{x})^2/S_{xx} = 0.110$ , and hence  $s\sqrt{n^{-1} + (x_0 - \bar{x})^2/S_{xx}} = 4.60$ . Then the distribution of the expected value  $\gamma$  at  $x=x_0$  is  $N(42.7, 4.60^2)$ . On the other hand, in single years in which the rainfall in November is 46.1, there is a greater variation in the December rainfall than the variance for the mean of  $4.60^2=21.2$  implies – in fact  $s\sqrt{1 + n^{-1} + (x_0 - \bar{x})^2/S_{xx}} = 14.6$  and the corresponding variance is  $14.6^2=213.2$ .

### 6.3.8 Multiple regression

Very often there is more than one explanatory variable, and we want to predict the value of  $y$  using the values of two or more variables  $x^{(1)}$ ,  $x^{(2)}$ , etc. It is not

difficult to adapt the method described earlier to estimate the parameters in a model such as  $y_i \sim N(\alpha + \beta^{(1)}(x_i^{(1)} - \bar{x}^{(1)}) + \beta^{(2)}(x_i^{(2)} - \bar{x}^{(2)}), \phi)$

although you will find some complications unless it happens that

$$\sum (x_i^{(1)} - \bar{x}^{(1)})(x_i^{(2)} - \bar{x}^{(2)}) = 0.$$

For this reason, it is best to deal with such *multiple regression* problems by using matrix analysis. Readers who are interested will find a brief introduction to this topic in Section 6.7, while a full account can be found in Box and Tiao (1992).

### 6.3.9 Polynomial regression

A difficult problem which will not be discussed in any detail is that of *polynomial regression*, that is, of fitting a model  $y \sim N(\eta_0 + \eta_1 x + \eta_2 x^2 + \dots + \eta_r x^r, \phi)$ ,

where all the parameters, *including the degree r* of the polynomial are unknown a priori. Some relevant references are Jeffreys (1961, Sections 5.9–5.92) and Sprent (1969, Sections 5.4, 5.5). There is also an interesting discussion in Meyer and Collier (1970, p. 114 *et seq.*) in which Lindley starts by remarking: I agree the problem of fitting a polynomial to the data is one that at the moment I can't fit very conveniently to the Bayesian analysis. I have prior beliefs in the smoothness of the polynomial. We need to express this idea quantitatively, but I don't know how to do it. We could bring in our prior opinion that some of the regression coefficients are very small.

Subsequently, a Bayesian approach to this problem has been developed by Young (1977).

## 6.4 Conjugate prior for the bivariate regression model

### 6.4.1 The problem of updating a regression line

In Section 6.3, we saw that with the regression line in the standard form  $y_i = \alpha + \beta(x_i - \bar{x})$  the joint posterior is  $p(\alpha, \beta, \phi | x, y) \propto \phi^{-(n+2)/2} \exp[-\frac{1}{2}\{S_{ee} + n(\alpha - a)^2 + S_{xx}(\beta - b)^2\}/\phi]$ .

For reasons that will soon emerge, we denote the quantities derived from the data with a prime as  $n'$ ,  $\bar{x}'$ ,  $\bar{y}'$ ,  $a'$ ,  $b'$ ,  $S'_{xx}$ ,  $S'_{ee}$ , etc. In the example on rainfall,

we found that  $n' = 10$  and

$$\bar{x}' = 57.8, \quad a' = \bar{y}' = 40.8, \quad b' = -0.161, \quad S'_{xx} = 13,539, \quad S'_{ee} = 1538.$$

Now suppose that we collect further data, thus

Year ( $i$ )	1981	1982	1983	1984	1985	1986
Nov ( $x_i$ )	34.1	62.0	106.9	34.1	68.3	81.0
Dec ( $y_i$ )	12.3	90.4	28.8	106.2	62.3	50.5

If this had been all the data available, we would have constructed a regression line based on data for  $n'' = 6$  years, with  $\bar{x}'' = 64.4, \quad a'' = \bar{y}'' = 58.4, \quad b'' = -0.381, \quad S''_{xx} = 3939, \quad S''_{ee} = 5815$ .

If, however, we had all 16 years data, then the regression line would have been based on data for  $n = 16$  years resulting in  $\bar{x} = 60.3, \quad a = \bar{y} = 47.4, \quad b = -0.184, \quad S_{xx} = 17641, \quad S_{ee} = 8841$ .

#### 6.4.2 Formulae for recursive construction of a regression line

By the sufficiency principle it must be possible to find  $\bar{x}, \bar{y}, b$ , etc., from  $\bar{x}', \bar{y}', b'$ , etc., and  $\bar{x}'', \bar{y}'', b''$ , etc. It is in fact not too difficult to show that  $n, \bar{x}$  and  $n = n' + n''$

$$\bar{x} = (n'\bar{x}' + n''\bar{x}'')/n$$

$$a = \bar{y} \text{ are given by } \bar{y} = (n'\bar{y}' + n''\bar{y}'')/n$$

and that if we define

$$n^h = (n'^{-1} + n''^{-1})^{-1}$$

$$S_{xx}^c = n^h(\bar{x}' - \bar{x}'')^2$$

$$S_{xy}^c = n^h(\bar{x}' - \bar{x}'')(\bar{y}' - \bar{y}'')$$

$$b^c = S_{xy}^c/S_{xx}^c$$

then  $S_{xx}$ ,  $b$  and  $S_{ee}$  are given by

$$S_{xx} = S'_{xx} + S''_{xx} + S_{xx}^c$$

$$b = (b'S'_{xx} + b''S''_{xx} + b^c S_{xx}^c)/S_{xx}$$

$$S_{ee} = S'_{ee} + S''_{ee} + [(b' - b'')^2 S'_{xx} S''_{xx} + (b'' - b^c)^2 S''_{xx} S_{xx}^c + (b^c - b')^2 S_{xx}^c S'_{xx}]/S_{xx}.$$

Of these formulae, the only one that is at all difficult to deduce is the last, which

$$S_{ee} = S_{yy} - b^2 S_{xx}$$

$$= S'_{yy} + S''_{yy} + S_{yy}^c - b^2 S_{xx}$$

$$\text{is established thus } = S'_{ee} + S''_{ee} + b'^2 S'_{xx} + b'' S''_{xx} + b^{c2} S_{xx}^c - b^2 S_{xx}$$

(it is easily checked that there is no term  $S_{ee}^c$ ). However,

$$\begin{aligned}
S_{xx}(b'^2 S'_{xx} + b''^2 S''_{xx} + b^c S^c_{xx} - b^2 S_{xx}) \\
&= (S'_{xx} + S''_{xx} + S^c_{xx})(b'^2 S'_{xx} + b''^2 S''_{xx} + b^c S^c_{xx}) - (b'S'_{xx} + b''S''_{xx} + b^c S^c_{xx})^2 \\
&= b'^2 S'_{xx} S''_{xx} + b''^2 S''_{xx} - 2b'b'' S'_{xx} S''_{xx} \\
&\quad + b''^2 S''_{xx} S^c_{xx} + b^c S^c_{xx} S''_{xx} - 2b''b^c S''_{xx} S^c_{xx} \\
&\quad + b^c S^c_{xx} S'_{xx} + b'^2 S^c_{xx} S'_{xx} - 2b^c b' S^c_{xx} S'_{xx} \\
&= (b' - b'')^2 S'_{xx} S''_{xx} + (b'' - b^c)^2 S''_{xx} S^c_{xx} + (b^c - b')^2 S^c_{xx} S'_{xx}
\end{aligned}$$

giving the result.

With the data in the example, it turns out that  $n = n' + n'' = 16$  and  $\bar{x}$  (a weighted mean of  $\bar{x}'$  and  $\bar{x}''$ ) is 60.3, and similarly  $\bar{y} = 47.4$ . Moreover  $n^h = 3.75$ ,  $S^c_{xx} = 163.35$ ,  $S^c_{xy} = 435.6$ ,  $b^c = 2.667$ , so that

$$S_{xx} = 17641, \quad b = -0.184, \quad S_{ee} = 8841$$

in accordance with the results quoted earlier obtained by considering all 16 years together.

### 6.4.3 Finding an appropriate prior

In the aforementioned analysis, our prior knowledge could be summarized by saying that if the regression line is put in the form  $y_i = \alpha + \beta(x_i - \bar{x}')$  then

$$p(\alpha, \beta, \phi) \propto \phi^{-(n'+2)/2} \exp[-\frac{1}{2}\{S'_{ee} + n'(\alpha - a')^2 + S'_{xx}(\beta - b')^2\}/\phi].$$

We then had observations (denoted  $x''$  and  $y''$ ) that resulted in a posterior which is such that if the regression line is put in the form  $y_i = \alpha + \beta(x_i - \bar{x})$  then

$$p(\alpha, \beta, \phi | x'', y'') \propto \phi^{-(n+2)/2} \exp[-\frac{1}{2}\{S'_{ee} + n(\alpha - a)^2 + S_{xx}(\beta - b)^2\}/\phi].$$

This, of course, gives a way of incorporating prior information into a regression model provided that it can be put into the form which occurs above. It is, however, often quite difficult to specify prior knowledge about a regression line unless, as in the case above, it is explicitly the result of previous data. Appropriate questions to ask to fix which prior of the class to use are as follows:

1. What number of observations is my present knowledge worth? Write the answer as  $n'$ .
2. What single point is the regression line most likely to go through? Write the answer as  $(\bar{x}', \bar{y}')$ .
3. What is the best guess as to the slope of the regression line? Write the answer as  $b'$ .

4. What is the best guess as to the variance of the observation  $y_i$  about the regression line? Write the answer as  $s'^2$  and find  $S'_{ee}$  as  $(n' - 2)s'^2$ .

5. Finally make  $S'_{xx}$  such that the estimated variances for the slope  $\beta$  and the intercept  $\alpha$  are in the ratio  $n'$  to  $S'_{xx}$ .

As noted above, it is difficult to believe that this process can be carried out in a very convincing manner, although the first three steps do not present as much difficulty as the last two. However, the case where information is received and then more information of the same type is used to update the regression line can be useful.

It is of course possible (and indeed simpler) to do similar things with the correlation coefficient.

## 6.5 Comparison of several means – the one way model

### 6.5.1 Description of the one way layout

Sometimes we want to compare more than two samples. We might, for example, wish to compare the performance of children from a number of schools at a standard test. The usual model for such a situation is as follows. We suppose that  $\theta = (\theta_1, \theta_2, \dots, \theta_I)$  is a vector of unknown parameters and that there are  $N = \sum K_i$  independent observations  $x_{ik} \sim N(\theta_i, \phi)$  ( $i = 1, 2, \dots, I; k = 1, 2, \dots, K_i$ )

from  $I$  independent populations with, however, a common variance  $\phi$ . For simplicity, we shall assume independent reference priors uniform in  $\theta_1, \theta_2, \dots, \theta_I$  and  $\log \phi$ , that is  $p(\theta, \phi) \propto 1/\phi$ .

The likelihood is

$$(2\pi\phi)^{-N/2} \exp(-\frac{1}{2}S/\phi),$$

where

$$S = \sum_i \sum_k (x_{ik} - \theta_i)^2$$

and so the posterior is

$$p(\theta, \phi|x) \propto \phi^{-N/2-1} \exp(-\frac{1}{2}S/\phi).$$

It is useful to define the following notation

$$x_{i*} = \sum_k x_{ik}/K_i$$

$$x_{**} = \sum_i \sum_k x_{ik}/N$$

$$\lambda = \sum K_i \theta_i / N = \sum \theta_i / I \quad \text{if } K_i = n \text{ for all } i$$

$$\mu_i = \theta_i - \lambda$$

$$\hat{\lambda}_i = x_{**}$$

$$\hat{\mu} = x_{i*} - x_{**}.$$

The reason for thinking of the  $\mu_i$  is that we are often concerned as to whether all the  $\theta_i$  are equal. If, for example, the  $x_{ik}$  represent yields of wheat on fields on which  $I$  different fertilizers have been used, then we are likely to be interested in whether the yields are on average all equal (or nearly so), that is,  $\theta_1 = \theta_2 = \dots = \theta_I$  or equivalently whether or not  $\mu_1 = \mu_2 = \dots = \mu_I = 0$ .

The  $\mu_i$  satisfy the condition

$$\sum_i K_i \mu_i = 0,$$

so that if we know the values of  $\mu_1, \dots, \mu_{I-1}$  we automatically know

$$\mu = (\mu_1, \dots, \mu_I).$$

Similarly the  $\hat{\mu}_i$  satisfy  $\sum K_i \hat{\mu}_i = 0$ .

## 6.5.2 Integration over the nuisance parameters

Since the Jacobian determinant of the transformation which takes  $\theta, \phi$  to  $\lambda, \mu_1, \mu_2, \dots, \mu_{I-1}, \phi$  consists of entries all of which are  $1/n$ , 1 or 0, its value is a

$$p(\lambda, \mu, \phi | x) \propto p(\lambda, \mu_1, \mu_2, \dots, \mu_{I-1}, \phi | x)$$

constant, and so

$$= p(\theta, \phi | x).$$

The thing to do now is to re-express  $S$  in terms of  $\lambda, \mu, \phi$ . Since  $x_{i*} = \hat{\lambda} + \hat{\mu}_i$  and  $\theta_i = \lambda + \mu_i$ , it follows that  $-(x_{ik} - \theta_i) = (\lambda - \hat{\lambda}) + (\mu_i - \hat{\mu}_i) + (x_{i*} - x_{ik})$ .

It is easily checked that sums of products of terms on the right vanish, and so it

$$S = \sum \sum (x_{ik} - \theta_i)^2$$

easily follows that  $= N(\lambda - \hat{\lambda})^2 + S_t(\mu) + S_e$ ,

where

$$S_t(\mu) = \sum K_i (\mu_i - \hat{\mu})^2, \quad S_e = \sum \sum (x_{i*} - x_{ik})^2.$$

It is also useful to define

$$v = N - I, \quad s^2 = S_e/v.$$

It follows that the posterior may be written in the form

$$p(\lambda, \mu, \phi | x) \propto \phi^{-N/2-1} \exp[-\frac{1}{2}\{N(\lambda - \hat{\lambda})^2 + S_t(\mu) + S_e\}/\phi].$$

As explained earlier, the value of  $\lambda$  is not usually of any great interest, and it is easily integrated out to give  $p(\mu, \phi | \mathbf{x}) \propto \phi^{-(N+1)/2} \exp[-\frac{1}{2}\{S_t(\mu) + S_e\}/\phi]$ .

The variance  $\phi$  can now be integrated out in just the same way as it was in Section 2.12 on ‘Normal mean and variance both unknown’ by reducing to a standard gamma function integral. The result is that

$$p(\mu | \mathbf{x}) \propto \{S_t(\mu) + S_e\}^{-(N-1)/2} \\ \propto \{1 + (I - 1)F(\mu)/v\}^{-(N-1)/2},$$

where

$$F(\mu) = \frac{S_t(\mu)/(I - 1)}{S_e/v} = \frac{\sum K_i(\mu_i - \hat{\mu}_i)^2/(I - 1)}{s^2}.$$

This is similar to a result obtained in one dimension (see Section 2.12 again; the situation there is not quite that we get by setting  $I = 1$  here because here  $\lambda$  has been integrated out). In that case we deduced that  $p(\mu | \mathbf{x}) \propto \{1 + t^2/v\}^{-(v+1)/2}$ ,

where

$$t^2 = \frac{K(\mu - \bar{x})^2}{s^2}.$$

By analogy with that situation, the posterior distribution for  $\mu$  is called the multivariate t distribution. It was discovered independently by Cornish (1954 and 1955) and by Dunnett and Sobel (1954). The constant of proportionality can be evaluated, but we will not need to use it.

It should be clear that the density is a maximum when  $\mu = \hat{\mu}$  and decreases as the distance from  $\mu$  to  $\hat{\mu}$ , and indeed an HDR for  $\mu$  is clearly a hyperellipsoid centred on  $\hat{\mu}$ , that is, it is of the form  $E(F) = \{\mu; F(\mu) \leq F\}$

in which the length of each of the axes is in a constant ratio to  $\sqrt{F}$ .

To find an HDR of any particular probability it therefore suffices to find the distribution of  $F(\mu)$ , and since  $F(\mu)$  is a ratio of sums of squares divided by appropriate numbers of degrees of freedom it seems reasonable to conjecture that  $F(\mu) \sim F_{I-1, v}$

which is indeed so.

### 6.5.3 Derivation of the F distribution

It is not really necessary to follow this proof that  $F(\mu)$  really has got an F distribution, but it is included for completeness.

$$\begin{aligned}
P\{F \leq F(\mu) \leq F + dF\} &= \int_{E(F+dF) \setminus E(F)} p(\mu|x) d\mu \\
&= \int_{E(F+dF) \setminus E(F)} \{1 + (I-1)F(\mu)/v\}^{-(N-1)/2} d\mu \\
&= \{1 + (I-1)F(\mu)/v\}^{-(N-1)/2} \int_{E(F+dF) \setminus E(F)} d\mu \\
&= \{1 + (I-1)F(\mu)/v\}^{-(N-1)/2} [V(F+dF) - V(F)],
\end{aligned}$$

where  $V(F)$  is the volume of the hyperellipsoid  $E(F)$ . At first sight it appears that this is  $I$ -dimensional, but because  $\sum K_i(\mu_i - \hat{\mu}_i) = 0$  it represents the intersection of a hyperellipsoid in  $I$  dimensions with a hyperplane through its centre, which is a hyperellipsoid in  $(I-1)$  dimensions. If this is not clear, it may help to note that an ordinary sphere in three-dimensional space cuts a plane in a circle, that is, a sphere in  $3-1=2$  dimensions. It follows that  $V(F) \propto (\sqrt{F})^{I-1} = F^{(I-1)/2}$ , and hence

$$V(F+dF) - V(F) \propto (F+dF)^{(I-1)/2} - F^{(I-1)/2} \propto F^{(I-1)/2-1} dF.$$

It follows that the density of  $F(\mu)$  is proportional to

$$F^{(I-1)/2-1} \{1 + (I-1)F/v\}^{-(N-1)/2}.$$

Comparing this with the standard form in Appendix A and noting that  $I-1+\nu=N-1$  it can be seen that indeed  $F(\mu) \sim F_{I-1,\nu}$ , as asserted.

## 6.5.4 Relationship to the analysis of variance

This relates to the classical approach to the one-way layout. Note that if

$$F(\mathbf{0}) = \frac{\sum K_i \hat{\mu}_i^2 / (I-1)}{s^2} = \frac{\sum K_i (x_{i\cdot} - x_{..})^2 / (I-1)}{\sum \sum (x_{ik} - x_{i\cdot})^2 / \nu},$$

then  $F(\mu) = F(\mathbf{0})$  at the point  $\mu = 0$  which represents no treatment effect. Consequently if  $\pi(\mu_0) = P\{F(\mu) \leq F(\mu_0)\}$ ,

then  $\pi(0)$  is the probability of an HDR which just includes  $\mu = 0$ . It is thus possible to carry out a significance test at level  $\alpha$  of the hypothesis that  $\mu = 0$  in the sense of Section 4.3 on ‘Lindley’s method’ by rejecting if and only if  $\pi(0) \geq 1 - \alpha$ .

This procedure corresponds exactly to the classical analysis of variance (ANOVA) procedure in which you construct a table as follows. First find

$$S_T = \sum \sum (x_{ik} - x_{..})^2,$$

$$S_t(\mathbf{0}) = \sum K_i (x_{i\cdot} - x_{..})^2.$$

It is convenient to write  $S_t$  for  $S_t(0)$ . Then find  $S_e$  by subtraction as it is easily shown that  $S_T = S_t + S_e$ .

In computing, it should be noted that it makes no difference if a constant is subtracted from each of the  $x_{ik}$  and that  $S_T$  and  $S_e$  can be found by

$$S_T = \sum \sum x_{ik}^2 - Nx_{..}^2 = \sum x_{ik}^2 - C,$$

$$S_e = \sum K_i x_{i..}^2 - Nx_{..}^2 = \sum T_i^2 / K_i - C,$$

where  $T_i = \sum x_{ik} = K_i x_{i..}$  is the total for treatment  $i$ ,  $G = \sum \sum x_{ik} = Nx_{..}$  is the grand total, and  $C=G^2/N$  is the ‘correction for error’. (Note that these formulae are subject to rounding error if used incautiously.) The value of  $F(0)$  is then found easily by setting out a table as follows: **ANOVA Table**

Source	Sum of squares	Degrees of freedom	Mean square	Ratio
Treatments	$S_t$	$I - 1$	$S_t/(I - 1)$	$F(0)$
Error	$S_e$	$v = N - I$	$s^2 = S_e/v$	
TOTAL	$S_T$	$N - 1$		

We will now consider an example.

### 6.5.5 Example

Cochran and Cox (1957, Section 4.13) quote the following data from an experiment on the effect of sulphur in reducing scab disease in potatoes. In addition to untreated plots which serve as a control, three amounts of dressing were compared: 300, 600 and 1200 pounds per acre. Both an autumn and a spring application of each treatment were tried, so that in all there were seven distinct treatments. The effectiveness of the treatments were measured by the ‘scab index’, which is (roughly speaking) the average percentage of the area of 100 potatoes taken at random from each plot that is affected with scab. The data

i	Treatment	$K_i$	Scab indexes $x_{ik}$		$T_i$	$x_{i..}$	$\hat{\mu}_i$
1	O	8	12 30 10 18 24 32 29	26	181	22.6	7.0
2	A3	4	9 9 16 4		38	9.5	-6.2
3	S3	4	30 7 21 9		67	16.8	1.1
4	A6	4	16 10 18 18		62	15.5	-0.2
5	S6	4	18 24 12 19		73	18.2	2.6
6	A12	4	10 4 4 5		23	5.8	-9.9
7	S12	4	17 7 16 17		57	14.2	-1.2

are as follows:

1	O	8	12 30 10 18 24 32 29	26	181	22.6	7.0
2	A3	4	9 9 16 4		38	9.5	-6.2
3	S3	4	30 7 21 9		67	16.8	1.1
4	A6	4	16 10 18 18		62	15.5	-0.2
5	S6	4	18 24 12 19		73	18.2	2.6
6	A12	4	10 4 4 5		23	5.8	-9.9
7	S12	4	17 7 16 17		57	14.2	-1.2

There are  $I = 7$  treatments and  $N = \sum K_i = 32$  observations, the grand total being  $G = 501$  (and the grand average  $x_{..}$  being 15.66), the crude sum of squares being  $\sum \sum x_{ik}^2 = 9939$  and the correction for error  $C=G^2/N=7844$ . Further  $S_T = 9939 - 7844 = 2095$

$S_t = 181^2/8 + (38^2 + 67^2 + 62^2 + 73^2 + 23^2 + 57^2)/4 - 7844 = 972$ ,  
and hence the analysis of variance table is as follows:

## ANOVA Table

Source	Sum of squares	Degrees of freedom	Mean square	Ratio
Treatments	972	6	162	3.60
Error	1123	25	45	
TOTAL	2095	31		

From tables of the F distribution an  $F_{6,25}$  variable exceeds 3.63 with probability 0.01. Consequently a 99% HDR is  $E(3.63) = \{\mu; F(\mu) \leq 3.63\}$ , so that  $\pi(0) \cong 0.99$  and, according to the methodology of Lindley's method, as described in Section 4.3, the data is very nearly enough to cause the null hypothesis of no treatment effect to be rejected at the 1% level.

The 99% HDR can be re-expressed by noting that  $\mu$  is in it if and only if  $F(\mu) \leq 3.63$  or  $\sum K_i(\mu_i - \hat{\mu}_i)^2 \leq 3.63s^2 = 3.63 \times 45 = 163$

that is, if and only if

$$2(\mu_1 - \hat{\mu}_1)^2 + \sum_{i=2}^7 (\mu_i - \hat{\mu}_i)^2 \leq 41.$$

It is of course difficult to visualize such sets, which is one reason why the significance test mentioned earlier is helpful in giving some ideas as to what is going on. However, as was explained when significance tests were first introduced, they should not be taken too seriously – in most cases, you would expect to see a treatment effect, even if only a small one. One point is that you can get some idea of the size of the treatment effect from the significance level.

## 6.5.6 Relationship to a simple linear regression model

A way of visualizing the analysis of variance in terms of the simple linear regression model was pointed out by Kelley (1927, p. 178); see also Novick and Jackson (1974, Section 4–7).

Kelley's work is relevant to a *random effects model* (sometimes known as a components of variance model or Model II for the analysis of variance). An idea of what this is can be gained by considering an example quoted by Scheffé (1959, Section 7.2). Suppose a machine is used by different workers on different days, being used by worker  $i$  on  $K_i$  days for  $i = 1, 2, \dots, I$ , and that the output when worker  $i$  uses it on day  $k$  is  $x_{ik}$ . Then it might be reasonable to suppose that  $x_{ik} = m_i + e_{ik}$ ,

where  $m_i$  is the ‘true’ mean for the  $i$ th worker and  $e_{ik}$  is his ‘error’ on the  $k$ th day.

We could then assume that the  $I$  workers are a random sample from a large labour pool, instead of contributing fixed if unknown effects. In such a case, all of our knowledge of the  $x_{ik}$  contributes to knowledge of the distribution of the  $m_i$ , and so if we want to estimate a particular  $m_i$  we should take into account the observations  $x_{i'k}$  for  $i' \neq i$  as well as the observations  $x_{ik}$ . Kelley's suggestion is that we treat the individual measurements  $x_{ik}$  as the explanatory variable and the treatment means  $x_{i\cdot}$  as the dependent variable, so that the model to be fitted is  $x_{i\cdot} = \eta_0 + \eta_1 x_{ik} + \varepsilon_{ik}$ ,

where the  $\varepsilon_{ik}$  are error terms of mean zero, or equivalently  $x_{i\cdot} = \alpha + \beta(x_{ik} - x_{\cdot\cdot})$ .

In terms of the notation, we used in connection with simple linear regression

$$S_{xx} = \sum \sum (x_{ik} - x_{\cdot\cdot})^2 = S_T$$

$$S_{yy} = \sum \sum (x_{i\cdot} - x_{\cdot\cdot})^2 = \sum K_i (x_{i\cdot} - x_{\cdot\cdot})^2 = S_t$$

$$S_{xy} = \sum \sum (x_{ik} - x_{\cdot\cdot})(x_{i\cdot} - x_{\cdot\cdot}) = \sum K_i (x_{i\cdot} - x_{\cdot\cdot})^2 = S_t$$

$$r = \frac{S_{xy}}{\sqrt{(S_{xx} S_{yy})}} = \sqrt{(S_t / S_T)}$$

$$1 - r^2 = (S_T - S_t) / S_T = S_e / S_T$$

$$S_{ee} = S_{yy}(1 - r^2) = S_t S_e / S_T.$$

In accordance with the theory of simple linear regression, we estimate  $\alpha$  and  $\beta$  by, respectively,  $a = x_{\cdot\cdot}$  and  $b = S_{xy} / S_{xx} = S_t / S_T$ ,

so that the regression line takes the form

$$x_{i\cdot} = (S_t x_{ik} + S_e x_{\cdot\cdot}) / S_T.$$

The point of this formula is that if you were to try one single replicate with another broadly similar treatment to those already tried, you could estimate the overall mean for that treatment not simply by the one observation you have for that treatment, but by a weighted mean of that observation and the overall mean of all observations available to date.

## 6.5.7 Investigation of contrasts

Often in circumstances where the treatment effect does not appear substantial you may want to make further investigations. Thus, in the aforementioned example about sulphur treatment for potatoes, you might want to see how the effect of any sulphur compares with none, that is, you might like an idea of the

size of  $d = \sum_{i=2}^7 \theta_i / 6 - \theta_1 = \sum_{i=2}^7 \mu_i / 6 - \mu_1$ .

More generally, it may be of interest to investigate any *contrast*, that is, any

linear combination  $d = \sum \delta_i \mu_i$  where  $\sum \delta_i = 0$ .

If we then write  $\hat{d} = \sum \delta_i \hat{\mu}_i$  and

$$K_d = \left( \sum \delta_i^2 / K_i \right)^{-1},$$

then it is not difficult to show that we can write

$$\begin{aligned} S_t(\mu) &= \sum K_i (\mu_i - \hat{\mu}_i)^2 \\ &= K_d (d - \hat{d})^2 + S'_t(\mu'), \end{aligned}$$

where  $S'_t$  is a quadratic much like  $S_t$  except that has one less dimension and  $\mu'$  consists of linear combinations of  $\mu - \mu'$ . It follows that  $p(\mu, \phi | x) \propto \phi^{-(N+1)/2} \exp[-\frac{1}{2}\{K_d(d - \hat{d})^2 + S'_t(\mu') + S_e\}/\phi]$ .

It is then possible to integrate over the  $I-2$  linearly independent components of  $\mu'$  to get  $p(d, \phi | x) \propto \phi^{-(N-I+3)/2} \exp[-\frac{1}{2}\{K_d(d - \hat{d})^2 + S_e\}/\phi]$ ,

and then to integrate  $\phi$  out to give

$$\begin{aligned} p(d|x) &\propto \{K_d(d - \hat{d})^2 + S_e\}^{-(N-I+1)/2} \\ &\propto \{1 + t^2/\nu\}^{-(\nu+1)/2} \end{aligned}$$

(remember that  $\nu = N - I$ ), where

$$t = \sqrt{K_d}(d - \hat{d})/s.$$

It follows that  $t \sim t_\nu$ .

For example, in the case of the contrast concerned with the main effect of sulphur,  $d = -14/6 - 7 = -9.3$  and  $K_d = \{6(1/6)^2/4 + 1^2/8\}^{-1} = 6$ , so that

$$\frac{\sqrt{K_d}(d - \hat{d})}{s} = \frac{\sqrt{6}(d + 9.33)}{\sqrt{45}} = \frac{d + 9.33}{2.74} \sim t_\nu,$$

so that, for example, as a  $t_{25}$  random variable is less than 2.060 in modulus with probability 0.95, a 95% HDR for  $d$  is between  $-9.33 \pm 2.74 \times 2.060$ , that is,  $(-15.0, -3.7)$ .

## 6.6 The two way layout

### 6.6.1 Notation

Sometimes we come across observations which can be classified in two ways. We shall consider a situation in which each of a number of treatments is applied to a number of plots in each of a number of blocks. However, the terminology need not be taken literally; the terms treatments and blocks are purely conventional and could be interchanged in what follows. A possible application is to the yields on plots on which different varieties of wheat have been sown

and to which different fertilizers are applied, and in this case either the varieties or the fertilizers could be termed treatments or blocks. Another is to an analysis of rainfall per hour, in which case the months of the year might be treated as the blocks and the hours of the day as the treatments (or vice versa).

We consider the simplest situation in which we have  $N=IJK$  observations  $x_{ijk} \sim N(\theta_{ij}, \phi)$  ( $i = 1, 2, \dots, I; j = 1, 2, \dots, J; k = 1, 2, \dots, K$ )

to which  $I$  treatments have been applied in  $J$  blocks, the observations having a common variance  $\phi$ . For simplicity, we assume independent reference priors uniform in the  $\theta_{ij}$  and  $\log \phi$ , that is,  $p(\theta, \phi) \propto 1/\phi$ .

The likelihood is

$$(2\pi\phi)^{-N/2} \exp(-\frac{1}{2}S/\phi),$$

where

$$S = \sum_i \sum_j \sum_k (x_{ijk} - \theta_{ij})^2$$

and so the posterior is

$$p(\theta, \phi | x) \propto \phi^{-N/2-1} \exp(-\frac{1}{2}S/\phi).$$

As in Section 6.5, we shall use dots to indicate averaging over suffices, so, for example,  $x_{\cdot j \cdot}$  is the average of  $x_{ijk}$  over  $i$  and  $k$  for fixed  $j$ . We write

$$\begin{aligned} \lambda &= \theta_{\cdot \cdot} & \hat{\lambda} &= x_{\cdot \cdot \cdot} \\ \tau_i &= \theta_{i \cdot} - \theta_{\cdot \cdot} & \hat{\tau}_i &= x_{i \cdot \cdot} - x_{\cdot \cdot \cdot} \\ \beta_j &= \theta_{\cdot j} - \theta_{\cdot \cdot} & \hat{\beta}_j &= x_{\cdot j \cdot} - x_{\cdot \cdot \cdot} \\ \kappa_{ij} &= \theta_{ij} - \theta_{i \cdot} - \theta_{\cdot j} + \theta_{\cdot \cdot} & \hat{\kappa}_{ij} &= x_{ij \cdot} - x_{i \cdot \cdot} - x_{\cdot j \cdot} + x_{\cdot \cdot \cdot} \end{aligned}$$

It is conventional to refer to  $\tau_i$  as the *main effect* of treatment  $i$  and to  $\beta_j$  as the main effect of block  $j$ . If  $\tau_i = 0$  for all  $i$  there is said to be no main effect due to treatments; similarly if  $\beta_j = 0$  for all  $j$  there is said to be no main effect due to blocks. Further,  $\kappa_{ij}$  is referred to as the interaction of the  $i$ th treatment with the  $j$ th block, and if  $\kappa_{ij} = 0$  for all  $i$  and  $j$  there is said to be no interaction between treatments and blocks. Note that the parameters satisfy the conditions  $\sum \tau_i = \sum \beta_j = 0$ ,

so that  $\tau$  is  $(I-1)$ -dimensional and  $\beta$  is  $(J-1)$ -dimensional. Similarly for all  $j$   $\sum_i \kappa_{ij} = 0$

and for all  $i$

$$\sum_j \kappa_{ij} = 0.$$

Because both of these imply  $\sum \sum \kappa_{ij} = 0$ , there are only  $I+J-1$  linearly independent constraints, so that  $\kappa$  is  $IJ-I-J+1=(I-1)(J-1)$ -dimensional.

## 6.6.2 Marginal posterior distributions

Because

$$-(x_{ijk} - \theta_{ij}) = (\lambda - \hat{\lambda}) + (\tau_i - \hat{\tau}_i) + (\beta_j - \hat{\beta}_j) + (\kappa_{ij} - \hat{\kappa}_{ij}) + (x_{ij*} - x_{ijk}),$$

the sum of squares  $S$  can be split up in a slightly more complicated way than in Section 6.5 as  $S = N(\lambda - \hat{\lambda})^2 + S_t(\tau) + S_b(\beta) + S_{tb}(\kappa) + S_e$ ,

where

$$S_t(\tau) = JK \sum (\tau_i - \hat{\tau}_i)^2$$

$$S_b(\beta) = IK \sum (\beta_j - \hat{\beta}_j)^2$$

$$S_{tb}(\kappa) = K \sum \sum (\kappa_{ij} - \hat{\kappa}_{ij})^2$$

$$S_e = \sum \sum \sum (x_{ij*} - x_{ijk})^2.$$

It is also useful to define

$$\nu = I(J-1) \quad \text{and} \quad s^2 = S_e/\nu.$$

After a change of variable as in Section 6.6, the posterior can be written as

$$p(\lambda, \tau, \beta, \kappa, \phi | x) \propto \phi^{-N/2-1} \exp[-\frac{1}{2}\{N(\lambda - \hat{\lambda})^2 + S_t(\tau) + S_b(\beta) + S_{tb}(\kappa) + S_e\}/\phi],$$

so that on integrating  $\lambda$  out

$$p(\tau, \beta, \kappa, \phi | x) \propto \phi^{-(N+1)/2-1} \exp[-\frac{1}{2}\{S_t(\tau) + S_b(\beta) + S_{tb}(\kappa) + S_e\}/\phi].$$

To investigate the effects of the treatments, we can now integrate over the  $\{(J-1)+(I-1)(J-1)\}$ -dimensional space of values of  $(\beta, \kappa)$  to get  $p(\tau, \phi | x) \propto \phi^{-(\nu+I+1)/2-1} \exp[-\frac{1}{2}\{S_t(\tau) + S_e\}/\phi]$ .

We can now integrate  $\phi$  out just as in Section 6.6 to get

$$p(\tau | x) \propto \{S_t(\tau) + S_e\}^{-(\nu+I-1)/2} \propto \{1 + (I-1)F_t(\tau)/\nu\}^{-(\nu+I-1)/2},$$

where

$$F_t(\tau) = \frac{S_t(\tau)/(I-1)}{S_e/\nu}.$$

Again as in Section 6.6, it can be shown that

$$F_t(\tau) \sim F_{I-1, \nu}$$

and so it is possible to conduct significance tests by Lindley's method and to find HDRs for  $\tau$ .

Similarly, it can be shown that  $F_b(\beta)$  (defined in the obvious way) is distributed as  $F_b(\beta) \sim F_{J-1, \nu}$ . Moreover, if  $F_{tb}(\kappa) = \frac{S_{tb}(\kappa)/(I-1)(J-1)}{S_e/\nu}$ , then  $F_{tb}(\kappa) \sim F_{(I-1)(J-1), \nu}$ . Thus, it is also possible to investigate the blocks effect and the interaction. It should be noted that it would rarely make sense to

believe in the existence of an interaction unless both main effects were there.

### 6.6.3 Analysis of variance

The analysis is helped by defining

$$S_T = \sum \sum \sum (x_{ijk} - \bar{x}_{...})^2 = \sum \sum \sum x_{ijk}^2 - C,$$

$$S_t = S_t(\mathbf{0}) = JK \sum T_i^2 - C,$$

$$S_b = S_b(\mathbf{0}) = IK \sum B_j^2 - C,$$

$$S_{tb} = S_{tb}(\mathbf{0}) = K \sum C_{ij}^2 - C - S_t - S_b$$

in which  $T_i$  is the treatment total  $JKx_{i..}$ , while  $B_j$  is the block total  $IKx_{.j..}$ , and  $C_{ij}$  is the cell total  $Kx_{ij..}$ . Finally,  $G$  is the grand total  $IJKx_{...}$ , and  $C$  is the correction factor  $G^2/N$ . The resulting ANOVA table is as follows:

**ANOVA Table**

Source	Sum of squares	Degrees of freedom	Mean square	Ratio
Treatments	$S_t$	$I - 1$	$S_t/(I - 1)$	$F_t(\mathbf{0})$
Blocks	$S_b$	$J - 1$	$S_b/(J - 1)$	$F_b(\mathbf{0})$
Interaction	$S_{tb}$	$(I - 1)(J - 1)$	$S_{tb}/(I - 1)$	$F_{tb}(\mathbf{0})$
Error	$S_e$	$v = IJ(K - 1)$	$s^2 = S_e/v$	
TOTAL	$S_T$	$N - 1$		

Other matters, such as the exploration of treatments or blocks contrasts, can be pursued as in the case of the one way model.

## 6.7 The general linear model

### 6.7.1 Formulation of the general linear model

All of the last few sections have been concerned with particular cases of the so-called general linear model. It is possible to treat them all at once in an approach using matrix theory. In most of this book, substantial use of matrix theory has been avoided, but if the reader has some knowledge of matrices this section may be helpful, in that the intention here is to put some of the models already considered into the form of the general linear model. An understanding of how these models can be put into such a framework, will put the reader in a good position to approach the theory in its full generality, as it is dealt with in such works as Box and Tiao (1992).

It is important to distinguish row vectors from column vectors. We write  $\mathbf{x}$  for

a column vector and  $x^T$  for its transpose; similarly if  $A$  is an  $n \times r$  matrix then  $A^T$  is its  $r \times n$  transpose. Consider a situation in which we have a column vector  $x$  of observations, so that  $x = (x_1, x_2, \dots, x_n)^T$  (the equation is written thus to save the excessive space taken up by column vectors). We suppose that the  $x_i$  are independently normally distributed with common variance  $\phi$  and a vector  $\mathbf{E}x$  of means satisfying  $\mathbf{E}x = A\theta$ ,

where  $\theta = (\theta_1, \theta_2, \dots, \theta_r)^T$  is a vector of unknown parameters and  $A$  is a known  $n \times r$  matrix.

In the case of the original formulation of the bivariate linear regression model in which, conditional on  $x_i$ , we have  $y_i \sim N(\eta_0 + \eta_1 x_i, \phi)$  then  $y$  takes the part of  $x$ ,  $r = 2$ ,  $\eta = (\eta_0, \eta_1)^T$  takes the part of  $\theta$  and  $A_0$  takes the part of  $A$  where

$$A_0 = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \quad A_0\eta = \begin{pmatrix} \eta_0 + \eta_1 x_1 \\ \eta_0 + \eta_1 x_2 \\ \vdots \\ \eta_0 + \eta_1 x_n \end{pmatrix}.$$

This model is reformulated in terms of  $\eta = (\alpha, \beta)^T$  and

$$A = \begin{pmatrix} 1 & x_1 - \bar{x} \\ 1 & x_2 - \bar{x} \\ \vdots & \vdots \\ 1 & x_n - \bar{x} \end{pmatrix} \quad A_0\eta = \begin{pmatrix} \alpha + \beta(x_1 - \bar{x}) \\ \alpha + \beta(x_2 - \bar{x}) \\ \vdots \\ \alpha + \beta(x_n - \bar{x}) \end{pmatrix}.$$

In the case of the one way model (where, for simplicity, we shall restrict ourselves to the case where  $K_i = K$  for all  $i$ ),  $n = N, r = I, \theta = (\theta_1, \theta_2, \dots, \theta_I)^T$  and

$$x = \begin{pmatrix} x_{11} \\ x_{12} \\ \vdots \\ x_{1K} \\ x_{21} \\ x_{22} \\ \vdots \\ x_{IK} \end{pmatrix} \quad A = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 1 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix} \quad A\theta = \begin{pmatrix} \theta_1 \\ \theta_1 \\ \vdots \\ \theta_1 \\ \theta_2 \\ \theta_2 \\ \vdots \\ \theta_I \end{pmatrix}.$$

The two way layout can be expressed similarly using a matrix of 0s and 1s. It is also possible to write the multiple regression model  $y_i \sim N(\eta_0 + x_{i1}\eta_1 + \dots + x_{ir}\eta_r, \phi)$  (the  $x_{ij}$  being treated as known) as a case of the general linear model.

## 6.7.2 Derivation of the posterior

Noting that for any vector  $u$  we have  $\sum u_i^2 = u^T u$ , we can write the likelihood function for the general linear model in the form

$$(2\pi\phi)^{-n/2} \exp\{-\frac{1}{2}(x - A\theta)^T(x - A\theta)/\phi\}.$$

Taking standard reference priors, that is,  $p(\theta, \phi) \propto 1/\phi$  the posterior is  $p(\theta, \phi|x) \propto \phi^{-n/2-1} \exp[-\frac{1}{2}(x - A\theta)^T(x - A\theta)/\phi]$ .

Now as  $(AB)^T = B^T A^T$  and scalars equal their own transposes

$$(x - A\theta)^T(x - A\theta) = \theta^T A^T A \theta - 2\theta^T A^T x + x^T x,$$

so that if  $\hat{\theta}$  is such that

$$A^T A \hat{\theta} = A^T x$$

(so that  $\theta^T A^T A \hat{\theta} = \theta^T A^T x$ ), that is, assuming  $A^T A$  is non-singular,  $\hat{\theta} = (A^T A)^{-1} A^T x$ ,

we have

$$\begin{aligned} S &= (x - A\theta)^T(x - A\theta) = \theta^T A^T A \theta - 2\theta^T A^T x + x^T x \\ &= (\theta - \hat{\theta})^T A^T A (\theta - \hat{\theta}) + x^T x - \hat{\theta}^T A^T A \hat{\theta} \\ &= S_t(\theta) + S_e, \end{aligned}$$

where

$$S_t(\theta) = (\theta - \hat{\theta})^T A^T A (\theta - \hat{\theta}), \quad S_e = x^T x - \hat{\theta}^T A^T A \hat{\theta}.$$

It is also useful to define

$$v = n - r, \quad s^2 = S_e/v.$$

Because  $S_t(\theta)$  is of the form  $u^T u$ , it is always non-negative, and it clearly vanishes if  $\theta = \hat{\theta}$ . Further,  $S_e$  is the minimum value of the sum of squares  $S$  and so is positive. It is sometimes worth noting that  $S_e = (x - A\hat{\theta})^T(x - A\hat{\theta})$  as is easily shown.

It follows that the posterior can be written as

$$p(\theta, \phi|x) \propto \phi^{-n/2-1} \exp[-\frac{1}{2}\{S_t(\theta) + S_e\}/\phi].$$

In fact, this means that for given  $\phi$ , the vector  $\theta$  has a multivariate normal distribution of mean  $\hat{\theta}$  and variance-covariance matrix  $A^T A$ .

If you are now interested in  $\theta$  as a whole you can now integrate with respect to

$$p(\theta|x) \propto \{S_t(\theta) + S_e\}^{-n/2}$$

$\phi$  to get  $\propto \{1 + r F(\theta)/v\}^{-n/2}$ ,

where

$$F(\theta) = \frac{S_t(\theta)/r}{S_e/v} = \frac{S_t(\theta)/r}{s^2}.$$

It may also be noted that the set

$$E(F) = \{\theta; F(\theta) \leq F\}$$

is a hyperellipsoid in  $r$ -dimensional space in which the length of each of the axes is in a constant ratio to  $\sqrt{F}$ . The argument of Section 6.5 on the one way layout can now be adapted to show that  $F(\theta) \sim F_{r,v}$ , so that  $E(F)$  is an HDR for  $\theta$  of probability  $p$  if  $F$  is the appropriate percentage point of  $F_{r,v}$ .

### 6.7.3 Inference for a subset of the parameters

However, it is often the case that most of the interest centres on a subset of the parameters, say on  $\theta_1, \theta_2, \dots, \theta_k$ . If so, then it is convenient to write  $\theta' = (\theta_1, \theta_2, \dots, \theta_k)$  and  $\theta'' = (\theta_{k+1}, \dots, \theta_r)$ . If it happens that  $S_t(\theta)$  splits into a sum  $S_t(\theta) = S'_t(\theta') + S''_t(\theta'')$

then it is easy to integrate

$$p(\theta', \theta'', \phi | x) \propto \phi^{-n/2-1} \exp[-\frac{1}{2}\{S'_t(\theta') + S''_t(\theta'') + S_e\}/\phi]$$

to get

$$p(\theta', \phi | x) \propto \phi^{-(n-r+k)/2-1} \exp[-\frac{1}{2}\{S'_t(\theta') + S_e\}/\phi],$$

and thus as  $n - r = \nu$

$$\begin{aligned} p(\theta' | x) &\propto \{S'_t(\theta') + S_e\}^{-(j+\nu)/2} \\ &\propto \{1 + kF'_t(\theta')/\nu\}^{-(k+\nu)/2}, \end{aligned}$$

where

$$F'_t(\theta') = \frac{S'_t(\theta')/j}{S_e/\nu}.$$

It is now easy to show that  $F'_t(\theta') \sim F_{k,\nu}$  and hence to make inferences for  $\theta'$ .

Unfortunately, the quadratic form  $S_t(\theta)$  does in general contain terms  $\theta_i \theta_j$ , where  $i \leq k$  but  $j > k$  and hence it does not in general split into  $S'_t(\theta) + S''_t(\theta'')$ . We will not discuss such cases further; useful references are Box and Tiao (1992), Lindley and Smith (1972) and Seber (2003).

### 6.7.4 Application to bivariate linear regression

The theory can be illustrated by considering the simple linear regression model. Consider first the reformulated version in terms of  $A$  and  $\beta = (\alpha, \beta)^T$ . In this

$$\text{case } A^T A = \begin{pmatrix} n & 0 \\ 0 & S_{xx} \end{pmatrix},$$

and the fact that this matrix is easy to invert is one of the underlying reasons why this reformulation was sensible. Also

$$\hat{\theta} = (A^T A)^{-1} A^T y = \begin{pmatrix} \bar{y} \\ S_{xy}/S_{xx} \end{pmatrix}$$

$$S_t(\theta) = (\theta - \hat{\theta})^T A^T A (\theta - \hat{\theta}) = n(\alpha - \bar{y})^2 + S_{xx}(\beta - S_{xy}/S_{xx})^2,$$

$$S_e = y^T y - \hat{\theta}^T A^T A \hat{\theta} = \sum y_i^2 - n\bar{y}^2 - S_{xx}(S_{xy}/S_{xx})^2 = S_{yy}^2 - S_{xy}^2/S_{xx},$$

so that  $S_e$  is what was denoted  $S_{ee}$  in Section 6.3 on bivariate linear regression.

If you are particularly interested in  $\alpha$ , then in this case the thing to do is to note

that the quadratic form splits with  $S_t'(\theta) = n(\alpha - \bar{y})^2$  and  $S_t''(\theta'') = S_{xx}(\beta - S_{xy}/S_{xx})^2$ . Consequently, the posterior distribution of  $\phi$  is given by  $\frac{v(\alpha - \bar{y})^2}{s^2} = \frac{S_t'(\theta')/1}{S_e/(n-2)} \sim F_{1,n-2}$ .

Since the square of a  $t_{n-2}$  variable can easily be shown to have an  $F_{1,n-2}$  distribution, this conclusion is equivalent to that of Section 6.3.

The greater difficulties that arise when  $A^T A$  is non-diagonal can be seen by following the same process through for the original formulation of the bivariate linear regression model in terms of  $A_0$  and  $\eta = (\eta_0, \eta_1)^T$ . In this case it is easy enough to find the posterior distribution of  $\eta_1$ , but it involves some rearrangement to get that of  $\eta_0$ .

## 6.8 Exercises on Chapter 6

1. The sample correlation coefficient between length and weight of a species of frog was determined at each of a number of sites. The results were as follows:

Site	1	2	3	4	5
Number of frogs	12	45	23	19	30
Correlation	0.631	0.712	0.445	0.696	0.535

Find an interval in which you are 95% sure that the correlation coefficient lies.

2. Three groups of children were given two tests. The numbers of children and the sample correlation coefficients between the two test scores in each group were as follows:

Number of children	45	34	49
Correlation	0.489	0.545	0.601

Is there any evidence that the association between the two tests differs in the three groups?

3. Suppose you have sample correlation coefficients  $r_1, r_2, \dots, r_k$  on the basis of sample sizes  $n_1, n_2, \dots, n_k$ . Give a 95% posterior confidence interval for  $\zeta = \tanh^{-1} \rho$ .

4. From the approximation

$$p(\rho|x, y) \propto (1 - \rho^2)^{n/2} (1 - \rho r)^{-n}$$

which holds for large  $n$ , deduce an expression for the log-likelihood  $L(\rho|x, y)$  and hence show that the maximum likelihood occurs when  $\rho = r$ . An approximation to the information can now be made by replacing  $r$  by  $\rho$  in the second derivative of the likelihood, since  $\rho$  is near  $r$  with high probability. Show that this approximation suggests a prior density of the form

$$p(\rho) \propto (1 - \rho^2)^{-1}.$$

5. Use the fact that

$$\int_0^\infty (\cosh t + \cos \theta)^{-1} dt = \theta / \sin \theta$$

(cf. Edwards, 1921, art. 180) to show that

$$p(\rho|x, y) \propto p(\rho)(1 - \rho^2)^{(n-1)/2} \frac{d^{n-2}}{d(\rho r)^{n-2}} \left( \frac{\arccos(-\rho r)}{\sqrt{1 - \rho^2 r^2}} \right).$$

6. Show that in the special case where the sample correlation coefficient  $r = 0$  and the prior takes the special form  $p(\rho) \propto (1 - \rho^2)^{k/2}$  the variable

$$\sqrt{(k+n+1)}\rho/(1-\rho^2)$$

has a Student's t distribution on  $k+n+1$  degrees of freedom.

**7.** By writing

$$\begin{aligned}\omega^{-1}(\omega + \omega^{-1} - 2\rho r)^{-(n-1)} &= \omega^{n-2}(1 - \rho^2)^{-(n-1)} \\ &\times [1 + (\omega - \rho r)^2(1 - \rho^2 r^2)^{-1}]^{-(n-1)}\end{aligned}$$

and using repeated integration by parts, show that the posterior distribution of  $\rho$  can be expressed as a finite series involving powers of  $\sqrt{(1 - \rho r)/(1 + \rho r)}$

and Student's t integrals.

**8.** By substituting

$$\cosh t - \rho r = \frac{1 - \rho r}{1 - u}$$

in the form

$$p(\rho|x, y) \propto p(\rho)(1 - \rho^2)^{(n-1)/2} \int_0^\infty (\cosh t - \rho r)^{-(n-1)} dt$$

for the posterior density of the correlation coefficient and then expanding

$$[1 - \frac{1}{2}(1 + \rho r)u]^{-\frac{1}{2}}$$

as a power series in  $u$ , show that the integral can be expressed as a series of beta functions. Hence, deduce that

$$p(\rho|x, y) \propto p(\rho)(1 - \rho^2)^{(n-1)/2}(1 - \rho r)^{-n+(3/2)} S_n(\rho r),$$

where

$$S_n(\rho r) = 1 + \sum_{l=1}^{\infty} \frac{1}{l!} \left(\frac{1 + \rho r}{8}\right)^l \prod_{s=1}^l \frac{(2s-1)^2}{(n - \frac{3}{2} + s)}.$$

**9.** Fill in the details of the derivation of the prior

$$p(\phi, \psi, \rho) \propto (\phi\psi)^{-1}(1 - \rho^2)^{-3/2}$$

from Jeffreys' rule as outlined at the end of Section 6.1.

**10.** The following data consist of the estimated gestational ages (in weeks) and weights (in grammes) of 12 female babies:

Age	40	36	40	38	42	39	40	37	36	38	39	40
Weight	3317	2729	2935	2754	3210	2817	3126	2539	2412	2991	2875	3231

Give an interval in which you are 90% sure that the gestational age of a particular such baby will lie if its weight is 3000 g, and give a similar interval in which the mean weight of all such babies lies.

**11.** Show directly from the definitions that, in the notation of Section 6.3,

$$S_{ee} = \sum (y_i - a - b(x_i - \bar{x}))^2.$$

**12.** Observations  $y_i$  for  $i = -m, -m+1, \dots, m$  are available which satisfy the regression model

$y_i \sim N(\alpha + \beta u_i + \gamma v_i, \phi)$ ,

where  $u_i = i^2 - \frac{1}{2}m(m+1)$ . Adopting the reference prior  $p(\alpha, \beta, \gamma, \phi) \propto 1/\phi$ , show that the posterior distribution of  $\alpha$  is such that

$$\frac{\alpha - \bar{y}}{s/\sqrt{n}} \sim t_{n-3},$$

where  $n=2m+1$ ,  $s^2=S_{ee}/(n-3)$  and

$$S_{ee} = S_{yy} - S_{uy}^2/S_{uu} - S_{vy}^2/S_{vv}$$

in which  $S_{yy}$ ,  $S_{uy}$ , etc., are defined by

$$S_{yy} = \sum (y_i - \bar{y})^2 \quad \text{and} \quad S_{uy} = \sum (u_i - \bar{u})(y_i - \bar{y}).$$

[Hint: Note that  $\sum u_i = \sum v_i = \sum u_i v_i = 0$ , and hence  $\bar{u} = \bar{v} = 0$  and  $S_{uy}=0$ .]

**13.** Fisher (1925b, Section 41) quotes an experiment on the accuracy of counting soil bacteria. In it, a soil sample was divided into four parallel samples, and from each of these after dilution seven plates were inoculated. The number of colonies on each plate is shown below. Do the results from the four samples agree within the limits of random sampling?

Plate / Sample	A	B	C	D
1	72	74	78	69
2	69	72	74	67
3	63	70	70	66
4	59	69	58	64
5	59	66	58	64
6	53	58	56	58
7	51	52	56	54

**14.** In the case of the data on scab disease quoted in Section 6.5, find a contrast measuring the effect of the season in which sulphur is applied and give an appropriate HDR for this contrast.

**15.** The data below [from Wishart and Sanders (1955, Table 5.6)] represent the weight of green produce in pounds made on an old pasture. There were three main treatments, including a control (O) consisting of the untreated land. In the other cases, the effect of a grass-land rejuvenator (R) was compared with the use of the harrow (H). The blocks were, therefore, composed of three plots each, and the experiment consisted of six randomized blocks placed side by side. The plan and yields were as follows:

O. H. R.		R. H. O.		O. R. H.		O. R. H.		H. O. R.		O. H. R.												
813	647	713		814	759	795		705	652	598		774	617	559		580	687	539		581	480	537

Derive an appropriate two-way analysis of variance.

**16.** Express the two-way layout as a particular case of the general linear model.

**17.** Show that the matrix  $A^+ = (A^T A)^{-1} A^T$  which arises in the theory of the general linear model is a *generalized inverse* of the (usually non-square) matrix  $A$  in that **a.**  $AA^+A = A$

- b.**  $A^+AA^+ = A^+$
- c.**  $(AA^+)^T = AA^+$
- d.**  $(A^+A)^T = A^+A.$

**18.** Express the bivariate linear regression model in terms of the original parameters  $\eta = (\eta_0, \eta_1)^T$  and the matrix  $A_0$  and use the general linear model to find the posterior distribution of  $\eta$ .

# Other topics

## 7.1 The likelihood principle

### 7.1.1 Introduction

This section would logically come much earlier in the book than it is placed, but it is important to have some examples of Bayesian procedures firmly in place before considering this material. The basic result is due to Birnbaum (1962), and a more detailed consideration of these issues can be found in Berger and Wolpert (1988).

The nub of the argument here is that in drawing any conclusion from an experiment only the actual observation  $x$  made (and not the other possible outcomes that might have occurred) is relevant. This is in contrast to methods by which, for example, a null hypothesis is rejected because the probability of a value as large as *or larger than* that actually observed is small, an approach which leads to Jeffreys' criticism that was mentioned in Section 4.1 when we first considered hypothesis tests, namely, that 'a hypothesis that may be true may be rejected because it has not predicted observable results that have not occurred'. Virtually all of the ideas discussed in this book abide by this principle, which is known as the *likelihood principle* (there are some exceptions, for example Jeffreys' rule is not in accordance with it). We shall show that it follows from two other principles, called the conditionality principle and the sufficiency principle, both of which are hard to argue against.

In this section, we shall write  $x$  for a particular piece of data, not necessarily one-dimensional, the density  $p(x|\theta)$  of which depends on an unknown parameter  $\theta$ . For simplicity, we shall suppose that  $x$  and  $\theta$  are discrete (although they may be more than one-dimensional). The triple  $E = \{\tilde{x}, \theta, p(x|\theta)\}$  represents the essential features of an experiment to gain information about  $\theta$ , and accordingly we shall refer to  $E$  as an experiment. Note that the random variable  $\tilde{x}$  is a feature of the experiment, not the particular value  $x$  that may be observed when the

experiment is carried out on a particular occasion. If such an experiment is carried out and the value  $x$  is observed, then we shall write  $Ev\{E, x, \theta\}$  for the evidence provided about the value of  $\theta$  by carrying out experiment  $E$  and observing the value  $x$ . This ‘evidence’ is not presumed to be in any particular form. To a Bayesian, it would normally be the posterior distribution of  $\theta$  or some feature of it, but for the moment we are not restricting ourselves to Bayesian inference, and a classical statistician might consider evidence to be made up of significance levels and confidence intervals, etc., while the notation does not rule out some form of evidence that would be new to both.

For example, you might be interested in the proportion  $\theta$  of defective articles coming from a factory. A possible experiment  $E$  would consist in observing a fixed number  $n$  of articles chosen at random and observing the number  $x$  defective, so that  $p(x|\theta)$  is a family of binomial densities. To have a definite experiment, it is necessary to give  $n$  a specific value, for example,  $n = 100$ ; once  $n$  is known  $E$  is fully determined. If we then observe that  $x = 3$ , then  $Ev\{E, 3, \theta\}$  denotes the conclusions we arrive at about the value of  $\theta$ .

### 7.1.2 The conditionality principle

The conditionality principle can be explained as the assertion that if you have decided which of two experiments you performed by tossing a coin, then once you tell me the *end result* of the experiment, it will not make any difference to any inferences I make about an unknown parameter  $\theta$  whether or not I know which way the coin landed and hence which experiment was actually performed (assuming that the probability of the coin’s landing ‘heads’ does not in any way depend on  $\theta$ ). For example, if we are told that an analyst has reported on the chemical composition of a sample, then it is irrelevant whether we had always intended to ask him or her to analyze the sample or had tossed a coin to decided whether to ask that scientist or the one in the laboratory next door to analyze it. Put this way, the principle should seem plausible, and we shall now try to formalize it.

We first need to define a *mixed* experiment. Suppose that there are two experiments,  $E_1 = \{\tilde{y}, \theta, p(y|\theta)\}$  and  $E_2 = \{\tilde{z}, \theta, p(z|\theta)\}$  and that the random variable  $\tilde{k}$  is such that  $p(k=1) = p(k=2) = \frac{1}{2}$ , whatever  $\theta$  is and independently of  $y$  and  $z$ . Then the mixed experiment  $E^*$  consists of carrying out  $E_1$  if  $k=1$  and  $E_2$  if  $k=2$ . It can also be defined as the triple  $\{\tilde{x}, \theta, p(x|\theta)\}$ , where

$$x = \begin{cases} (1, y) & \text{if } k = 1 \\ (2, z) & \text{if } k = 2 \end{cases}$$

and

$$p(x|\theta) = \begin{cases} \frac{1}{2}p(y|\theta) & \text{if } k = 1, \text{ so } x = (1, y) \\ \frac{1}{2}p(z|\theta) & \text{if } k = 2, \text{ so } x = (2, z). \end{cases}$$

We only need to assume the following rather weak form of the principle:

*Weak conditionality principle.* If  $E_1, E_2$  and  $E^*$  are as defined earlier, then

$$\text{Ev}\{E^*, x, \theta\} = \begin{cases} \text{Ev}\{E_1, y, \theta\} & \text{if } k = 1, \text{ so } x = (1, y) \\ \text{Ev}\{E_2, z, \theta\} & \text{if } k = 2, \text{ so } x = (2, z), \end{cases}$$

that is, the evidence about  $\theta$  from  $E^*$  is just the evidence from the experiment actually performed.

### 7.1.3 The sufficiency principle

The sufficiency principle says that if  $t(x)$  is sufficient for  $\theta$  given  $x$ , then any inference we may make about  $\theta$  may be based on the value of  $t$ , and once we know that we have no need of the value of  $x$ . We have already seen in Section 2.9 that Bayesian inference satisfies the sufficiency principle. The form in which the sufficiency principle will be used in this section is as follows:

#### 7.1.3.1 Weak sufficiency principle

Consider the experiment  $E = \{\tilde{x}, \theta, p(x|\theta)\}$  and suppose that  $t=t(x)$  is sufficient for  $\theta$  given  $x$ . Then if  $t(x_1)=t(x_2)$

$$\text{Ev}\{E, x_1, \theta\} = \text{Ev}\{E, x_2, \theta\}.$$

This clearly implies that, as stated in Corollary 2.1 in Section 2.9.3, ‘For any prior distribution, the posterior distribution of  $\theta$  given  $x$  is the same as the posterior distribution of  $\theta$  given a sufficient statistic  $t$ ’. In Bayesian statistics inference is based on the posterior distribution, but this principle makes it clear that even if we had some other method of arriving at conclusions,  $x_1$  and  $x_2$  would still lead to the same conclusions.

### 7.1.4 The likelihood principle

For the moment, we will state what the likelihood principle is – its implications will be explored later.

#### 7.1.4.1 Likelihood principle

Consider two different experiments  $E_1 = \{\tilde{y}, \theta, p(y|\theta)\}$  and  $E_2 = \{\tilde{z}, \theta, p(z|\theta)\}$ , where  $\theta$  is the same quantity in each experiment. Suppose that there are particular possible outcomes  $y^*$  of experiment  $E_1$  and  $z^*$  of  $E_2$  such that

$$p(y^*|\theta) = cp(z^*|\theta)$$

for some constant  $c$ , that is, the likelihoods of  $\theta$  as given by these possible outcomes of the two experiments are proportional, so that

$$l(\theta|y^*) \propto l(\theta|z^*).$$

Then

$$\text{Ev}\{E_1, y^*, \theta\} = \text{Ev}\{E_2, z^*, \theta\}.$$

The following theorem [due to Birnbaum (1962)] shows that the likelihood principle follows from the other two principles described earlier.

**Theorem 7.1** The likelihood principle follows from the weak conditionality principle and the weak sufficiency principle.

*Proof.* If  $E_1$  and  $E_2$  are the two experiments about  $\theta$  figuring in the statement of the likelihood principle, consider the mixed experiment  $E^*$  which arose in connection with the weak conditionality principle. Define a statistic  $t$  by

$$t = t(x) = \begin{cases} (1, y^*) & \text{if } k = 2 \text{ and } z = z^* \\ x & \text{otherwise.} \end{cases}$$

(Note that if experiment 2 is performed and we observe the value  $z^*$  then by the assumption of the likelihood principle there is a value  $y^*$  such that  $p(y^*|\theta) = cp(z^*|\theta)$  so we can take this value of  $y^*$  in the proof.) Now note that if  $t \neq (1, y^*)$  then

$$p(x|t, \theta) = \begin{cases} 1 & \text{if } t = t(x) \\ 0 & \text{otherwise} \end{cases}$$

whereas if  $t = x = (1, y^*)$  then

$$p(x|t, \theta) = \frac{\frac{1}{2}p(y^*|\theta)}{\frac{1}{2}p(y^*|\theta) + \frac{1}{2}p(z^*|\theta)} = \frac{c}{1+c}$$

and if  $t = (1, y^*)$  but  $x = (2, z^*)$  then

$$p(x|t, \theta) = \frac{\frac{1}{2}p(z^*|\theta)}{\frac{1}{2}p(y^*|\theta) + \frac{1}{2}p(z^*|\theta)} = \frac{1}{1+c}$$

while for  $t = (1, y^*)$  and all other  $x$  we have  $p(x|t, \theta) = 0$ . In no case does  $p(x|t, \theta)$  depend on  $\theta$  and hence, from the definition given when sufficiency was first introduced in Section 2.9,  $t$  is sufficient for  $\theta$  given  $x$ . It follows from the weak sufficiency principle that  $\text{Ev}\{E^*, (1, y^*), \theta\} = \text{Ev}\{E^*, (2, z^*), \theta\}$ . But the weak conditionality principle now ensures that

$$\begin{aligned} \text{Ev}\{E_1, y^*, \theta\} &= \text{Ev}\{E^*, (1, y^*), \theta\} = \text{Ev}\{E^*, (2, z^*), \theta\} \\ &= \text{Ev}\{E_2, z^*, \theta\} \end{aligned}$$

establishing the likelihood principle. ■

**Corollary 7.1** If  $E = \{\tilde{x}, \theta, p(x|\theta)\}$  is an experiment, then  $Ev\{E, x, \theta\}$  should depend on  $E$  and  $x$  only through the likelihood

$$l(\theta|x) \propto p(x|\theta).$$

*Proof.* For any one particular value  $x_1$  of  $x$  define

$$y = \begin{cases} 1 & \text{if } x = x_1 \\ 0 & \text{otherwise,} \end{cases}$$

so that  $P(y = 1|\theta) = p(x_1|\theta)$  (since we have assumed for simplicity that everything is discrete this will not, in general, be zero). Now let the experiment  $E_1$  consist simply of observing  $y$ , that is, of noting whether or not  $x=x_1$ . Then the likelihood principle ensures that  $Ev\{E, x_1, \theta\} = Ev\{E_1, 1, \theta\}$ , and  $E_1$  depends solely on  $p(x_1|\theta)$  and hence solely on the likelihood of the observation actually made. ■

**Converse 7.1** If the likelihood principle holds, then so do the weak conditionality principle and the weak sufficiency principle.

*Proof.* Using the notation introduced earlier for the mixed experiment, we see that if  $x=(1, y)$  then

$$p(x|\theta) = \frac{1}{2}p(y|\theta)$$

and so by the likelihood principle  $Ev\{E^*, x, \theta\} = Ev\{E_1, y, \theta\}$ , implying the weak conditionality principle. Moreover, if  $t$  is a sufficient statistic and  $t(x_1)=t(x_2)$ , then  $x_1$  and  $x_2$  have proportional likelihood functions, so that the likelihood principle implies the weak sufficiency principle. ■

### 7.1.5 Discussion

From the formulation of Bayesian inference as ‘posterior is proportional to prior times likelihood,’ it should be clear that Bayesian inference obeys the likelihood principle. It is not logically necessary that if you find the arguments for the likelihood principle convincing, you have to accept Bayesian inference, and there are some authors, for example, Edwards (1992), who have argued for a non-Bayesian form of inference based on the likelihood. Nevertheless, I think that Savage was right in saying in the discussion on Birnbaum (1962) that ‘... I suspect that that once the likelihood principle is widely recognized, people will not long stop at that halfway house but will go forward and accept the implications of personalistic probability for statistics’.

Conversely, much of classical statistics notably fails to obey the likelihood principle – any use of tail areas (e.g. the probability of observing a value as large as that seen *or greater*) evidently involves matters other than the likelihood of

the observations actually made. Another quotation from Savage, this time from Savage *et al.* (1962), may help to point to some of the difficulties that arise in connection with confidence intervals.

Imagine, for example, that two Meccans carefully drawn at random differ from each other in height by only 0.01 mm. Would you offer 19 to 1 odds that the standard deviation of the height of Meccans is less than 1.13 mm? That is the 95 per cent upper confidence limit computed with one degree of freedom. No, I think you would not have enough confidence in that limit to offer odds of 1 to 1.

In fact, the likelihood principle has serious consequences for both classical and Bayesian statisticians and some of these consequences will be discussed in the Sections 7.2–7.4. For classical statisticians, one of the most serious is the *stopping rule principle*, while for Bayesians one of the most serious is that Jeffreys' rule for finding reference priors is incompatible with the likelihood principle.

## 7.2 The stopping rule principle

### 7.2.1 Definitions

We shall restrict ourselves to a simple situation, but it is possible to generalize the following account considerably; see Berger and Wolpert (1988, Section 4.2). Basically, in this section, we will consider a sequence of experiments which can be terminated at any stage in accordance with a rule devised by the experimenter (or forced upon him).

Suppose that the observations  $x_1, x_2, \dots$  are independently and identically distributed with density  $p(x|\theta)$  and let

$$\begin{aligned} x^{(m)} &= (x_1, x_2, \dots, x_m) \\ \bar{x}_m &= (x_1 + x_2 + \dots + x_m)/m. \end{aligned}$$

We say that  $s$  is a *stopping rule* or a *stopping time* if it is a random variable whose values are finite natural numbers  $(1, 2, 3, \dots)$  with probability one, and is such that whether or not  $s > m$  depends solely on  $x^{(m)}$ . In a sequential experiment  $E$  we observe the values  $x_1, x_2, \dots, x_s$  where  $s$  is such a stopping rule and then stop. The restriction on the distribution of  $s$  means simply that whether or not you decide to stop cannot depend on future observations (unless you are clairvoyant), but only on the ones you have available to date.

## 7.2.2 Examples

**1. Fixed sample size from a sequence of Bernoulli trials.** Suppose that the  $x_i$  are independently 1 with probability  $\pi$  (representing ‘success’) or 0 with probability  $1 - \pi$  (representing ‘failure’). If  $s=n$  where  $n$  is a constant, then we have the usual situation which gives rise to the binomial distribution for the total number of successes.

**2. Stopping after the first success in Bernoulli trials.** With the  $x_i$  as in the previous example, we could stop after the first success, so that  
$$s = \min\{m; x_m = 1\}.$$

Because the probability that  $s > n$  is  $(1 - \pi)^n$  which tends to 0 as  $n \rightarrow \infty$ , this is finite with probability 1.

**3. A compromise between the first two examples.** With the  $x_i$  as in the previous two examples, we could stop after the first success if that occurs at or before the  $n$ th trial, but if there has not been a success by then, stop at the  $n$ th trial, so that

$$s = \min\{n, \min\{m; x_m = 1\}\}.$$

**4. Fixed size sample from the normal distribution.** If the  $x_i$  are independently  $N(0, 1)$  and  $s=n$  where  $n$  is a constant, then we have a case which has arisen often before of a sample of fixed size from a normal distribution.

**5. Stopping when a fixed number of standard deviations from the mean.** Still taking  $x_i \sim N(0, 1)$ , we could have

$$s = \min\{m; |\bar{x}_m| > c/\sqrt{m}\}$$

which, as  $\bar{x}_m \sim N(0, 1/m)$ , means stopping as soon as we observe a value of  $\bar{x}_m$  that is at least  $c$  standard deviations from the mean. It is not obvious in this case that  $s$  is finite with probability 1, but it follows from the law of the iterated logarithm, a proof of which can be found in any standard text on probability.

## 7.2.3 The stopping rule principle

The *stopping rule principle* is that in a sequential experiment, if the observed value of the stopping rule is  $m$ , then the evidence  $Ev\{E, x^{(m)}, \theta\}$  provided by the experiment about the value of  $\theta$  should not depend on the stopping rule.

Before deciding whether it is valid, we must consider what it means. It asserts, for example, that if you observe ten Bernoulli trials, nine of which result in failure and only the last in success, then any inference about the probability  $\pi$  of

successes cannot depend on whether the experimenter had all along intended to carry out ten trials and had, in fact, observed one success, or whether he or she had intended to stop the experiment immediately the first success was observed. Thus, it amounts to an assertion that all that matters is what actually happened and not the intentions of the experimenter if something else had happened.

**Theorem 7.2** The stopping rule principle follows from the likelihood principle, and hence is a logical consequence of the Bayesian approach.

*Proof.* If the  $x_i$  are discrete random variables, then it suffices to note that the likelihood

$$l(\theta|x^{(s)}) \propto p(x_1|\theta) p(x_2|\theta) \dots p(x_s|\theta)$$

which clearly does not depend on the stopping rule. There are some slight complications in the continuous case, which are largely to do with measure theoretic complications, and in particular with events of probability zero, but a general proof from the so-called relative likelihood principle is more or less convincing; for details, see Berger and Wolpert (1988, Sections 3.4.3 and 4.2.6). ■

## 7.2.4 Discussion

The point about this is as follows. A classical statistician is supposed to choose the stopping rule before the experiment and then follow it exactly. In actual practice, the ideal is often not adhered to; an experiment can end because the data already looks good enough, or because there is no more time or money, and yet the experiment is often analyzed as if it had a fixed sample size. Although stopping for some reasons would be harmless, statisticians who stop ‘when the data looks good’, a process which is sometimes described as *optional* (or *optimal*) stopping, can produce serious errors if used in a classical analysis.

It is often argued that a single number which is a good representation of our knowledge of a parameter should be unbiased, that is, should be such that its expectation over repeated sampling should be equal to that parameter. Thus, if we have a sample of fixed size from a Bernoulli distribution [example (1), mentioned before], then  $E\bar{x}_n = \pi$ , so that  $\bar{x}_n$  is in that sense a good estimator of  $\pi$ . However, if the stopping rule in example (2) or that in example (3), is used, then the proportion  $\bar{x}_s$  will, on average, be more than  $\pi$ . If, for example, we take example (3) with  $n = 2$ , then

$$\begin{aligned} E\bar{x}_s &= P(x_1 = 1) E(\bar{x}_1 | x_1 = 1, \pi) + P(x_1 = 0|\pi) E(\bar{x}_2 | x_1 = 0, \pi) \\ &= \pi \times 1 + (1 - \pi) \times \left\{ \frac{1}{2} \times \pi + 0 \times (1 - \pi) \right\} \\ &= \pi + \frac{1}{2}\pi(1 - \pi). \end{aligned}$$

Thus, if a classical statistician who used the proportion of successes actually observed as an estimator of the probability  $\pi$  of success, would be accused of ‘making the probability of success look larger than it is’.

The stopping rule principle also plays havoc with classical significance tests. A particular case can be constructed from example (5) above with, for example,  $c = 2$ . If a classical statistician were to consider data from an  $N(\theta, 1)$  population in which (unknown to him or her)  $\theta = 0$ , then because  $s$  is so constructed that, necessarily, the value of  $|\bar{x}_s|$  is at least  $c$  standard deviations from the mean, a single sample of a fixed size equal to this would necessarily lead to a rejection of the null hypothesis that  $\theta = 0$  at the 5% level. By taking other values of  $c$ , it can be seen that a crafty classical statistician could arrange to reject a null hypothesis that was, in fact, true, at any desired significance level.

It can thus be seen that the stopping rule principle is very hard to accept from the point of view of classical statistics. It is for these reasons that Savage said that

I learned the stopping rule principle from Professor Barnard, in conversation in the summer of 1952. Frankly, I then thought it a scandal that anyone in the profession could advance an idea so patently wrong, even as today I can scarcely believe some people can resist an idea so patently right (Savage *et al.*, 1962, p. 76).

From a Bayesian viewpoint, there is nothing to be said for unbiased estimates, while a test of a sharp null hypothesis would be carried out in quite a different way, and if (as is quite likely if in fact  $\theta = 0$ ) the sample size resulting in example (5) were very large, then the posterior probability that  $\theta = 0$  would remain quite large. It can thus be seen that if the stopping rule is seen to be plausible, and it is difficult to avoid it in view of the arguments for the likelihood principle in the last section, then Bayesian statisticians are not embarrassed in the way that classical statisticians are.

## 7.3 Informative stopping rules

### 7.3.1 An example on capture and recapture of fish

A stopping rule  $s$  is said to be *informative* if its distribution depends on  $\theta$  in such a way that it conveys information about  $\theta$  in addition to that available from the values of  $x_1, x_2, \dots, x_s$ . The point of this section is to give a non-trivial example

of an informative stopping rule; the example is due to Roberts (1967).

Consider a capture–recapture situation for a population of fish in a lake. The total number  $N$  of fish is unknown and is the parameter of interest (i.e. it is the  $\theta$  of the problem). It is known that  $R$  of the fish have been captured tagged and released, and we shall write  $S$  for the number of untagged fish. Because  $S=N-R$  and  $R$  is known, we can treat  $S$  as the unknown parameter instead of  $N$ , and it is convenient to do so. A random sample of  $n$  fish is then drawn (without replacement) from the lake. The sample yields  $r$  tagged fish and  $s=n-r$  untagged ones.

Assume that there is an unknown probability  $\pi$  of catching each fish independently of each other. Then the stopping rule is given by the binomial distribution as

$$p(n|R, S, \pi) = \binom{R+S}{n} \pi^n (1-\pi)^{R+S-n} \quad (n = 0, 1, 2, \dots, R+S),$$

so that  $\pi$  is a nuisance parameter such that  $0 \leq \pi \leq 1$ . Note that this stopping rule is informative, because it depends on  $N=R+S$ .

Conditional on  $R$ ,  $N$ ,  $\pi$  and  $n$ , the probability of catching  $r$  tagged fish out of  $n=r+s$  is given by the hypergeometric distribution

$$p(r|R, S, \pi, n) = \frac{\binom{R}{r} \binom{S}{s}}{\binom{R+S}{r+s}}.$$

Because we know  $r$  and  $s$  if and only if we know  $r$  and  $n$ , it follows that

$$\begin{aligned} p(r, s|R, S, \pi) &= p(r|R, S, \pi, n)p(n|R, S, \pi) \\ &= \frac{\binom{R}{r} \binom{S}{s}}{\binom{R+S}{r+s}} \binom{R+S}{r+s} \pi^{r+s} (1-\pi)^{R+S-r-s} \\ &= \binom{R}{r} \binom{S}{s} \pi^{r+s} (1-\pi)^{R+S-r-s}. \end{aligned}$$

### 7.3.2 Choice of prior and derivation of posterior

We assume that not much is known about the number of the fish in the lake *a priori*, and we can represent this by an improper prior

$$p(S) \propto 1.$$

On the other hand, in the process of capturing the first sample  $R$  for tagging, some knowledge will have been gained about the probability  $\pi$  of catching a fish. Suppose that this knowledge can be represented by a beta prior, so that  $\pi \sim \text{Be}(r', R' - r')$ , that is,

$$p(\pi) \propto \pi^{r'-1} (1-\pi)^{R'-r'-1}$$

independently of  $S$ . It follows that

$$\begin{aligned} p(S, \pi | R, r, s) &\propto p(S)p(\pi)p(r, s|R, S, \pi) \\ &= \pi^{r'-1}(1-\pi)^{R'-r'-1} \binom{R}{r} \binom{S}{s} \pi^{r+s}(1-\pi)^{R+S-r-s} \\ &\propto \pi^{(r''-1)-1}(1-\pi)^{(R''-r'')-1} \left[ \binom{S}{s} \pi^{s+1}(1-\pi)^{S-s} \right], \end{aligned}$$

where

$$r'' = r + r', \quad R'' = R + R'.$$

It follows that for given  $\pi$  the distribution of  $S$  is such that  $S-s$  has a negative binomial distribution  $\text{NB}(s+1, \pi)$  (see Appendix A). Summing over  $S$  from  $s$  to  $\infty$ , it can also be seen that

$$p(S, \pi | R, r, s) \propto \pi^{(r''-1)-1}(1-\pi)^{(R''-r'')-1},$$

so that the posterior for  $\pi$  is  $\text{Be}(r''-1, R''-r'')$ .

To find the unconditional distribution of  $S$ , it is necessary to integrate the joint posterior for  $S$  and  $\pi$  over  $\pi$ . It can be shown without great difficulty that the result is that

$$p(S | R, r, s) = \binom{S}{s} \frac{\text{B}(r''+s, R''-r''+S-s)}{\text{B}(r''-1, R''-r'')},$$

where  $\text{B}(\alpha, \beta)$  is the usual beta function. This distribution is sometimes known as the *beta-Pascal distribution*, and its properties are investigated by Raiffa and Schlaifer (1961, Section 7.11). It follows from there that the posterior mean of  $S$  is

$$\mathbb{E}S = (s+1) \left( \frac{R''-2}{r''-2} \right) - 1$$

from which the posterior mean of  $N$  follows as  $N=R+S$ .

### 7.3.3 The maximum likelihood estimator

A standard classical approach would seek to estimate  $S$  or equivalently  $N$  by the maximum likelihood estimator, that is, by the value of  $N$  which maximizes

$$p(r, s | R, S, \pi) = \frac{\binom{R}{r} \binom{S}{s}}{\binom{R+S}{r+s}}.$$

Now it is easily shown that

$$p(r, s | R, S, \pi) / p(r, s | R, S-1, \pi) = (R+S-r-s)S/(S-s)(R+S)$$

and this increases as a function of  $S$  until it reaches unity when  $(r+s)S=(R+S)s$  and then decreases, so that the maximum likelihood estimator of  $S$  is

$$\hat{S} = Rs/r.$$

### 7.3.4 Numerical example

As a numerical example, suppose that the original catch was  $R = 41$  fish and that the second sample results in  $r = 8$  tagged and  $s = 24$  untagged fish. Suppose further that the prior for the probability  $\pi$  of catching a fish is  $\text{Be}(2, 23)$ , so that

$$R = 41, \quad r = 8, \quad s = 24, \quad R' = 25, \quad r' = 2$$

(so that  $R'' = 66$  and  $r'' = 10$ ). Then the posterior mean of  $S$  is

$$\mathbb{E}S = 25 \times 64/8 - 1 = 199,$$

and hence that of  $N$  is  $R + \mathbb{E}S$ , that is,  $41 + 199 = 240$ . On the other hand, the same data with a reference prior  $\text{Be}(0, 0)$  for  $\pi$  (i.e.  $r' = R' = 0$ ) results in a posterior mean for  $S$  of

$$\mathbb{E}S = 25 \times 39/6 - 1 = 161.5,$$

and hence that of  $N$  is  $41 + 161.5 = 202.5$ .

Either of these answers is notably different from the maximum likelihood answer that a classical statistician would be likely to quote, which is

$$\hat{S} = 41 \times 24/8 = 123$$

resulting in  $\hat{N} = 41 + 123 = 164$ . The conclusion is that an informative stopping rule can have a considerable impact on the conclusions, and (though this is scarcely surprising) that prior beliefs about the nuisance parameter  $\pi$  make a considerable difference.

## 7.4 The likelihood principle and reference priors

### 7.4.1 The case of Bernoulli trials and its general implications

Care should be taken when using reference priors as a representation of prior ignorance. We have already seen in Section 2.4 on ‘Dominant likelihoods’ that the improper densities which often arise as reference priors should be regarded as approximations, reflecting the fact that our prior beliefs about an unknown parameter (or some function of it) are more or less uniform over a wide range. A different point to be aware of is that some ways of arriving at such priors, such as Jeffreys’ rule, depend on the experiment that is to be performed, and so on *intentions*. (The same objection applies, of course, to arguments based on data translated likelihoods.) Consequently, an analysis using such a prior is not in accordance with the likelihood principle.

To make this clearer, consider a sequence of independent trials, each of which

results in success with probability  $\pi$  or failure with probability  $1 - \pi$  (i.e. a sequence of Bernoulli trials). If we look at the number of successes  $x$  in a fixed number  $n$  of trials, so that

$$p(x|\pi) \propto \binom{n}{x} \pi^x (1 - \pi)^{n-x} \quad (x = 0, 1, \dots, n)$$

then, as was shown in Section 3.3, Jeffreys' rule results in an arc-sine distribution

$$p(\pi) \propto \pi^{-\frac{1}{2}} (1 - \pi)^{-\frac{1}{2}}$$

for the prior.

Now suppose that we decide to observe the number of failures  $y$  before the  $m$ th success. Evidently, there will be  $m$  successes and  $y$  failures, and the probability of any particular sequence with that number of successes and failures is  $\pi^m (1 - \pi)^y$ . The number of such sequences is  $\binom{m+y-1}{y}$ , because the  $y$  failures and  $m-1$  of the successes can occur in any order, but the sequence must conclude with a success. It follows that

$$p(y|\pi) = \binom{m+y-1}{y} \pi^m (1 - \pi)^y,$$

that is, that  $y \sim \text{NB}(m, \pi)$  has a negative binomial distribution (see Appendix A). For such a distribution

$$L(\pi|y) = m \log \pi + y \log(1 - \pi) + \text{constant},$$

so that

$$\partial^2 L / \partial \pi^2 = -m/\pi^2 - y/(1 - \pi)^2.$$

Because  $Ey = m(1 - \pi)/\pi$ , it follows that

$$I(\pi|y) = m/\pi^2 + m/\pi(1 - \pi) = m/\pi^2(1 - \pi),$$

so that Jeffreys' rule implies that we should take a prior

$$p(\pi) \propto \pi^{-1}(1 - \pi)^{-\frac{1}{2}},$$

that is,  $\pi \sim \text{Be}(0, \frac{1}{2})$  instead of  $\pi \sim \text{Be}(\frac{1}{2}, \frac{1}{2})$

## 7.4.2 Conclusion

Consequently, on being told that an experiment resulted in, say, ten successes and ten failures, Jeffreys' rule does not allow us to decide which prior to use until we know whether the experimental design involved a fixed number of trials, or waiting until a fixed number of successes, or some other method. This clearly violates the likelihood principle (cf. Lindley, 1971a, Section 12.4); insofar as they appear to include Jeffreys' work, it is hard to see how Berger and Wolpert (1988, Section 4.1.2) come to the conclusion that ‘... use of

noninformative priors, purposely not involving subjective prior opinions ... is consistent with the LP [Likelihood Principle']). Some further difficulties inherent in the notion of a uniform reference prior are discussed in Hill (1980) and in Berger and Wolpert (1988).

However, it has been argued that a reference prior should express ignorance *relative to* the information which can be supplied by a particular experiment; see Box and Tiao (1992, Section 1.3). In any case, provided they are used critically, reference priors can be very useful, and, of course, if there is a reasonable amount of detail, the precise form of the prior adopted will not make a great deal of difference.

## 7.5 Bayesian decision theory

### 7.5.1 The elements of game theory

Only a very brief account of this important topic is included here; readers who want to know more should begin by consulting Berger (1985) and Ferguson (1967).

The elements of decision theory are very similar to those of the mathematical theory of games as developed by von Neumann and Morgenstern (1953), although for statistical purposes one of the players is nature (in some sense) rather than another player. Only those aspects of the theory of games which are strictly necessary are given here; an entertaining popular account is given by Williams (1966). A two-person zero-sum game  $(\Theta, A, L)$  has the following three basic elements:

1. A non-empty set  $\Theta$  of possible states of nature  $\theta$ , sometimes called the *parameter space*;
2. A non-empty set  $A$  of *actions* available to the statistician;
3. A *loss function*  $L$ , which defines the loss  $\mathcal{L}(\theta, a)$  which a statistician suffers if he takes action  $a$  when the true state of nature is  $\theta$  (this loss being expressed as a real number).

A *statistical decision problem* or a statistical game is a game  $(\Theta, A, \mathcal{L})$  coupled with an experiment whose result  $x$  lies in a sample space  $\mathcal{X}$  and is randomly distributed with a density  $p(x|\theta)$  which depends on the state  $\theta \in \Theta$  ‘chosen’ by nature. The data  $x$  can be, and usually is, more than one-dimensional.

Now suppose that on the basis of the result  $x$  of the experiment, the statistician

chooses an action  $d(x) \in A$ , resulting in a random loss  $L(\theta, d(x))$ . Taking expectations over possible outcomes  $x$  of the experiment, we get a *risk function*

$$\begin{aligned} R(\theta, d) &= \mathbb{E}\mathcal{L}(\theta, d(x)) \\ &= \int \mathcal{L}(\theta, d(x)) p(x|\theta) dx \end{aligned}$$

which depends on the true state of nature and the form of the function  $d$  by which the action to be taken once the result of the experiment is known is determined. It is possible that this expectation may not exist, or may be infinite, but we shall exclude such cases and define a (nonrandomized) *decision rule* or a *decision function* as any function  $d$  for which  $R(\theta, d)$  exists and is finite for all  $\theta \in \Theta$ .

An important particular case of an action dependent on the outcome of an experiment is that of a *point estimators* of a parameter  $\theta$ , that is, to find a single number  $\hat{\theta} = \hat{\theta}(x)$  from the data which in some way best represents a parameter under study.

For classical statisticians, an important notion is admissibility. An estimator  $\hat{\theta}^*$  is said to *dominate* an estimator  $\hat{\theta}$  if

$$R(\theta, \hat{\theta}^*) \leq R(\theta, \hat{\theta})$$

for *all*  $\theta$  with strict inequality for *at least one* value of  $\theta$ . An estimator  $\hat{\theta}$  is said to be *inadmissible* if it is dominated by some other action  $\hat{\theta}^*$ . The notion of admissibility is clearly related to that of Pareto optimality. [‘Pareto optimality’ is a condition where no one is worse off in one state than another but someone is better off, and there is no state ‘Pareto superior’ to it (i.e. in which more people would be better off without anyone being worse off).]

From the point of view of classical statisticians, it is very undesirable to use inadmissible actions. From a Bayesian point of view, admissibility is not generally very relevant. In the words of Press (1989, Section 2.3.1),

Admissibility requires that we average over all possible values of the observable random variables (the expectation is taken with respect to the observables). In experimental design situations, statisticians must be concerned with the performance of estimators for many possible situational repetitions and for many values of the observables, and then admissibility is a reasonable Bayesian performance criterion. In most other situations, however, statisticians are less concerned with performance of an estimator over many possible samples that have yet to be observed than they are with the performance of their estimator conditional upon having observed this

particular data set conditional upon having observed this particular data set and conditional upon all prior information available. For this reason, in non-experimental design situations, admissibility is generally not a compelling criterion for influencing our choice of estimator.

For the moment we shall, however, follow the investigation from a classical standpoint.

From a Bayesian viewpoint, we must suppose that we have prior beliefs about  $\theta$  which can be expressed in terms of a prior density  $p(\theta)$ . The Bayes risk  $r(d)$  of the decision rule  $d$  can then be defined as the expectation of  $R(\theta, d)$  over all possible values of  $\theta$ , that is,

$$r(d) = \mathbb{E} R(\theta, d) = \int R(\theta, d) p(\theta) d\theta.$$

It seems sensible to minimize one's losses, and accordingly a *Bayes decision rule*  $d$  is defined as one which minimizes the Bayes risk  $r(d)$ . Now

$$\begin{aligned} r(d) &= \int R(\theta, d) p(\theta) d\theta \\ &= \iint \mathcal{L}(\theta, d(x)) p(x|\theta) p(\theta) dx d\theta \\ &= \iint \mathcal{L}(\theta, d(x)) p(x, \theta) dx d\theta \\ &= \int \left\{ \int \mathcal{L}(\theta, d(x)) p(\theta|x) d\theta \right\} p(x) dx. \end{aligned}$$

It follows that if the posterior expected loss of an action  $a$  is defined by

$$\rho(a, x) = \int \mathcal{L}(\theta, a) p(\theta|x) d\theta$$

then the Bayes risk is minimized if the decision rule  $d$  is so chosen that  $\rho(d(x), x)$  is a minimum for all  $x$  (technically, for those who know measure theory, it need only be a minimum for almost all  $x$ ).

Raiffa and Schlaifer (1961, Sections 1.2–1.3) refer to the overall minimization of  $r(d)$  as the *normal form* of Bayesian analysis and to the minimization of  $\rho(d(x), x)$  for all  $x$  as the *extensive form*; the aforementioned remark shows that the two are equivalent.

When a number of possible prior distributions are under consideration, one sometimes finds that the term Bayes rule as such is restricted to rules restricting from *proper* priors, while those resulting from *improper* priors are called *generalized* Bayes rules. Further extensions are mentioned in Ferguson (1967).

## 7.5.2 Point estimators resulting from quadratic loss

A Bayes decision rule in the case of point estimation is referred to as a *Bayes estimator*. In such problems, it is easiest to work with *quadratic loss*, that is, with a *squared-error* loss function

$$\mathcal{L}(\theta, a) = (\theta - a)^2.$$

In this case,  $\rho(a, x)$  is the *mean square error*, that is,

$$\begin{aligned}\rho(a, x) &= \int [\theta - a]^2 p(\theta|x) d\theta \\ &= \int [(\theta - E(\theta|x)) + (E(\theta|x) - a)]^2 p(\theta|x) d\theta \\ &= \int \{(\theta - E(\theta|x))^2 p(\theta|x) d\theta \\ &\quad + 2(E(\theta|x) - a) \int \{(\theta - E(\theta|x)) p(\theta|x) d\theta \\ &\quad + \{E(\theta|x) - a\}^2.\end{aligned}$$

The second term clearly vanishes, so that

$$\rho(a, x) = V(\theta|x) + \{E(\theta|x) - a\}^2$$

which is a minimum when  $a = E(\theta|x)$ , so that a Bayes estimator  $d(x)$  is the posterior mean of  $\theta$ , and in this case  $\rho(d(x), x)$  is the posterior variance of  $\theta$ .

### 7.5.3 Particular cases of quadratic loss

As a particular case, if we have a single observation  $x \sim N(\theta, \phi)$ , where  $\phi$  is known and our prior for  $\theta$  is  $N(\theta_0, \phi_0)$ , so that our posterior is  $N(\theta_1, \phi_1)$  (cf. Section 2.2 on ‘Normal prior and likelihood’), then an estimate of  $\theta$  that minimizes quadratic loss is the posterior mean  $\theta_1$ , and if that estimate is used the mean square error is the posterior variance  $\phi$ .

For another example, suppose that  $x \sim P(\lambda)$ , that is, that  $x$  has a Poisson distribution of mean  $\lambda$ , and that our prior density for  $\lambda$  is  $p(\lambda)$ . First note that the predictive density of  $x$  is

$$\begin{aligned}p(x) &= \int p(x, \lambda) d\lambda = \int p(x|\lambda) p(\lambda) d\lambda \\ &= \int (x!)^{-1} \lambda^x \exp(-\lambda) p(\lambda) d\lambda.\end{aligned}$$

To avoid ambiguity in what follows,  $p_{\tilde{x}}(x)$  is used for this predictive distribution, so that  $p_{\tilde{x}}(z)$  just denotes  $\int (z!)^{-1} \lambda^z \exp(-\lambda) p(\lambda) d\lambda$ . Then as

$$\begin{aligned}p(\lambda|x) &= p(\lambda)p(x|\lambda)/p_{\tilde{x}}(x) \\ &= (x!)^{-1} \lambda^x \exp(-\lambda) p(\lambda)/p_{\tilde{x}}(x),\end{aligned}$$

it follows that the posterior mean is

$$\begin{aligned}\mathbb{E}(\lambda|x) &= \int \lambda (x!)^{-1} \lambda^x \exp(-\lambda) p(\lambda) d\lambda / p_{\tilde{x}}(x) \\ &= (x+1) \int \{(x+1)!\}^{-1} \lambda^{x+1} \exp(-\lambda) p(\lambda) d\lambda / p_{\tilde{x}}(x) \\ &= (x+1) p_{\tilde{x}}(x+1) / p_{\tilde{x}}(x).\end{aligned}$$

We shall return to this example in Section 7.8 in connection with empirical Bayes methods.

Note incidentally that if the prior is  $\lambda \sim S_0^{-1} \chi_\nu^2$ , then the posterior is  $(S_0 + 2)^{-1} \chi_{\nu+2x}^2$  (as shown in Section 3.4 on ‘The Poisson distribution’), so that in this particular case

$$\mathbb{E}(\lambda|x) = (\nu + 2x)/(S_0 + 2).$$

### 7.5.4 Weighted quadratic loss

However, you should not go away with the conclusion that the solution to all problems of point estimation from a Bayesian point of view is simply to quote the posterior mean – the answer depends on the loss function. Thus, if we take as loss function a *weighted quadratic loss*, that is,

$$\mathcal{L}(\theta, a) = w(\theta)(\theta - a)^2,$$

then

$$\rho(a, x) = \int w(\theta)[\theta - a]^2 p(\theta|x) dx.$$

If we now define

$$E^w(\theta|x) = \frac{\mathbb{E}(w(\theta)\theta|x)}{\mathbb{E}(w(\theta)|x)} = \frac{\int w(\theta)\theta p(\theta|x) d\theta}{\int w(\theta)p(\theta|x) d\theta},$$

then similar calculations to those above show that

$$\rho(a, x) = \mathbb{E}[w(\theta)\{\theta - a\}^2|x] + [E^w(\theta|x) - a]^2,$$

and hence that a Bayes decision results if

$$d(x) = E^w(\theta|x),$$

that is,  $d(x)$  is a weighted posterior mean of  $\theta$ .

### 7.5.5 Absolute error loss

A further answer results if we take as loss function the absolute error

$$\mathcal{L}(\theta, a) = |\theta - a|$$

in which case

$$\rho(a, x) = \int |\theta - a| p(\theta|x) d\theta$$

is sometimes referred to as the *mean absolute deviation* or MAD. In this case,

any median  $m(x)$  of the posterior distribution given  $x$ , that is, any value such that

$$\mathbb{P}(\theta \leq m(x) | x) \geq \frac{1}{2} \quad \text{and} \quad \mathbb{P}(\theta \geq m(x) | x) \geq \frac{1}{2}$$

is a Bayes rule. To show this suppose that  $d(x)$  is any other rule and, for definiteness, that  $d(x) > m(x)$  for some particular  $x$  (the proof is similar if  $d(x) < m(x)$ ). Then

$$L(\theta, m(x)) - L(\theta, d(x)) = \begin{cases} m(x) - d(x) & \text{if } \theta \leq m(x) \\ 2\theta - (m(x) + d(x)) & \text{if } m(x) < \theta < d(x) \\ d(x) - m(x) & \text{if } \theta \geq d(x) \end{cases}$$

while for  $m(x) < \theta < d(x)$

$$2\theta - (m(x) + d(x)) < \theta - m(x) < d(x) - m(x),$$

so that

$$L(\theta, m(x)) - L(\theta, d(x)) \leq \begin{cases} m(x) - d(x) & \text{if } \theta \leq m(x) \\ d(x) - m(x) & \text{if } \theta > m(x), \end{cases}$$

and hence on taking expectations over  $\theta$

$$\begin{aligned} \rho(m(x), x) - \rho(d(x), x) &\leq \{m(x) - d(x)\}\mathbb{P}(\theta \leq m(x) | x) \\ &\quad + \{d(x) - m(x)\}\mathbb{P}(\theta > m(x) | x) \\ &= \{d(x) - m(x)\}\{-\mathbb{P}(\theta \leq m(x) | x)\} \\ &\quad + 1 - \mathbb{P}(\theta \leq m(x) | x) \\ &\leq 0 \end{aligned}$$

from which it follows that taking the posterior median is indeed the appropriate Bayes rule for this loss function. More generally, a loss function  $L(\theta, a)$  which is  $v|\theta - a|$  if  $\theta \geq a$  but  $u|\theta - a|$  if  $\theta < a$  results in a Bayes estimator which is a  $v/(v+u)$  fractile of the posterior distribution.

## 7.5.6 Zero-one loss

Yet another answer results from the loss function

$$\mathcal{L}(\theta, a) = \begin{cases} 0 & \text{if } |\theta - a| \leq \varepsilon \\ 1 & \text{if } |\theta - a| > \varepsilon \end{cases}$$

which results in

$$\rho(a, x) = \mathbb{P}(|\theta - a| > \varepsilon | x) = 1 - \mathbb{P}(|\theta - a| \leq \varepsilon | x).$$

Consequently, if a modal interval of length  $2\varepsilon$  is defined as an interval

$$(mod^\varepsilon(x) - \varepsilon, mod^\varepsilon(x) + \varepsilon)$$

which has highest probability for given  $\varepsilon$ , then the midpoint  $mod^\varepsilon(x)$  of this interval is a Bayes estimate for this loss function. [A modal interval is, of course, just another name for a highest density region (HDR) except for the presumption that  $\varepsilon$  will usually be small.] If  $\varepsilon$  is fairly small, this value is clearly very close to the posterior mode of the distribution, which in its turn will be close to the

maximum likelihood estimator if the prior is reasonably smooth.

Thus all three of mean, median and mode of the posterior distribution can arise as Bayes estimators for suitable loss functions (namely, quadratic error, absolute error and zero-one error, respectively).

### 7.5.7 General discussion of point estimation

Some Bayesian statisticians pour scorn on the whole idea of point estimators; see Box and Tiao (1992 Section A5.6). There are certainly doubtful points about the preceding analysis. It is difficult to be convinced in any particular case that a particular loss function represents real economic penalties in a particular case, and in many scientific contexts, it is difficult to give any meaning at all to the notion of the loss suffered by making a wrong point estimate. Certainly, the same loss function will not be valid in all cases. Moreover, even with quadratic loss, which is often treated as a norm, there are problems which do not admit of an easy solution. If, for example,  $x_1, x_2, \dots, x_n \sim N(1/\theta, 1)$ , then it would seem reasonable to estimate  $\theta$  by  $1/\bar{x}$ , and yet the mean square error  $E(1/\bar{x} - \theta)^2$  is infinite. Of course, such decision functions are excluded by requiring that the risk function should be finite, but this is clearly a case of what Good (1965, Section 6.2) referred to as ‘adhockery’.

Even though there are cases (e.g. where the posterior distribution is bimodal) in which there is no sensible point estimator, I think there are cases where it is reasonable to ask for such a thing, though I have considerable doubts as to whether the ideas of decision theory add much to the appeal that quantities such as the posterior mean, median or mode have in themselves.

## 7.6 Bayes linear methods

### 7.6.1 Methodology

Bayes linear methods are closely related to point estimators resulting from quadratic loss. Suppose that we restrict attention to decision rules  $d(x)$  which are constrained to be a *linear* function  $d(x) = \alpha + \beta y$  of some known function  $y=y(x)$  of  $x$  and seek for a rule which, subject to this constraint, has minimum Bayes risk  $r(d)$ . The resulting rule will not usually be a Bayes rule, but will not, on the other hand, necessitate a complete specification of the prior distribution. As we have seen that it can be very difficult to provide such a specification, there are

real advantages to Bayes linear methods. To find such an estimator we need to minimize

$$\begin{aligned} r(d) &= \iint (\theta - \alpha - \beta y)^2 p(\theta|x) p(x) d\theta dx \\ &= E\{(\theta - E\theta) - \beta(y - Ey) + (E\theta - \alpha - \beta y)\}^2 \\ &= V\theta - 2\beta C(\theta, y) + \beta^2 V_y + (E\theta - \alpha - \beta E_y)^2 \end{aligned}$$

(since cross terms involving  $E\theta - \alpha - \beta E_y$  clearly vanish). By setting  $\partial r/\partial\alpha = 0$ , we see that the values  $\hat{\alpha}$  and  $\hat{\beta}$  which minimize  $r$  satisfy

$$\hat{\alpha} = E\theta - \hat{\beta} E_y,$$

and then setting  $\partial r/\partial\beta = 0$ , we see that

$$\hat{\beta} = (V_y)^{-1} C(\theta, y),$$

so that the Bayes linear estimator is

$$d(x) = E\theta + (V_y)^{-1} C(\theta, y)(y - E_y).$$

It should be noted that, in contrast to Bayes decision rules, Bayes linear estimators do *not* depend solely on the observed data  $x$  – they also depend on the *distribution* of the data through  $E_y$ ,  $V_y$  and  $C(\theta, y)$ . For that reason they violate the likelihood principle.

## 7.6.2 Some simple examples

### 7.6.2.1 Binomial mean

Suppose that  $x \sim B(n, \pi)$  and that  $y=x$ . Then using the results at the end of Section 1.5 on ‘Means and variances’

$$Ex = E(E(x|\pi)) = nE\pi$$

$$Vx = E(V(x|\pi)) + V(E(x|\pi)) = En\pi(1-\pi) + V(n\pi) = nE\pi(1-\pi) + n^2V\pi.$$

Since  $E(\pi x) = E(\pi E(x|\pi)) = nE\pi^2$ , we see that

$$C(\pi, x) = E(\pi x) - E\pi Ex = nE\pi^2 - n(E\pi)^2 = nV\pi,$$

so that

$$d(x) = E\pi + (Vx)^{-1} C(\pi, x)(x - Ex) = \lambda(x/n) + (1-\lambda)E\pi.$$

where

$$\lambda = n(Vx)^{-1} C(\pi, x) = \frac{V\pi}{V\pi + n^{-1} E\pi(1-\pi)}.$$

Note that the resulting posterior estimator for  $\pi$  depends only on  $E\pi$  and  $V\pi$ . This is an advantage if you think you can be precise enough about your prior knowledge of  $\pi$  to specify these quantities but find difficulty in giving a full specification of your prior which would be necessary for you to find the Bayes

estimator for quadratic loss  $E(\pi|x)$ ; the latter cannot be evaluated in terms of a few summary measures of the prior distribution. On the other hand, we have had to use, for example, the fact that  $E(x|\pi) = n\pi$  thus taking into account observations which might have been, but were not in fact, made, in contravention of the likelihood principle.

### 7.6.2.2 Negative binomial distribution

Suppose that  $z \sim NB(m, \pi)$  and that  $y=z$ . Then similar formulae are easily deduced from the results in Appendix A. The fact that different formulae from those in the binomial case result when  $m=x$  and  $z=n-x$ , so that in both cases we have observed  $x$  successes and  $n - x$  failures, reflects the fact that this method of inference does not obey the likelihood principle.

### 7.6.2.3 Estimation on the basis of the sample mean

Suppose that  $x_1, x_2, \dots, x_n$  are such that  $E(x_i|\theta) = \theta$ ,  $V(x_i|\theta) = \phi$  and  $C(x_i, x_j|\theta) = \kappa$  but that you know nothing more about the distribution of the  $x_i$ . Then  $E(\bar{x}|\theta) = \theta$  and

$$\begin{aligned} V(\bar{x}|\theta) &= \frac{1}{n^2} V\left(\sum_i x_i\right) = \frac{1}{n^2} \left\{ \sum_i Vx_i + \sum_i \sum_j C(x_i, x_j) \right\} \\ &= n^{-2} [n\phi + n(n-1)\kappa], \end{aligned}$$

so that using the results at the end of Section 1.5 on ‘Means and variances’  $E\bar{x} = E(E(\bar{x}|\theta)) = E\theta$  and

$$V\bar{x} = E[V(\bar{x}|\theta)] + V(E(\bar{x}|\theta)) = n^{-1} [\phi + (n-1)\kappa] + V\theta.$$

Since  $E(\theta\bar{x}) = E(\theta E(\bar{x}|\theta)) = E\theta^2$ , we see that

$$C(\theta, \bar{x}) = E(\theta\bar{x}) - E\theta E\bar{x} = E\theta^2 - (E\theta)^2 = V\theta.$$

It follows that

$$d(x) = E\theta + (V\bar{x})^{-1} C(\theta, \bar{x})(\bar{x} - E\bar{x}) = \lambda\bar{x} + (1-\lambda)E\theta,$$

where

$$\lambda = nV\theta / \{[\phi + (n-1)\kappa] + nV\theta\}.$$

### 7.6.3 Extensions

Bayes linear methods can be applied when there are several unknown parameters. A brief account can be found in O’Hagan (1994, Section 6.48 et seq.) and full coverage is given in Goldstein and Wooff (2007).

## 7.7 Decision theory and hypothesis testing

### 7.7.1 Relationship between decision theory and classical hypothesis testing

It is possible to reformulate hypothesis testing in the language of decision theory. If we want to test  $H_0 : \theta \in \Theta_0$  versus  $H_1 : \theta \in \Theta_1$ , we have two actions open to us, namely,

$$a_0 : \text{accept } H_0 \quad \text{and} \quad a_1 : \text{accept } H_1.$$

As before, we shall write  $\pi_0$  and  $\pi_1$  for the prior probabilities of  $H_0$  and  $H_1$  and  $p_0$  and  $p_1$  for their posterior probabilities and

$$B = \frac{(p_0/p_1)}{(\pi_0/\pi_1)}$$

for the Bayes factor. We also need the notation

$$\rho_0(\theta) = p(\theta)/\pi_0 \quad \text{for } \theta \in \Theta_0$$

$$\rho_1(\theta) = p(\theta)/\pi_1 \quad \text{for } \theta \in \Theta_1,$$

where  $p(\theta)$  is the prior density function.

Now let us suppose that there is a loss function  $\mathcal{L}(\theta, a)$  defined by

$$\begin{array}{lll} a \setminus \theta & \theta \in \Theta_0 & \theta \in \Theta_1 \\ a_0 & 0 & 1 \\ a_1 & 1 & 0, \end{array}$$

so that the use of a decision rule  $d(x)$  results in a posterior expected loss function

$$\rho(a_0, x) = p_1$$

$$\rho(a_1, x) = p_0,$$

so that a decision  $d(x)$  which minimizes the posterior expected loss is just a decision to accept the hypothesis with the greater posterior probability, which is the way of choosing between hypotheses suggested when hypothesis testing was first introduced.

More generally, if there is a ‘0– $K_i$ ’ loss function, that is,

$$\begin{array}{lll} a \setminus \theta & \theta \in \Theta_0 & \theta \in \Theta_1 \\ a_0 & 0 & K_0 \\ a_1 & K_1 & 0, \end{array}$$

then the posterior expected losses of the two actions are

$$\rho(a_0, x) = K_0 p_1,$$

$$\rho(a_1, x) = K_1 p_0,$$

so that a Bayes decision rule results in rejecting the null hypothesis, that is, in taking action  $a_i$ , if and only if  $K_i p_0 < K_0 p_1$ , that is,

$$B = \frac{(p_0/p_1)}{(\pi_0/\pi_1)} < \frac{(K_0/K_1)}{(\pi_0/\pi_1)}.$$

In the terminology of classical statistics, this corresponds to the use of a rejection region

$$R = \{x; B < K_0\pi_1/K_1\pi_0\}.$$

When hypothesis testing was first introduced in Section 4.1, we noted that in the case where  $\Theta_0 = \{\theta_0\}$  and  $\Theta_1 = \{\theta_1\}$  are simple hypotheses, then Bayes theorem implies that

$$B = \frac{p(x|\theta_0)}{p(x|\theta_1)},$$

so that the rejection region takes the form

$$R = \{x; p(x|\theta_0)/p(x|\theta_1) < K_0\pi_1/K_1\pi_0\},$$

which is the likelihood ratio test prescribed by Neyman–Pearson theory. A difference is that in the Neyman–Pearson theory, the ‘critical value’ of the rejection region is determined by fixing the size  $\alpha$ , that is, the probability that  $x$  lies in the rejection region  $R$  if the null hypothesis is true, whereas in a decision theoretic approach, it is fixed in terms of the loss function and the prior probabilities of the hypotheses.

## 7.7.2 Composite hypotheses

If the hypotheses are composite (i.e. not simple), then, again as in Section 4.1 on ‘Hypothesis testing’,

$$B = \frac{(p_0/p_1)}{(\pi_0/\pi_1)} = \frac{\int_{\theta \in \Theta_0} p(x|\theta)\rho_0(\theta) d\theta}{\int_{\theta \in \Theta_1} p(x|\theta)\rho_1(\theta) d\theta},$$

so that there is still a rejection region that can be interpreted in a similar manner. However, it should be noted that classical statisticians faced with similar problems are more inclined to work in terms of a likelihood ratio

$$\frac{\max_{\theta \in \Theta_0} p(x|\theta)}{\max_{\theta \in \Theta_1} p(x|\theta)}$$

(cf. Lehmann, 1986, Section 1.7). In fact, it is possible to express quite a lot of the ideas of classical statistics in a language involving loss functions.

It may be noted that it is easy to extend the above discussion about dichotomies (i.e. situations where a choice has to be made between two hypotheses) to deal with trichotomies or polytomies, although some theories of statistical inference find choices between more than two hypotheses difficult to deal with.

## 7.8 Empirical Bayes methods

### 7.8.1 Von Mises' example

Only a very brief idea about empirical Bayes methods will be given in this chapter; more will be said about this topic in Chapter 8 and a full account can be found in Maritz and Lwin (1989). One of the reasons for this brief treatment is that, despite their name, very few empirical Bayes procedures are, in fact, Bayesian; for a discussion of this point see, for example, Deely and Lindley (1981).

The problems we will consider in this section are concerned with a sequence  $x_i$  of observations such that the distribution of the  $i$ th observation  $x_i$  depends on a parameter  $\theta_i$ , typically in such a way that  $p(x_i|\theta_i)$  has the same functional form for all  $i$ . The parameters  $\theta_i$  are themselves supposed to be a random sample from some (unknown) distribution, and it is this unknown distribution that plays the role of a prior distribution and so accounts for the use of the name of Bayes. There is a clear contrast with the situation in the rest of the book, where the prior distribution represents our prior beliefs, and so by definition it cannot be unknown. Further, the prior distribution in empirical Bayes methods is usually given a frequency interpretation, by contrast with the situation arising in true Bayesian methods.

One of the earliest examples of an empirical Bayes procedure was due to von Mises (1942). He supposed that in examining the quality of a batch of water for possible contamination by certain bacteria,  $m = 5$  samples of a given volume were taken, and he was interested in determining the probability  $\theta$  that a sample contains at least one bacterium. Evidently, the probability of  $x$  positive result in the 5 samples is

$$p(x|\theta) = \binom{5}{x} \theta^x (1-\theta)^{5-x}$$

for a given value of  $\theta$ . If the same procedure is to be used with a number of batches of different quality, then the predictive distribution (denoted  $p_{\tilde{x}}(x)$  to avoid ambiguity) is

$$p_{\tilde{x}}(x) = \int \binom{5}{x} \theta^x (1-\theta)^{5-x} p(\theta) d\theta,$$

where the density  $p(\theta)$  represents the variation of the quality  $\theta$  of batches. [If  $p(\theta)$  comes from the beta family, and there is no particular reason why it should, then  $p_{\tilde{x}}(x)$  is a beta-binomial distribution, as mentioned at the end of Section 3.1

on ‘The binomial distribution’]. In his example, von Mises wished to estimate the density function  $p(\theta)$  on the basis of  $n = 3420$  observations.

### 7.8.2 The Poisson case

Instead of considering the binomial distribution further, we shall consider a problem to do with the Poisson distribution which, of course, provides an approximation to the binomial distribution when the number  $m$  of samples is large and the probability  $\theta$  is small. Suppose that we have observations  $x_i \sim P(\lambda_i)$  where the  $\lambda_i$  have a distribution with a density  $p(\lambda)$ , and that we have available  $n$  past observations, among which  $f_n(x)$  were equal to  $x$  for  $x = 0, 1, 2, \dots$ . Thus,  $f_n(x)$  is an empirical frequency and  $f_n(x)/n$  is an estimate of the predictive density  $p_{\tilde{x}}(x)$ . As  $x$  has a Poisson distribution for given  $\lambda$

$$\begin{aligned} p_{\tilde{x}}(x) &= \int p(x, \lambda) d\lambda = \int p(x|\lambda)p(\lambda) d\lambda \\ &= \int (x!)^{-1}\lambda^x \exp(-\lambda)p(\lambda) d\lambda. \end{aligned}$$

Now suppose that, with this past data available, a new observation  $\xi$  is made, and we want to say something about the corresponding value of  $\lambda$ . In Section 7.5 on ‘Bayesian decision theory’, we saw that the posterior mean of  $\lambda$  is

$$E(\lambda|\xi) = \frac{(\xi + 1)p_{\tilde{x}}(\xi + 1)}{p_{\tilde{x}}(\xi)}.$$

To use this formula, we need to know the prior  $p(\lambda)$  or at least to know  $p_{\tilde{x}}(\xi)$  and  $p_{\tilde{x}}(\xi + 1)$ , which we do not know. However, it is clear that a reasonable estimate of  $p_{\tilde{x}}(\xi)$  is  $(f_n(\xi) + 1)/(n + 1)$ , after allowing for the latest observation. Similarly, a reasonable estimate for  $p_{\tilde{x}}(\xi + 1)$  is  $f_n(\xi + 1)/(n + 1)$ . It follows that a possible point estimate for the current value of  $\lambda$ , corresponding to the value  $E(\lambda|x)$  resulting from a quadratic loss function, is

$$\delta_n = \frac{(\xi + 1)f_n(\xi + 1)}{f_n(\xi) + 1}.$$

This formula could be used in a case like that investigated by von Mises if the number  $m$  of samples taken from each batch were fairly large and the probability  $\theta$  that a sample contained at least one bacterium were fairly small, so that the Poisson approximation to the binomial could be used.

This method can easily be adapted to any case where the posterior mean  $E(\theta|\xi)$  of the parameter of interest takes the form

$$\frac{c(\xi)p_{\tilde{x}}(\xi + 1)}{p_{\tilde{x}}(\xi)},$$

and there are quite a number of such cases (Maritz and Lwin, 1989, Section 1.3).

Going back to the Poisson case, if it were known that the underlying distribution  $p(\lambda)$  were of the form  $S_0^{-1} \chi_\nu^2$  for some  $S_0$  and  $\nu$ , then it is known (cf. Section 7.5) that

$$E(\lambda|\xi) = (\nu + 2\xi)/(S_0 + 2).$$

In this case, we could use  $x_1, x_2, \dots, x_n$  to estimate  $S_0$  and  $\nu$  in some way, by, say,  $\hat{S}_0$  and  $\hat{\nu}$ , giving an alternative point estimate for the current value of

$$(\hat{\nu} + 2)/(\hat{S}_0 + 2).$$

The advantage of an estimate like this is that, because, considered as a function of  $\xi$ , it is smoother than  $\delta_n$ , it could be expected to do better. This is analogous with the situation in regression analysis, where a fitted regression line can be expected to give a better estimate of the mean of the dependent variable  $y$  at a particular value of the independent variable  $x$  than you would get by concentrating on values of  $y$  obtained at that single value of  $x$ . On the other hand, the method just described does depend on assuming a particular form for the prior, which is probably not justifiable. There are, however, other methods of producing a ‘smoother’ estimate.

Empirical Bayes methods can also be used for testing whether a parameter  $\theta$  lies in one or another of a number of sets, that is, for hypothesis testing and its generalizations.

## 7.9 Exercises on Chapter 7

1. Show that in any experiment  $E$  in which there is a possible value  $y$  for the random variable  $\tilde{x}$  such that  $p_{\tilde{x}}(y|\theta) = 0$ , then if  $z$  is any other possible value of  $\tilde{x}$ , the statistic  $t=t(x)$  defined by

$$t(x) = \begin{cases} z & \text{if } x = y \\ x & \text{if } x \neq y \end{cases}$$

is sufficient for  $\theta$  given  $x$ . Hence, show that if  $\tilde{x}$  is a continuous random variable, then a naïve application of the weak sufficiency principle as defined in Section 7.1 would result in  $Ev\{E, y, \theta\} = Ev\{E, z, \theta\}$  for any two possible values  $y$  and  $z$  of  $\tilde{x}$ .

2. Consider an experiment  $E = \{\tilde{x}, \theta, p(x|\theta)\}$ . We say that *censoring* (strictly speaking, fixed censoring) occurs with censoring mechanism  $g$  (a known function of  $x$ ) when, instead of  $\tilde{x}$ , one observes  $y=g(x)$ . A typical example occurs when we report  $x$  if  $x < k$  for some fixed  $k$ , but otherwise simply report that  $x \geq k$ . As a result, the experiment really performed is  $E^g = \{\tilde{y}, \theta, p(y|\theta)\}$ . A second method with censoring mechanism  $h$  is said to be *equivalent* to the first when

$$g(x) = g(x') \text{ if and only if } h(x) = h(x').$$

As a special case, if  $g$  is one-to-one then the mechanism is said to be equivalent to no censoring. Show that if two censoring mechanisms are equivalent, then the likelihood principle implies that

$$Ev\{E^g, x, \theta\} = Ev\{E^h, x, \theta\}.$$

3. Suppose that the density function  $p(x|\theta)$  is defined as follows for  $x = 1, 2, 3, \dots$  and  $\theta = 1, 2, 3, \dots$ . If  $\theta$  is even, then

$$p(x|\theta) = \begin{cases} \frac{1}{3} & \text{if } x = \theta/2, 2\theta \text{ or } 2\theta + 1 \\ 0 & \text{otherwise} \end{cases}$$

if  $\theta$  is odd but  $\theta \neq 1$ , then

$$p(x|\theta) = \begin{cases} \frac{1}{3} & \text{if } x = (\theta - 1)/2, 2\theta \text{ or } 2\theta + 1 \\ 0 & \text{otherwise} \end{cases}$$

while if  $\theta = 1$  then

$$p(x|\theta) = \begin{cases} \frac{1}{3} & \text{if } x = \theta, 2\theta \text{ or } 2\theta + 1 \\ 0 & \text{otherwise.} \end{cases}$$

Show that, for any  $x$  the data intuitively give equal support to the three possible values of  $\theta$  compatible with that observation, and hence that on likelihood grounds any of the three would be a suitable estimate. Consider, therefore, the three possible estimators  $d_1$ ,  $d_2$  and  $d_3$  corresponding to the

smallest, middle and largest possible  $\theta$ . Show that

$$p(d_2 = 1) = \begin{cases} \frac{1}{3} & \text{when } \theta \text{ is even} \\ 0 & \text{otherwise,} \end{cases}$$

$$p(d_3 = 1) = \begin{cases} \frac{1}{3} & \text{when } \theta \text{ is odd but } \theta \neq 1 \\ 0 & \text{otherwise} \end{cases}$$

but that

$$p(d_1 = 1) = \begin{cases} 1 & \text{when } \theta = 1 \\ \frac{2}{3} & \text{otherwise} \end{cases}$$

Does this apparent discrepancy cause any problems for a Bayesian analysis (due to G. Monette and D. A. S. Fraser)?

**4.** A drunken soldier, starting at an intersection O in a city which has square blocks, staggers around a random path trailing a taut string. Eventually, he stops at an intersection (after walking at least one block) and buries a treasure. Let  $\theta$  denote the path of the string from O to the treasure. Letting N, S, E and W stand for a path segment one block long in the indicated direction, so that  $\theta$  can be expressed as a sequence of such letters, say  $\theta = NNESWSWW$ . (Note that NS, SN, EW and WE cannot appear as the taut string would be rewound). After burying the treasure, the soldier walks one block further in a random direction (still keeping the string taut). Let  $X$  denote this augmented path, so that  $X$  is one of  $\theta N$ ,  $\theta S$ ,  $\theta E$  and  $\theta W$ , each with probability  $\frac{1}{4}$ . You observe  $X$  and are then to find the treasure. Show that if you use a reference prior  $p(\theta) \propto 1$  for all possible paths  $\theta$ , then all four possible values of  $\theta$  given  $X$  are equally likely. Note, however, that intuition would suggest that  $\theta$  is three times as likely to extend the path as to backtrack, suggesting that one particular value of  $\theta$  is more likely than the others after  $X$  is observed (due to M. Stone).

**5.** Suppose that, starting with a fortune of  $f_0$  units, you bet  $a$  units each time on evens at roulette (so that you have a probability of 18/37 of winning at Monte Carlo or 18/38 at Las Vegas) and keep a record of your fortune  $f_n$  and the difference  $d_n$  between the number of times you win and the number of times you lose in  $n$  games. Which of the following are stopping rules?

- a.** The last time  $n$  at which  $f_n \geq f_0$ .
- b.** The first time that you win in three successive games.
- c.** The value of  $n$  for which  $f_n = \max_{\{0 \leq k < \infty\}} f_k$ .

**6.** Suppose that  $x_1, \dots$  is a sequential sample from an  $N(\theta, 1)$  distribution and it is desired to test  $H_0 : \theta = \theta_0$  versus  $H_1 : \theta \neq \theta_0$ . The experimenter reports that

he used a proper stopping rule and obtained the data 3, -1, 2, 1.

a. What could a frequentist conclude?

b. What could a Bayesian conclude?

7. Let  $x_1, x_2, \dots$  be a sequential sample from a Poisson distribution  $P(\lambda)$ . Suppose that the stopping rule is to stop sampling at time  $n \geq 2$  with probability

$$\sum_{i=1}^{n-1} x_i > \sum_{i=1}^n x_i$$

for  $n = 2, 3, \dots$  (define  $0/0=1$ ). Suppose that the first five observations are 3, 1, 2, 5, 7 and that sampling then stops. Find the likelihood function for  $\lambda$  (Berger, 1985).

8. Show that the mean of the beta-Pascal distribution

$$p(S|R, r, s) = \binom{S}{s} \frac{B(r'' + s, R'' - r'' + S - s)}{B(r'' - 1, R'' - r'')}$$

is given by the formula in Section 7.3, namely,

$$\mathbb{E}S = (s + 1) \left( \frac{R'' - 2}{r'' - 2} \right) - 1.$$

9. Suppose that you intend to observe the number  $x$  of successes in  $n$  Bernoulli trials and the number  $y$  of failures before the  $n$ th success after the first  $n$  trials, so that  $x \sim B(n, \pi)$  and  $y \sim NB(n, \pi)$ . Find the likelihood function  $L(\pi|x, y)$  and deduce the reference prior that Jeffreys' rule would suggest for this case.

10. The negative of loss is sometimes referred to as *utility*. Consider a gambling game very unlike most in that you are bound to win at least £2, and accordingly in order to be allowed to play, you must pay an entry fee of £e. A coin is tossed until it comes up heads, and if this occurs for the first time on the  $n$ th toss, you receive £2<sup>n</sup>. Assuming that the utility to you of making a gain of £x is  $u(x)$ , find the expected utility of this game, and then discuss whether it is plausible that  $u(x)$  is directly proportional to  $x$ . [The gamble discussed here is known as the *St Petersburg Paradox*. A fuller discussion of it can be found in Leonard and Hsu (2001, Chapter 4).]

11. Suppose that you want to estimate the parameter  $\pi$  of a binomial distribution  $B(n, \pi)$ . Show that if the loss function is

$$\mathcal{L}(\theta, a) = (\theta - a)^2 / \{\theta(1 - \theta)\},$$

then the Bayes rule corresponding to a uniform [i.e.  $Be(1, 1)$ ] prior for  $\pi$  is given by  $d(x)=x/n$  for any  $x$  such that  $0 < x < n$ , that is, the maximum likelihood estimator. Is  $d(x)=x/n$  a Bayes rule if  $x = 0$  or  $x=n$ ?

**12.** Let  $x \sim B(n, \pi)$  and  $y \sim B(n, \rho)$  have independent binomial distributions of the same index but possibly different parameters. Find the Bayes rule corresponding to the loss

$$\mathcal{L}(\pi, \rho, a) = (\pi - \rho - a)^2$$

when the priors for  $\pi$  and  $\rho$  are independent uniform distributions.

**13.** Investigate possible point estimators for  $\pi$  on the basis of the posterior distribution in the example in the subsection of Section 2.10 headed ‘Mixtures of conjugate densities’.

**14.** Find the Bayes rule corresponding to the loss function

$$\mathcal{L}(\theta, a) = \begin{cases} u(a - \theta) & \text{if } a \leq \theta \\ v(\theta - a) & \text{if } a \geq \theta. \end{cases}$$

**15.** Suppose that your prior for the proportion  $\pi$  of defective items supplied by a manufacturer is given by the beta distribution  $Be(2, 12)$ , and that you then observe that none of a random sample of size 6 is defective. Find the posterior distribution and use it to carry out a test of the hypothesis  $H_0: \pi < 0.1$  using

a. a ‘0 – 1’ loss function, and

b. the loss function

$a \setminus \theta$	$\theta \in \Theta_0$	$\theta \in \Theta_1$
$a_0$	0	1
$a_1$	2	0

**16.** Suppose there is a loss function  $\mathcal{L}(\theta, a)$  defined by

$a \setminus \theta$	$\theta \in \Theta_0$	$\theta \in \Theta_1$
$a_0$	0	10
$a_1$	10	0
$a_2$	3	3.

On the basis of an observation  $x$  you have to take action  $a_0$ ,  $a_1$  or  $a_2$ . For what values of the posterior probabilities  $p_0$  and  $p_1$  of the hypotheses  $H_0: \theta \in \Theta_0$  and  $H_1: \theta \in \Theta_1$  would you take each of the possible actions?

**17.** A child is given an intelligence test. We assume that the test result  $x$  is  $N(\theta, 100)$  where  $\theta$  is the true intelligence quotient of the child, as measured by the test (in other words, if the child took a large number of similar tests, the average score would be  $\theta$ ). Assume also that, in the population as a whole,  $\theta$  is distributed according to an  $N(100, 225)$  distribution. If it is desired, on the basis of the intelligence quotient, to decide whether to put the child into a slow, average or fast group for reading, the actions available are:

a.  $a_1$ : Put in slow group, that is, decide  $\theta \in \Theta_1 = (0, 90)$

b.  $a_1$ : Put in average group, that is, decide  $\theta \in \Theta_2 = [90, 100]$

c.  $a_1$ : Put in fast group, that is, decide  $\theta \in \Theta_3 = (100, \infty)$ .

A loss function  $\mathcal{L}(\theta, a)$  of the following form might be deemed appropriate:

$a \setminus \theta$	$\theta \in \Theta_1$	$\theta \in \Theta_2$	$\theta \in \Theta_3$
$a_1$	0	$\theta - 90$	$2(\theta - 90)$
$a_2$	$90 - \theta$	0	$\theta - 110$
$a_3$	$2(110 - \theta)$	$110 - \theta$	0.

Assume that you observe that the test result  $x = 115$ . By using tables of the normal distribution and the fact that if  $\phi(t)$  is the density function of the standard normal distribution, then  $\int t\phi(t) dt = -\phi(t)$ , find is the appropriate action to take on the basis of this observation. [See Berger (1985, Sections 4.2–4.4)].

**18.** In Section 7.8, a point estimator  $\delta_n$  for the current value  $\lambda$  of the parameter of a Poisson distribution was found. Adapt the argument to deal with the case where the underlying distribution is geometric, that is

$$p(x|\pi) = \pi(1-\pi)^x.$$

Generalize to the case of a negative binomial distribution, that is,

$$p(x|\pi) = \binom{n+x-1}{x} \pi^n (1-\pi)^x.$$

# Hierarchical models

## 8.1 The idea of a hierarchical model

### 8.1.1 Definition

So far, we have assumed that we have a single known form to our prior distribution. Sometimes, however, we feel uncertain about the extent of our prior knowledge. In a typical case, we have a first stage in which observations  $x$  have a density  $p(x|\theta)$  which depends on  $r$  unknown parameters  $\theta = (\theta_1, \theta_2, \dots, \theta_r)$  for which we have a prior density  $p(\theta)$ . Quite often we make one or more assumptions about the relationships between the different parameters  $\theta_i$ , for example, that they are independently and identically distributed [sometimes abbreviated i.i.d.] or that they are in increasing order. Such relationships are often referred to as *structural*.

In some cases, the structural prior knowledge is combined with a standard form of Bayesian prior belief about the parameters of the structure. Thus, in the case where the  $\theta_i$  are independently and identically distributed, their common distribution might depend on a parameter  $\eta$ , which we often refer to as a *hyperparameter*. We are used to this situation in cases where  $\eta$  is known, but sometimes it is unknown. When it is unknown, we have a second stage in which we suppose that we have a *hyperprior*  $p(\eta)$  expressing our beliefs about possible values of  $\eta$ . In such a case, we say that we have a *hierarchical prior*; for the development of this idea, see Good (1980). It should be noted that the difficulty of specifying a second stage prior has made common the use of noninformative priors at the second stage (cf. Berger, 1985, Sections 3.6 and 4.6.1).

In Lindley's words in his contribution to Godambe and Sprott (1971),

The type of problem to be discussed ... is one in which there is a substantial amount of data whose probability structure depends on several parameters of the same type. For example, an agricultural trial involving many varieties, the parameters being the varietal means, or an educational test performed on

many subjects, with their true scores as the unknowns. In both these situations the parameters are related, in one case by the common circumstances of the trial, in the other by the test used, so that a Bayesian solution, which is capable of including such prior feelings of relationship, promises to show improvements over the usual techniques.

There are obvious generalizations. For one thing, we might have a vector  $\eta$  of hyperparameters rather than a single hyperparameter. For another, we sometimes carry this process to a third stage, supposing that the prior for  $p(\eta)$  depends on one or more hyper-hyperparameters  $\mu$  and so takes the form  $p(\eta|\mu)$ . If  $\mu$  is unknown, then we have a hyper-hyperprior density  $p(\mu)$  representing our beliefs about possible values of  $\mu$ .

All of this will become clearer when we consider some examples. Examples from various fields are given to emphasize the fact that hierarchical models arise in many different contexts.

## 8.1.2 Examples

### 8.1.2.1 Hierarchical Poisson model

In Section 7.8 on ‘Empirical Bayes methods’, we considered a case where we had observations  $x_i \sim P(\lambda_i)$  where the  $\lambda_i$  have a distribution with a density  $p(\lambda)$ , and then went on to specialize to the case where  $p(\lambda)$  was of the conjugate form  $S_0^{-1} \chi_\nu^2$  for some  $S_0$  and  $\nu$ . This is a structural relationship as defined earlier in which the parameters are  $\lambda = (\lambda_i)$  and the hyperparameters are  $\eta = (S_0, \nu)$ . To fit this situation into the hierarchical framework we only need to take a prior distribution for the hyperparameters. Since they are both in the range  $(0, \infty)$ , one possibility might be to take independent reference priors  $p(S_0) \propto 1/S_0$  and  $p(\nu) \propto 1/\nu$  or proper priors close to these over a large range.

### 8.1.2.2 Test scores

Suppose that a number of individuals take intelligence tests (‘IQ tests’) on which their scores are normally distributed with a known variance  $\phi$  but with a mean which depends on the ‘true abilities’ of the individuals concerned, so that  $x_i \sim N(\theta_i, \phi)$ . It may well happen that the individuals come from a population in which the true abilities are (at least to a reasonable approximation) normally distributed, so that  $\theta_i \sim N(\mu, \psi)$ . In this case the hyperparameters are  $\eta = (\mu, \psi)$ .

If informative priors are taken at this stage, a possible form would be to take  $\mu$  and  $\psi$  as independent with  $\mu \sim N(\lambda, \psi)$  and  $1/\psi \sim S_0^{-1} \chi_\nu^2$  for suitable values of the hyper-hyperparameters  $\lambda$ ,  $\psi$ ,  $S_0$  and  $\nu$ .

### 8.1.2.3 Baseball statistics

The *batting average* of a baseball player is defined as the number of hits  $S_i$  divided by the number of times at bat; it is always a number between 0 and 1. We will suppose that each of  $r$  players have been  $n$  times at bat and that the batting average of the  $i$ th player  $Y_i = S_i/n$  is such that  $S_i \sim B(n, \pi_i)$ , so that using the inverse root-sine transformation (see Section 3.2 on ‘Reference prior for the

$$X_i = 2\sqrt{n} \sin^{-1} \sqrt{Y_i}$$

binomial likelihood’), we see that if  $\theta_i = 2\sqrt{n} \sin^{-1} \sqrt{\pi_i}$ ,

then to a good approximation

$$X_i \sim N(\theta_i, 1).$$

We might then suppose that

$$\theta_i \sim N(\mu, \psi).$$

Finally, we suppose that  $\psi$  is *known* and that the prior knowledge of  $\mu$  is *weak*, so that over the range over which the likelihood is appreciable, the prior density of  $\mu$  is constant (cf. Section 2.5 on ‘Locally uniform priors’).

This example is considered by Efron and Morris (1975 and 1977); we give further consideration to it in Section 8.3. These authors also consider an example arising from data on the incidence of toxoplasmosis (a disease of the blood that is endemic in much of Central America) among samples of various sizes from 36 cities in El Salvador.

### 8.1.2.4 Poisson process with a change point

Suppose that we have observations  $x_i$  for  $i = 1, 2, \dots, n$  which represent the number of times a rare event has occurred in each of  $n$  equal time intervals and that we have reason to believe that the frequency of this event has changed abruptly from one level to another at some intermediate value of  $i$ . We might then be interested in deciding whether there really *is* evidence of such an abrupt change and, if so, then investigating when it took place.

To model this situation, we suppose that  $x_i \sim P(\lambda)$  for  $i = 1, 2, \dots, k$  while  $x_i \sim P(\mu)$  for  $i = k + 1, \dots, n$ . We then take independent priors for the parameters  $\theta = (\lambda, \mu, k)$  such that

**a.**  $k \sim UD(1, n)$ , that is,  $k$  has a discrete uniform distribution on  $[1, n]$ ; **b.**  $\lambda = U/\gamma$  where  $U$  has an exponential distribution of mean 1 (or equivalently a one-parameter gamma distribution with parameter 1, so that  $2U \sim \chi_2^2$ ); **c.**  $\mu = V/\delta$  where  $V$  is independent of  $U$  and has the same distribution.

These distributions depend on the two parameters  $\gamma$  and  $\delta$ , so that  $\eta = (\gamma, \delta)$  are hyperparameters.

Finally, we suppose that  $\gamma$  and  $\delta$  have independent prior distributions which are multiples of chi-squared, so that for suitable values of the parameters  $\xi$ ,  $\zeta$ ,  $\alpha$  and  $\beta$  we have  $p(\gamma) \propto \gamma^{\xi/2-1} \exp(-\alpha\gamma)$  and  $p(\delta) \propto \delta^{\zeta/2-1} \exp(-\beta\delta)$ .

This situation is a slightly simplified version of one described by Carlin *et al.* (1992), Tanner (1996, Sections 6.2.2 and 6.2.3) and Carlin and Louis (2008, Chapter 5, Exercises 8–10) as a model for the numbers of coal-mining disasters (defined as accidents resulting in the deaths of ten or more miners) in Great Britain for the years from 1851 to 1962 inclusive. We shall consider this example in detail in 9.4 on ‘The Gibbs sampler’.

### 8.1.2.5 Risk of tumour in a group of rats

In a study of tumours among laboratory rats of type ‘F344’, the probability of tumours in different groups of rats is believed to vary because of differences between rats and experimental conditions among the experiments. It may well be reasonable to suppose that the probabilities come from a beta distribution, but it is not clear *a priori* which prior beta distribution to take.

In this case, the number of rats  $y_i$  that develop tumours in the  $i$ th group which is of size  $n_i$  is such that  $y_i \sim B(n_i, \theta_i)$ , while  $\theta_i \sim Be(\alpha, \beta)$ . We then take some appropriate hyperprior distribution for the hyperparameters  $\eta = (\alpha, \beta)$ ; it has been suggested that a suitable noninformative hyperprior is  $p(\alpha, \beta) \propto (\alpha + \beta)^{-5/2}$ .

This example is discussed in detail and the above hyperprior is derived in Gelman *et al.* (2004, Sections 5.1 and 5.3).

### 8.1.2.6 Vaccination against Hepatitis B

In a study of the effect of vaccination for Hepatitis B in the Gambia, it was supposed that  $y_{ij}$ , the log anti-HB titre (the amount of surface-antibody in blood samples) in the  $j$ th observation for the  $i$ th infant, could be modelled as follows:

$$y_{ij} \sim N(\theta_{ij}, \phi)$$

$$\theta_{ij} = \alpha_i + \beta_i(\log t_{ij} - \log 730)$$

$$\alpha_i \sim N(\alpha_0, \phi_\alpha)$$

$$\beta_i \sim N(\beta_0, \phi_\beta)$$

(cf. Gilks *et al.*, 1993, or Spiegelhalter, *et al.*, 1996, Section 2.2). Here, the parameters are  $\theta = (\theta_{ij})$  and the hyperparameters are  $\eta = (\alpha, \beta, \phi)$ . The hyperprior distributions for  $\alpha$  and  $\beta$  are independent normals, and we take a reference prior for  $\phi$ . Further, we have hyper-hyperparameters  $(\alpha_0, \beta_0, \phi_\alpha, \phi_\beta)$  for which we take reference priors, so that  $p(\alpha_0, \beta_0, \phi_\alpha, \phi_\beta) \propto 1/\phi_\alpha \phi_\beta \quad (\phi_\alpha, \phi_\beta > 0)$ .

Actually, Gilks *et al.* take proper priors which over reasonable values of

$$\alpha_0 \sim N(0, 10,000) \quad \beta_0 \sim N(0, 10,000)$$

$$(\alpha_0, \beta_0, \phi_\alpha, \phi_\beta) \text{ and } \phi \text{ behave similarly, namely, } \phi_\alpha \sim G(0.01, 0.01) \quad \phi_\beta \sim G(0.01, 0.01).$$

### 8.1.3 Objectives of a hierarchical analysis

The objectives of hierarchical analyses vary. We can see this by considering the way in which the examples described in Subsection 8.1.2, headed ‘Examples’ might be analyzed.

In the case of the hierarchical Poisson model, the intention is to estimate the density function  $p(\lambda)$  or equivalently the hyperparameters  $\eta = (S_0, \nu)$ . Similarly, in the case of the example on rat tumours, the main interest lies in finding the joint posterior density of the hyperparameters  $\alpha$  and  $\beta$ .

In the cases of the test example and the baseball example, the interest lies in estimating the parameters  $\theta$  as well as possible, while the hyperparameters  $\mu$  and  $\psi$  are of interest mainly as tools for use in estimating  $\theta$ .

The main interest in the case of the Poisson process with a change point could quite well lie in determining whether there really *is* a change point, and, assuming that there is, finding out where it occurs as closely as possible.

However, the models could be explored with other objectives. For example, in the intelligence test example we might be interested in the predictive distribution  $p(x) = \int \int p(x|\theta)p(\theta|\eta) d\theta d\eta$ , which represents the overall distribution of ‘IQ’ in the population under consideration. Similarly, in the case of the Poisson distribution with a change point, we might be interested in the extent of the change (presuming that there is one), and hence in  $\lambda/\mu$ , that is, in the distribution of a *function* of the parameters.

### 8.1.4 More on empirical Bayes methods

The empirical Bayes method, of which a very short account was given in 7.8, is often employed in cases where we have a structural relationship as described at the start of this section. Suppose for definiteness that we have a straightforward two-stage model in which the density  $p(x|\theta)$  of the observations depends on parameters  $\theta = (\theta_i)$  which themselves are independent observations from a density  $p(\theta|\eta)$ , so that we have a posterior density  $p(x|\eta) = \int \dots \int p(x|\theta) p(\theta|\eta) d\theta_1 d\theta_2 \dots d\theta_n$ .

We then estimate  $\eta$  by the method of maximum likelihood or by some other method of classical statistics. Note that this method makes no use of a prior distribution for the hyperparameter  $\eta$ .

## 8.2 The hierarchical normal model

### 8.2.1 The model

Suppose that

$$\theta = (\theta_1, \dots, \theta_r)$$

is a vector of fixed, unknown parameters and that

$$X = (X_1, \dots, X_r)$$

is a vector of independent observations such that

$$X_i \sim N(\theta_i, \phi).$$

Of course, the  $X_i$  could each be means of a number of observations.

For the moment, we shall suppose that  $\phi$  is *known*, so, after a suitable normalization, we can suppose that  $\phi = 1$ .

It is useful to establish some notation for use later on. We shall consider a fixed origin

$$\mu = (\mu_1, \dots, \mu_r),$$

and we will write

$$\bar{X} = (X_1 + \dots + X_r)/r$$

$$S_0 = \sum_{i=1}^r X_i^2 \quad S_1 = \sum_{i=1}^r (X_i - \mu_i)^2 \quad S = \sum_{i=1}^r (X_i - \bar{X})^2$$

and

$$\mathbf{1} = (1, \dots, 1)$$

for a vector of  $r$  elements all equal to unity.

We suppose that on the basis of our knowledge of the  $X_i$  we form estimates  $\hat{\theta}_i$

of the  $\theta_i$  and write  $\widehat{\boldsymbol{\theta}} = (\widehat{\theta}_1, \widehat{\theta}_2, \dots, \widehat{\theta}_r)$ .

In general, our estimates will not be exactly right and we will adopt a decision theoretic approach as described in Section 7.5 on ‘Bayesian decision theory’. In particular, we shall suppose that by estimating the parameters we suffer a loss

$$\mathcal{L}(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}) = \frac{1}{r} \sum_{i=1}^r (\theta_i - \widehat{\theta}_i)^2.$$

We recall that the risk function is defined as

$$R(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}) = E \mathcal{L}(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}).$$

For our problem the ‘obvious’ estimator (ignoring the hierarchical structure which will be introduced later) is  $\widehat{\boldsymbol{\theta}}^0 = \mathbf{X}$

and indeed since the log-likelihood is

$$\text{constant} - \frac{1}{2} \sum_{i=1}^r (X_i - \theta_i)^2,$$

it is the maximum likelihood estimator. It is clearly unbiased.

It is easy to find the risk of this obvious estimator – it is

$$R(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}^0) = E \frac{1}{r} \sum_{i=1}^r (X_i - \theta_i)^2 = 1.$$

## 8.2.2 The Bayesian analysis for known overall mean

To express this situation in terms of a hierarchical model, we need to suppose that the parameters  $\boldsymbol{\theta} = (\theta_i)$  come from some population, and the simplest possibility is to suppose that  $\theta_i \sim N(\mu, \phi) \quad (i = 1, 2, \dots, r)$

in which case it is convenient to take  $\mu = \mu^1$ . With the additional structure assumed for the means, the problem has the structure of a situation variously described as a random effects model, Model II or a components of variance model (cf. Eisenhart *et al.*, 1947, or Scheffé, 1959, Section 7.2, n.7). We are, however, primarily interested in the means  $\theta_i$  and not in the variance components  $\phi$  and  $\psi$ , at least for the moment.

It follows that the posterior distribution of  $\theta_i$  given  $X_i \sim N(\theta_i, 1)$  is  $\theta_i | X_i \sim N(\theta_*, \phi_*)$ , where (writing  $\lambda = 1/(\phi + 1)$ )

$$\phi_* = \frac{1}{\phi^{-1} + 1} = 1 - \lambda,$$

and

$$\theta_* = \phi_*(\mu/\phi + X_i) = \lambda\mu + (1 - \lambda)X_i$$

(cf. Section 2.2 on ‘Normal Prior and Likelihood’).

To minimize the expectation  $\rho(\hat{\theta}, X)$  of the loss  $\mathcal{L}(\theta, \hat{\theta})$  over the posterior distribution of  $\theta$ , it is clearly necessary to use the Bayes estimator  $\hat{\theta} = \hat{\theta}^B$ , where

$$\hat{\theta}^B = \mathbb{E}(\theta|X) = \lambda\mu + (1 - \lambda)X,$$

the posterior mean of  $\theta$  given  $X$  (see the subsection of Section 7.5 on ‘Point estimators resulting from quadratic loss’). Further, if we do this, then the value

$$\rho(\hat{\theta}^B, X) = \frac{1}{r} \sum_{i=1}^r \mathcal{V}(\theta_i|X_i) = 1 - \lambda.$$

of this posterior expected loss is

It follows that the Bayes risk

$$r(\hat{\theta}^B) = \mathbb{E}\rho(\hat{\theta}^B, X)$$

(the expectation being taken over values of  $X$ ) is  $r(\hat{\theta}^B) = 1 - \lambda$ .

We note that if instead we use the maximum likelihood estimator  $\hat{\theta}^0 = X$ , then the posterior expected loss is increased by an amount  $\frac{1}{r} \sum_{i=1}^r [X_i - \mathbb{E}(\theta_i|X_i)]^2 = \lambda^2 \frac{1}{r} \sum_{i=1}^r (X_i - \mu)^2 = \lambda^2 \frac{S_1}{r}$ ,

which is always positive, so that

$$\rho(\hat{\theta}^0, X) = 1 - \lambda + \lambda^2 \frac{S_1}{r}.$$

Further, since the unconditional distribution of  $X_i$  is evidently  $N(\mu, \phi + 1)$ , so that  $S_1/(\phi + 1) \sim \chi_r^2$ , its expectation over repeated sampling (the Bayes risk) is  $r(\hat{\theta}^0) = 1 - \lambda + \lambda^2(\phi + 1) = 1$ .

This is, in fact, obvious since we can also write

$$r(\hat{\theta}) = \mathbb{E} R(\theta, \hat{\theta}),$$

where the expectation is over  $\theta$ , and since for the maximum likelihood estimator  $R(\theta, \hat{\theta}^0) = 1$  for all  $\theta$ , we have  $r(\hat{\theta}^0) = \mathbb{E}1 = 1$ .

We can thus see that use of the Bayes estimator  $\hat{\theta}^B$  always diminishes the posterior loss, and that the amount ‘saved’ by its use averages out at  $\lambda$  over repeated sampling.

### 8.2.3 The empirical Bayes approach

Typically, however, you will not know  $\phi$  (or equivalently  $\lambda$ ). In such a situation, you can attempt to estimate it from the data. Since the  $X_i$  have an unconditional distribution which is  $N(\mu, \phi + 1)$ , it is clear that  $S_1$  is a sufficient statistic for  $\phi$ , or

equivalently for  $\lambda$ , which is such that  $S_1/(\phi + 1) \sim \chi_r^2$  or  $S_1 \sim \frac{1}{\lambda} \chi_r^2$ ,

so that if we define

$$\hat{\lambda} = \frac{r-2}{S_1},$$

then using the probability density of a chi-squared distribution (as given in

$$\begin{aligned}\mathbb{E}\hat{\lambda} &= \int \frac{r-2}{S_1} \frac{\lambda^{r/2}}{2^{r/2}\Gamma(r/2)} S_1^{r/2-1} \exp(-\frac{1}{2}\lambda S_1) dS_1 \\ &= (r-2) \frac{\lambda^{r/2}}{2^{r/2}\Gamma(r/2)} \frac{2^{r/2-1}\Gamma(r/2-1)}{\lambda^{r/2-1}}\end{aligned}$$

Appendix A)  $= \lambda$ ,

so that  $\hat{\lambda}$  is an unbiased estimator of  $\lambda$ .

Now consider the effect of using the *empirical Bayes* estimator

$$\hat{\theta}^{EB} = \hat{\lambda}\mu\mathbf{1} + (1 - \hat{\lambda})X,$$

which results from replacing  $\lambda$  by  $\hat{\lambda}$  in the expression for  $\hat{\theta}^B = \mathbb{E}(\theta|X)$ . If we use this, then the value of the posterior expected loss exceeds that incurred by the

$$\begin{aligned}\frac{1}{r} \sum_{i=1}^r [\hat{\theta}_i^{EB} - \mathbb{E}(\theta_i|X_i)]^2 &= (\hat{\lambda} - \lambda)^2 \frac{1}{r} \sum_{i=1}^r (X_i - \mu)^2 \\ &= (\hat{\lambda} - \lambda)^2 \frac{S_1}{r},\end{aligned}$$

Bayes rule  $\hat{\theta}^B$  by an amount

which is always positive, so that

$$\rho(\hat{\theta}^{EB}, X) = 1 - \lambda + (\hat{\lambda} - \lambda)^2 \frac{S_1}{r}.$$

Further, if we write  $U = \lambda S_1$  (so that  $\hat{\lambda} = \lambda(r-2)/U$  and  $U \sim \chi_r^2$ ), then we see that the expectation of  $(\hat{\lambda} - \lambda)^2 S_1/r$  over repeated sampling is  $\frac{\lambda}{r} \left\{ \mathbb{E} \frac{(r-2)^2}{U} - 2(r-2) + \mathbb{E}U \right\} = \frac{\lambda}{r} \{(r-2) - 2(r-2) + r\} = \frac{2\lambda}{r}$ .

It follows that the Bayes risk resulting from the use of the empirical Bayes estimator is  $r(\hat{\theta}^{EB}) = 1 - \frac{r-2}{r}\lambda$

as opposed to  $1 - \lambda$  for the Bayes estimator or 1 for the maximum likelihood estimator.

## 8.3 The baseball example

Efron and Morris's example on baseball statistics was outlined in Section 8.1. As their primary data, they take the number of times hits  $S_i$  or equivalently the batting averages  $Y_i = S_i/n$  of  $r=18$  major league players as they were recorded after  $n=45$  times at bat in the 1970 season. These were, in fact, all the players who happened to have batted exactly 45 times the day the data were tabulated. If  $X_i$

and  $\theta_i$  are as in Section 8.1, so that approximately  $X_i \sim N(\theta_i, 1)$ , then we have a case of the hierarchical normal model. With the actual data, we have

$$r = 18, \quad \bar{Y} = 0.2654, \quad \bar{X} = 7.221, \quad S = 18.96,$$

and so with

$$\lambda = \frac{r - 3}{S} = 0.7910,$$

the empirical Bayes estimator for the  $\theta_i$  takes the form  $\hat{\theta}^{EB} = \lambda \bar{X} \mathbf{1} + (1 - \lambda) X$  so giving estimates

$$\hat{\pi}_i^{EB} = \sin^2 \left( \frac{\hat{\theta}_i^{EB}}{2\sqrt{n}} \right).$$

We can test how well an estimator performs by comparing it with the observed batting averages. We suppose that the  $i$ th player had  $T_i$  hits and was at bat  $m_i$  times, so that his batting average for the remainder of the season was  $p_i = T_i/m_i$ . If we write  $\delta_i = 2\sqrt{n} \sin^{-1} \sqrt{p_i}$ ,

we could consider a mean square error

$$\sum (\theta_i - \delta_i)^2 / r$$

or more directly

$$\sum (\hat{\pi}_i^{EB} - p_i)^2 / r.$$

In either case, it turns out that the empirical Bayes estimator appears to be about three and a half times better than the ‘obvious’ (maximum likelihood) estimator which ignores the hierarchical model and just estimates each  $\theta_i$  by the corresponding  $X_i$ . The original data and the resulting estimates are tabulated in [Table 8.1](#).

**Table 8.1** Data for the baseball example.

$i$	Name	$S_i$	$T_i$	$m_i$	$Y_i$	$\hat{\pi}_i^{EB}$	$p_i$	$X_i$	$\theta_i$
1	Clemente	18	127	367	0.400	0.290	0.346	9.186	8.438
2	F. Robinson	17	127	426	0.378	0.286	0.298	8.881	7.749
3	F. Howard	16	144	521	0.356	0.281	0.276	8.571	7.427
4	Johnstone	15	61	275	0.333	0.277	0.222	8.258	6.579
5	Berry	14	114	418	0.311	0.273	0.273	7.938	7.372
6	Spencer	14	126	466	0.311	0.273	0.270	7.938	7.337
7	Kessinger	13	154	586	0.289	0.268	0.263	7.613	7.221
8	L. Alvarado	12	29	138	0.267	0.264	0.210	7.280	6.389
9	Santo	11	137	510	0.244	0.259	0.269	6.938	7.310
10	Swoboda	11	46	200	0.244	0.259	0.230	6.938	6.711
11	Unser	10	73	277	0.222	0.254	0.264	6.586	7.233
12	Williams	10	69	270	0.222	0.254	0.256	6.586	7.111
13	Scott	10	132	435	0.222	0.254	0.303	6.586	7.827
14	Petrocelli	10	142	538	0.222	0.254	0.264	6.586	7.239
15	E Rodriguez	10	42	186	0.222	0.254	0.226	6.586	6.644
16	Campaneris	9	159	558	0.200	0.249	0.285	6.220	7.555
17	Munson	8	129	408	0.178	0.244	0.316	5.839	8.012
18	Alvis	7	14	70	0.156	0.239	0.200	5.439	6.220

So there is evidence that in at least some practical case, use of the hierarchical model and a corresponding empirical Bayes estimator is genuinely worth while.

## 8.4 The Stein estimator

This section is about an aspect of classical statistics which is related to the aforementioned discussion, but an understanding of it is by no means necessary for developing a knowledge of Bayesian statistics per se. The Bayesian analysis of the hierarchical normal model is continued in Section 8.5.

One of the most puzzling and provocative results in classical statistics in the past half century was Stein's startling discovery (see Stein, 1956, and James and Stein, 1961) that the ‘obvious’ estimator  $\hat{\theta}^0$  of the multivariate normal mean is inadmissible if  $r \geq 3$ . In fact if  $c$  is any constant with  $0 < c < 2(r - 2)$ , then

$$\mu + \left(1 - \frac{c}{S_1}\right)(X - \mu)$$

dominates  $\hat{\theta}^0$ . The best value of  $c$  is  $r-2$ , leading to the *James–Stein estimator*  $\hat{\theta}^{JS} = \mu + \left(1 - \frac{r-2}{S_1}\right)(X - \mu)$ .

Because it may be considered as a weighted mean of  $\mu$  and  $X$ , it is often called a *shrinkage estimator* which ‘shrinks’ the ordinary estimator  $\hat{\theta}^0$  towards  $\mu$ , despite the fact that if  $S_1 < r-2$  it ‘shrinks’ past  $\mu$ . Note, incidentally, that points

which are initially far from  $\mu$  are little affected by this shrinkage. Of course, this ties in with the results of Section 8.3 because the James–Stein estimator  $\hat{\theta}^{JS}$  has turned out to be just the same as the empirical Bayes estimator  $\hat{\theta}^{EB}$ .

In fact, it can be shown that the risk of  $\hat{\theta}^{JS}$  is

$$R(\theta, \hat{\theta}^{JS}) = 1 - \frac{r-2}{r} E\left(\frac{r-2}{S_1}\right).$$

The expectation on the right-hand side depends on  $\theta$  and  $\mu$ , but as  $S_1 \geq 0$  it must be non-negative, so that  $\hat{\theta}^{JS}$  dominates  $\hat{\theta}^0$ , that is,  $R(\theta, \hat{\theta}^{JS}) < R(\theta, \hat{\theta}^0) = 1$

for all  $\theta$ . It turns out that  $S_1$  has a distribution which depends solely on  $r$  and the

$$\text{quantity } \gamma = \sum_{i=1}^r (\theta_i - \mu_i)^2,$$

a fact which can be proved by considering an orthogonal transformation of the

$$\text{variates } X_i - \mu_i \text{ to variates } W_i \text{ such that } W_1 = \frac{1}{\sqrt{\gamma}} \sum_{i=1}^r (\theta_i - \mu_i)(X_i - \mu_i).$$

Evidently if  $\gamma = 0$  then  $S_1 \sim \chi_r^2$ , and in general we say that  $S_1$  has a *non-central chi-squared distribution on r degrees of freedom with non-centrality parameter γ*. We denote this by  $S_1 \sim \chi_r'^2(\gamma)$ .

It is fairly obvious that as  $\gamma \rightarrow \infty$  typical values of  $S_1 = \sum(X_i - \mu_i)^2$  will tend to infinity and we will get  $R(\theta, \hat{\theta}^{JS}) = 1$

whereas when  $\gamma = 0$  the variate  $S_1$  has a central  $\chi^2$  distribution on  $r$  degrees of freedom (no parameters are estimated), so  $E\frac{1}{S_1} = \frac{1}{r-2}$ ,

and hence

$$R(\theta, \hat{\theta}^{JS}) = \frac{2}{r}$$

which, particularly for large values of  $r$ , is notably less than the risk of the obvious estimator.

In the particular case where the arbitrary origin is taken at  $\mathbf{0}$  the James–Stein estimator takes the form  $\hat{\theta}^{JS} = \left(1 - \frac{r-2}{S_0}\right)X$

but it is important to note that this is only a special case.

Variants of the James–Stein estimator have been derived. For example, if  $c$  is any constant with  $0 < c < 2(r-3)$ ,

then

$$\bar{X}\mathbf{1} + \left(1 - \frac{c}{S}\right)(X - \bar{X}\mathbf{1})$$

dominates  $\theta_0$ , this time provided  $r \geq 4$  (loss of one dimension as a result of

estimating a mean is something we are used to in statistics). The best value of  $c$  in this case is  $k-3$ , leading to the *Efron–Morris estimator*  
 $\hat{\theta}^{EM} = \bar{X}\mathbf{1} + \left(1 - \frac{r-3}{S}\right)(\mathbf{X} - \bar{X}\mathbf{1}).$

In this case the ‘shrinkage’ is towards the overall mean.

In the case of the Efron–Morris estimator, it can be shown (see Lehmann, 1983, Section 4.6) that the risk of  $\hat{\theta}^{EM}$  is  $R(\theta, \hat{\theta}^{EM}) = 1 - \frac{r-3}{r} \mathbb{E}\left(\frac{r-3}{S}\right).$

Since  $S$  has a central  $\chi^2$  distribution on  $r-1$  degrees of freedom,  $\mathbb{E}\frac{1}{S} = \frac{1}{r-3}$ , and hence

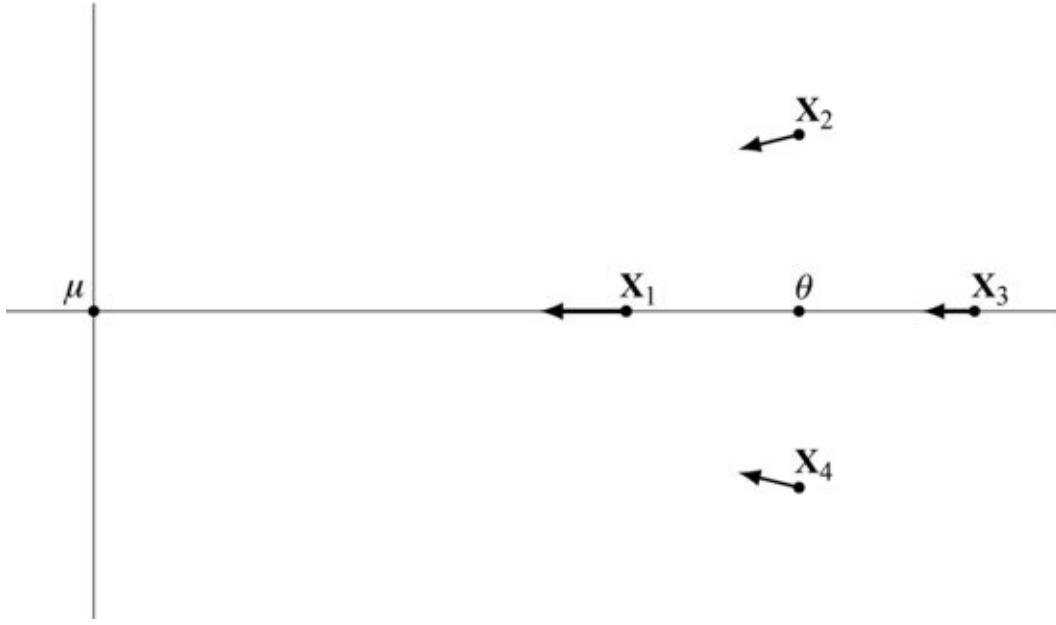
$$R(\theta, \hat{\theta}^{EM}) = \frac{3}{r}$$

which, particularly for large values of  $r$ , is again notably less than the risk of the obvious estimator.

When we consider using such estimates in practice we encounter the ‘speed of light’ rhetorical question, Do you mean that if I want to estimate tea consumption in Taiwan, I will do better to estimate simultaneously the speed of light and the weight of hogs in Montana?

The question then arises as to why this happens. Stein’s own explanation was that the sample distance squared of  $\mathbf{X}$  from  $\mu$ , that is  $\sum(X_i - \mu_i)^2$ , overestimates the squared distance of  $\theta$  from  $\mu$  and hence that the estimator  $\mathbf{X}$  could be improved by bringing it nearer  $\mu$  (whatever  $\mu$  is). Following an idea due to Lawrence Brown, the effect was illustrated as shown in [Figure 8.1](#) in a paper by Berger (1980, Figure 2, p. 736).

[Figure 8.1](#) Shrinkage estimators.



The four points  $X_1, X_2, X_3$  and  $X_4$  represent a spherical distribution centred at  $\theta$ .

Consider the effect of shrinking these points as shown. The points  $X_1$  and  $X_3$  move, *on average*, slightly further away from  $\theta$ , but the points  $X_2$  and  $X_4$  move slightly closer (while distant points hardly move at all). In three dimensions, there are a further two points (not on the line between  $\theta$  and  $\mu$ ) that are shrunk closer to  $\theta$ .

Another explanation that has been offered is that  $\hat{\theta}^{JS}$  can be viewed as a ‘pre-test’ estimator: if one performs a preliminary test of the hypothesis that  $\theta = \mu$  and then uses  $\hat{\theta} = \mu$  or  $\hat{\theta} = X$  depending on the outcome of the test, then the resulting estimator is a weighted average of  $\mu$  and  $X$  of which  $\hat{\theta}^{JS}$  is a smoothed version, although why this particular smoothing is to be used is not obvious from this chain of reasoning (cf. Lehmann, 1983, Section 4.5).

### 8.4.1 Evaluation of the risk of the James–Stein estimator

We can prove that the James–Stein estimator has the risk quoted earlier, namely

$$R(\theta, \hat{\theta}^{JS}) = 1 - \frac{r-2}{r} \mathbb{E} \left( \frac{r-2}{S_1} \middle| \theta \right).$$

[An alternative approach can be found in Lehmann (1983, Sections 4.5 and 4.6).]

$$\gamma = \sum_{i=1}^r (\theta_i - \mu_i)^2$$

$$g(\gamma) = R(\theta, \widehat{\theta}^{JS})$$

$$h(\gamma) = 1 - \frac{r-2}{r} \mathbb{E} \left( \frac{r-2}{S_1} \middle| \theta \right),$$

We proceed by writing

where the expectations are over repeated sampling for fixed  $\theta$ . The function  $g$  depends on  $\gamma$  alone by spherical symmetry about  $\mu$ . Similarly, the function  $h$  depends on  $\gamma$  alone since  $S_1 \sim \chi_r'^2(\gamma)$ . We note that because the unconditional distribution of  $S_1$  is  $\lambda^{-1} \chi_r^2$ , we have  $\mathbb{E}h(\gamma) = 1 - \frac{r-2}{r}\lambda$ ,

the expectation being taken over values of  $\theta$  or over values of  $\gamma$ , that is,  $\mathbb{E}g(\gamma) = \mathbb{E}R(\theta, \widehat{\theta}^{JS}) = r(\widehat{\theta}^{JS}) = \mathbb{E}h(\gamma)$

using the result at the very end of Section 8.2 and bearing in mind that  $\widehat{\theta}^{JS} = \widehat{\theta}^{EB}$ . Now writing  $k = g - h$  we have  $\mathbb{E}k(\gamma) = 0$ ,

and hence

$$\int_0^\infty k(\gamma) \gamma^{r/2-1} \exp\left(-\frac{1}{2}\frac{\gamma}{\phi}\right) d\gamma = 0$$

for all  $\phi$ , which can happen only if  $k$  vanishes identically by the uniqueness of Laplace transforms.

## 8.5 Bayesian analysis for an unknown overall mean

In Section 8.2, we derived the posterior for  $\theta$  supposing that a priori  $\theta_i \sim N(\mu, \phi)$  ( $i = 1, 2, \dots, r$ ),

where  $\mu$  was known. We shall now go on to an approach introduced by Lindley (1969) and developed in his contribution to Godambe and Sprott (1971) and in Lindley and Smith (1972) for the case where  $\mu$  is unknown.

We suppose that

$$x_{ij} \sim N(\theta_i, \phi) \quad (i = 1, 2, \dots, r; \quad j = 1, 2, \dots, n_i)$$

are independent given the  $\theta_i$  and  $\phi$ . This is the situation which arises in one way analysis of variance (analysis of variance between and within groups). In either of the practical circumstances described above, the means  $\theta_i$  will be thought to be alike. More specifically, the joint distribution of these means must have the property referred to by de Finetti (1937 or 1974–1975, Section 11.4) as *exchangeability*; that is, the joint distribution remains invariant under any

permutation of the suffices. A famous result in de Finetti (1937) [for a good outline treatment see Bernardo and Smith (1994, Sections 4.2 and 4.3)] says that exchangeability implies that the  $\theta_i$  have the probability structure of a random sample from a distribution. It might seem sensible to add the additional assumption that this distribution is normal (as we often do in statistics). It would then be appropriate to assume that  $\theta_i \sim N(\mu, \psi)$ ,  
the  $\theta_i$  being assumed independent for given  $\mu$  and  $\psi$ .

To complete the specification of the prior distribution, it is necessary to discuss  $\mu$ ,  $\phi$  and  $\psi$ . For the moment, we shall suppose that the two variances are *known* and that the prior knowledge of  $\mu$  is *weak*, so that, over the range for which the likelihood is appreciable, the prior density of  $\mu$  is constant (cf. Section 2.5).

We thus have

$$p(x|\theta) \propto \exp\left(-\frac{1}{2} \sum_i \sum_j (x_{ij} - \theta_i)^2 / \phi\right),$$

$$p(\theta|\mu) \propto \exp\left(-\frac{1}{2} \sum_i (\theta_i - \mu)^2 / \psi\right),$$

$$p(\mu) \propto 1,$$

so that

$$p(x, \theta, \mu) \propto \exp\left(-\frac{1}{2} \sum_i \sum_j (x_{ij} - \theta_i)^2 / \phi - \frac{1}{2} \sum_i (\theta_i - \mu)^2 / \psi\right).$$

We shall show that we can write the posterior distribution in the form

$$p(\theta|x) \propto \exp\left(-\frac{1}{2} \sum_i (a_i + b)(\theta_i - t_i)^2 - \sum_i \sum_{j \neq i} b(\theta_i - t_i)(\theta_j - t_j)\right),$$

where the  $t_i$  are defined by

$$t_j = w_j x_j + (1 - w_j) \bar{x}$$

in which

$$w_j = \frac{n_j \psi}{n_j \psi + \phi} = \frac{n_j / \phi}{n_j / \phi + 1 / \psi}.$$

We thus see that the posterior means of the  $\theta_j$  take the form of a weighted average of the mean  $x_j$  (the least-squares estimate) and an overall mean  $\bar{x}$ , depending in a natural way on the sample sizes and the ratio of the two variance components. The effect of this weighted average is to shift all the estimates for the sample mean towards the overall mean. It is clear that these estimates are of the same type as the Efron–Morris (or Stein) estimators derived earlier.

The proof of this result is given in Section 8.6, but can be omitted by readers willing to take the result for granted. It must be admitted that the result is mainly of theoretical interest because it is difficult to think of real-life cases where both  $\phi$  and  $\psi$  are known.

In the case where  $\phi$  and  $\psi$  are unknown and conjugate (inverse chi-squared) priors are taken for them, somewhat similar results are possible with  $\phi$  and  $\psi$  in the expression for  $w_i$  replaced by suitable estimators; the details can be found in Lindley's contribution to Godambe and Sprott (1971). Unfortunately, while it is possible to use a reference prior  $p(\phi) \propto 1/\phi$  for  $\phi$ , there are severe difficulties about using a similar prior for  $\psi$ . In the words of Lindley, *op. cit.*, The difficulty can be viewed mathematically by remarking that if a prior proportional to  $1/\psi$  ... which is improper ... – is used, then the posterior remains improper whatever size of sample is taken. Heuristically it can be seen that the between-sample variance provides information directly about  $\psi + \phi/n_i$ , – that is, confounded with  $\phi$  – and not about  $\psi$  itself, so that the extreme form of the prior cannot be overcome by sampling.

We shall discuss numerical methods for use in connection with the hierarchical normal model in Sections 9.2 and 9.4.

### 8.5.1 Derivation of the posterior

Because

$$\begin{aligned} \sum_i \sum_j (x_{ij} - \theta_i)^2 &= \sum_i n_i (x_{i*} - \theta_i)^2 + \sum_i \sum_j (x_{ij} - x_{i*})^2 \\ &= \sum_i n_i (x_{i*} - \theta_i)^2 + \text{constant} \end{aligned}$$

where  $x_{i*} = \sum_j x_{ij}/n_i$ , we see that

$$\begin{aligned} p(\theta, \mu | x) &\propto p(x, \theta, \mu) \\ &\propto \exp\left(-\frac{1}{2} \sum_i n_i (x_{i*} - \theta_i)^2 / \phi - \frac{1}{2} \sum_i (\theta_i - \mu)^2 / \psi\right). \end{aligned}$$

Noting that (because  $\sum_i (\theta_i - \bar{\theta}) = 0$ )

$$\begin{aligned} \exp\left(-\frac{1}{2} \sum_i (\theta_i - \mu)^2 / \psi\right) &= \exp\left(-\frac{1}{2} \sum_i \{(\theta_i - \bar{\theta}) - (\mu - \bar{\theta})\}^2 / \psi\right) \\ &\propto \exp\left(-\frac{1}{2} \sum_i (\theta_i - \bar{\theta})^2 / \psi - \frac{1}{2} r(\mu - \bar{\theta})^2 / \psi\right), \end{aligned}$$

we can integrate over  $\mu$  to get

$$\begin{aligned}
p(\theta|x) &\propto \exp\left(-\frac{1}{2}\sum_i n_i(x_{i*} - \theta_i)^2/\phi - \frac{1}{2}\sum_i (\theta_i - \bar{\theta})^2/\psi\right) \\
&\propto \exp\left(-\frac{1}{2}\sum_i n_i(x_{i*} - \theta_i)^2/\phi - \frac{1}{2}\sum_i \theta_i^2/\psi + \frac{1}{2}r\bar{\theta}^2/\psi\right).
\end{aligned}$$

Minus twice the coefficient of  $\theta_i^2$  in the above exponential is  $\frac{n_i}{\phi} + \left(1 - \frac{1}{r}\right)\frac{1}{\psi}$ , while the coefficient of  $\theta_i\theta_j$  is

$$\frac{1}{r}\frac{1}{\psi}$$

from which it follows that if we set

$$\begin{aligned}
a_i &= \frac{n_i}{\phi} + \frac{1}{\psi} = \frac{n_i\psi + \phi}{\phi\psi}, \\
b &= -\frac{1}{r\psi}
\end{aligned}$$

we can write the posterior distribution in the form

$$p(\theta|x) \propto \exp\left(-\frac{1}{2}\sum_i(a_i + b)(\theta_i - t_i)^2 - \sum\sum_{i \neq j} b(\theta_i - t_i)(\theta_j - t_j)\right),$$

where the  $t_i$  are yet to be determined.

By equating coefficients of  $\theta_j$  we see that

$$\begin{aligned}
\frac{n_j x_{j*}}{\phi} &= (a_j + b)t_j + b \sum_{i \neq j} t_i = a_j + br\bar{t} \\
&= \left\{ \frac{n_j\psi + \phi}{\phi\psi} \right\} t_j - \frac{1}{\psi}\bar{t}
\end{aligned}$$

where

$$\bar{t} = \frac{\sum_j t_j}{r}.$$

Writing

$$w_j = \frac{n_j\psi}{n_j\psi + \phi} = \frac{n_j/\phi}{n_j/\phi + 1/\psi},$$

it follows that

$$\begin{aligned}
t_j &= w_j x_{j*} + (1 - w_j)\bar{t} \\
r\bar{t} &= \sum_j w_j x_{j*} + \left(r - \sum_j w_j\right)\bar{t},
\end{aligned}$$

so that

$$\bar{t} = \bar{x}$$

where

$$\bar{x} = \frac{\sum_j w_j x_j}{\sum_j w_j}$$

and so

$$t_j = w_j x_j + (1 - w_j) \bar{x}.$$

We have thus proved that the posterior means of the  $\theta_j$  do indeed take the form of a weighted average of the mean  $x_j$ . (the least-squares estimate) and an overall mean  $\bar{x}$ , depending in a natural way on the sample sizes and the ratio of the two variance components and so shift all the estimates for the sample mean towards the overall mean.

A further discussion of related matters can be found in Leonard and Hsu (2001, Section 6.3).

## 8.6 The general linear model revisited

### 8.6.1 An informative prior for the general linear model

This section follows on from Section 6.7 on ‘The general linear model’ and like that section presumes a knowledge of matrix theory.

We suppose as in that section that

$$p(x|\theta) = (2\pi\phi)^{-n/2} \exp\left\{-\frac{1}{2}(x - A\theta)^T(x - A\theta)/\phi\right\}$$

(where  $\theta$  is  $r$ -dimensional, so  $A$  is  $n \times r$ ), but this time we take a non-trivial prior for  $\theta$ , namely  $p(\theta|\mu) = (2\pi\psi)^{-r/2} \exp\left\{-\frac{1}{2}(\theta - B\mu)^T(\theta - B\mu)/\psi\right\}$

(where  $\mu$  is  $s$ -dimensional, so  $B$  is  $r \times s$ ). If the hyperparameters are known, we may as well take  $r=s$  and  $B = I$ , and in practice dispense with  $B$ , but although for the moment we assume that  $\mu$  is known, in due course we shall let  $\mu$  have a distribution, and it will then be useful to allow other values for  $B$ .

Assuming that  $\mu$ ,  $\phi$  and  $\psi$  are known, the log of the posterior density is (up to

$$\begin{aligned} \log p(\theta|x) &= -\frac{1}{2} \frac{(x - A\theta)^T(x - A\theta)}{\phi} - \frac{1}{2} \frac{(\theta - B\mu)^T(\theta - B\mu)}{\psi} \\ &= -\frac{1}{2} \frac{x^T x}{\phi} + \frac{\theta^T A^T x}{\phi} - \frac{1}{2} \frac{\theta^T A^T A \theta}{\phi} \\ &\quad - \frac{1}{2} \frac{\theta^T \theta}{\psi} + \frac{\theta^T B \mu}{\psi} - \frac{1}{2} \frac{\mu^T B^T B \mu}{\psi}. \end{aligned}$$

an additive constant)

Differentiating with respect to the components of  $\theta$ , we get a set of equations

$$\frac{\partial L}{\partial \theta} = \frac{A^T x}{\phi} - \frac{A^T A \theta}{\phi} + \frac{B \mu}{\psi} - \frac{\theta}{\psi}$$

which can be written as one vector equation

Equating this to zero to find the mode of the posterior distribution, which by symmetry equals its mean, we get  $(A^T A + \frac{\phi}{\psi} I) \hat{\theta} = A^T x + \frac{\phi}{\psi} B \mu$ , so that

$$\hat{\theta} = (A^T A + k I)^{-1} (A^T x + k B \mu), \quad \text{where } k = \frac{\phi}{\psi}.$$

In particular, if  $\mu$  is taken as zero, so that the vector of prior means vanishes, then this takes the form  $\hat{\theta} = (A^T A + k I)^{-1} A^T x$ .

The usual least squares estimators reappear if  $\psi = \infty$ .

## 8.6.2 Ridge regression

This result is related to a technique which has become popular in recent years among classical statisticians which is known as *ridge regression*. This was originally developed by Hoerl and Kennard (1970), and a good account of it can be found in the article entitled ‘Ridge Regression’ in Kotz *et al.* (2006); alternatively, see Weisberg (2005, Section 11.2). Some further remarks about the connection between ridge regression and Bayesian analysis can be found in Rubin (1988).

What they pointed out was that the appropriate (least squares) point estimator for  $\theta$  was  $\hat{\theta} = (A^T A)^{-1} A^T x$ .

From a classical standpoint, it then matters to find the variance–covariance matrix of this estimator in repeated sampling, which is easily shown to be  $\mathcal{V}(\hat{\theta}) = [(A^T A)^{-1} A^T]^T \mathcal{V}(x) (A^T A)^{-1} A^T = \phi (A^T A)^{-1}$

(since  $\mathcal{V}x = \phi I$ ), so that the sum of the variances of the regression coefficients  $\theta_j$  is  $\text{Trace}(\mathcal{V}(\hat{\theta})) = \phi \text{Trace}\{(A^T A)^{-1}\}$

(the trace of a matrix being defined as the sum of the elements down its main diagonal) and the mean square error in estimating  $\theta$  is  $MSE = \phi \text{Trace}\{(A^T A)^{-1}\}$

However, there can be considerable problems in carrying out this analysis. It has been found that the least squares estimates are sometimes inflated in magnitude, sometimes have the wrong sign, and are sometimes unstable in that radical changes to their values can result from small changes or additions to the data. Evidently if  $\text{Trace}\{(A^T A)^{-1}\}$  is large, so is the mean-square error, which we can summarize by saying that the poorer the conditioning of the  $A^T A$  matrix, the worse the deficiencies referred to above are likely to be. The suggestion of Hoerl

and Kennard was to add small positive quantities to the main diagonal, that is to replace  $A^T A$  by  $A^T A + kI$  where  $k > 0$ , so obtaining the estimator  $\hat{\theta} = (A^T A + kI)^{-1} A^T x$

which we derived earlier from a Bayesian standpoint. On the other hand, Hoerl and Kennard have some rather *ad hoc* mechanisms for deciding on a suitable value for  $k$ .

### 8.6.3 A further stage to the general linear model

We now explore a genuinely hierarchical model. We supposed that  $x \sim N(A\theta, \phi I)$ , or slightly more generally that  $x \sim N(A\theta, \Phi)$

(see the description of the multivariate normal distribution in Appendix A). Further a priori  $\theta \sim N(B\mu, \psi I)$ , or slightly more generally  $\theta \sim N(B\mu, \Psi)$ .

At the next stage, we can suppose that our knowledge of  $\mu$  is vague, so that  $p(\mu) \propto 1$ . We can then find the marginal density of  $\theta$  as

$$p(\theta) \propto \int p(\mu) p(\theta|\mu) d\mu$$

$$\begin{aligned} &= \int \exp \left\{ -\frac{1}{2}(\theta - B\mu)^T \Psi^{-1} (\theta - B\mu) \right\} d\mu \\ &= \int \exp \left\{ -\frac{1}{2}\theta^T \Psi^{-1} \theta - \frac{1}{2}\mu^T B^T \Psi^{-1} B\mu + \mu^T B^T \Psi^{-1} \theta \right\} d\mu \\ &= \int \exp \left\{ -\frac{1}{2}\theta^T \Psi^{-1} \theta + \frac{1}{2}\mu_0^T B^T \Psi^{-1} B\mu_0 \right\} \\ &\quad \times \exp \left\{ -\frac{1}{2}(\mu - \mu_0)^T B^T \Psi^{-1} B(\mu - \mu_0) \right\} d\mu \end{aligned}$$

on completing the square by taking  $\mu_0$  such that  $B^T \Psi^{-1} \theta = B^T \Psi^{-1} B\mu_0$ , that is

$$\begin{aligned} \mu_0 &= (B^T \Psi^{-1} B)^{-1} B^T \Psi^{-1} \theta, \\ \mu_0^T B^T \Psi^{-1} B\mu_0 &= \theta^T \Psi^{-1} B (B^T \Psi^{-1} B)^{-1} B^T \Psi^{-1} \theta. \end{aligned}$$

Since the second exponential is proportional to a normal density, it integrates to a constant and we can deduce that  $p(\theta) \propto \exp \left\{ -\frac{1}{2}\theta^T H^{-1} \theta \right\}$ ,

that is  $\theta \sim N(0, H)$ , where

$$H^{-1} = \Psi^{-1} - \Psi^{-1} B (B^T \Psi^{-1} B)^{-1} B^T \Psi^{-1}.$$

We can then find the posterior distribution of  $\theta$  given  $x$  as

$$p(\theta|x) \propto p(\theta) p(x|\theta),$$

$$\begin{aligned} &\propto \exp \left\{ -\frac{1}{2}(x - A\theta)^T \Phi^{-1} (x - A\theta) - \frac{1}{2}\theta^T H^{-1} \theta \right\} \\ &\propto \exp \left\{ -\frac{1}{2}\theta^T (A^T \Phi^{-1} A + H^{-1}) \theta + \theta^T A^T x \right\}. \end{aligned}$$

Again completing the square, it is easily seen that this posterior distribution is

$$\mathbf{K}^{-1} = \mathbf{A}^T \Phi^{-1} \mathbf{A} + \mathbf{H}^{-1},$$

$N(\nu, K)$  where  $\nu = \mathbf{K} \mathbf{A}^T \mathbf{x}$ .

## 8.6.4 The one way model

If we take the formulation of the general linear model much as we discussed it in

$$\mathbf{x} = \begin{pmatrix} x_{11} \\ x_{12} \\ \vdots \\ x_{1n_1} \\ x_{21} \\ x_{22} \\ \vdots \\ x_{rn_r} \end{pmatrix}; \quad \mathbf{A} = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 1 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix},$$

Section 6.7, so that

we note that  $\mathbf{A}^T \mathbf{A} = \text{diag}\{n_1, \dots, n_r\}$ . We assume that the  $x_i$  are independent and have variance  $\phi$ , so that  $\Phi$  reduces to  $\phi I$  and hence  $\mathbf{A}^T \Phi^{-1} \mathbf{A} = \phi^{-1} \mathbf{A}^T \mathbf{A}$ . The situation where we assume that the  $\theta_j$  are independently  $N(\mu, \psi)$  fits into this situation if we take  $B = 1$  (an  $r$ -dimensional column vector of 1s, so that  $1^T 1 = r$  while  $11^T$  is an  $r \times r$  matrix with 1s everywhere) and have just one scalar hyperparameter  $\mu$  of which we have vague prior knowledge. Then  $\Psi$  reduces to  $\psi I$  and  $(B^T \Psi^{-1} B)^{-1}$  to  $\psi/r$  giving

$$\mathbf{H}^{-1} = \Psi^{-1} - \Psi^{-1} B (B^T \Psi^{-1} B)^{-1} B^T \Psi^{-1} = (1/\psi) I - (1/r\psi) 11^T$$

$$\mathbf{K}^{-1} = \mathbf{A}^T \Phi^{-1} \mathbf{A} + \mathbf{H}^{-1} = \text{diag}\{n_1/\phi + 1/\psi, \dots, n_r/\phi + 1/\psi\} - (1/r\psi) 11^T,$$

and so  $\mathbf{K}^{-1}$  has diagonal elements  $a_i + b$  and all off-diagonal elements equal to  $b$ ,

$$\text{where } a_i = \frac{n_i}{\phi} + \frac{1}{\psi}, \quad b = -\frac{1}{r\psi}.$$

These are of course the same values we found in Section 8.5 earlier. It is, of course, also possible to deduce the form of the posterior means found there from the approach used here.

## 8.6.5 Posterior variances of the estimators

Writing

$$\mathbf{D} = \text{diag}\{\sqrt{a_1}, \sqrt{a_2}, \dots, \sqrt{a_m}\}$$

and  $u = \sqrt{(-b)} \mathbf{D}^{-1} 1$  (remember that  $b < 0$ ) it is easily seen that  $\Sigma^{-1} = \mathbf{D}(I - uu^T)\mathbf{D}$ , and hence

$$\boldsymbol{\Sigma} = \mathbf{D}^{-1}(\mathbf{I} - \mathbf{u}\mathbf{u}^T)^{-1}\mathbf{D}^{-1} = \mathbf{D}^{-1} \left( \mathbf{I} + \frac{1}{1 + \mathbf{u}^T \mathbf{u}} \mathbf{u}\mathbf{u}^T \right) \mathbf{D}^{-1}$$

using the Sherman–Morrison formula for the inverse of  $\mathbf{I} + \mathbf{v}\mathbf{v}^T$  with  $\mathbf{v} = -\mathbf{u}$ . [This result is easily established; in case of difficulty, refer to Miller (1987, Section 3) or Horn and Johnson (1991, Section 0.7.4).] Consequently the

posterior variance of  $\theta_i$  is  $\Sigma_{ii} = \frac{1}{a_i} \left( 1 - \frac{1}{1 + \mathbf{u}^T \mathbf{u}} \frac{b}{a_i} \right)$ .

Now substituting  $a_i = n_i/\phi + 1/\psi$  and  $b = -1/r\psi$ , we see that

$$1 - \frac{1}{1 + \mathbf{u}^T \mathbf{u}} \frac{b}{a_i} \leq 1 + \frac{1/r\psi}{a_i} \leq 1 + \frac{1/r\psi}{n_i/\phi} = \frac{\phi}{n_i} \left( \frac{n_i}{\phi} + \frac{1}{r\psi} \right) \leq \frac{\phi}{n_i} a_i$$

from which it follows that

$$\Sigma_{ii} \leq \frac{\phi}{n_i}.$$

We thus confirm that the incorporation of prior information has resulted in a reduction of the variance.

## 8.7 Exercises on Chapter 8

1. Show that the prior

$$p(\alpha, \beta) \propto (\alpha + \beta)^{-5/2}$$

suggested in connection with the example on risk of tumour in a group of rats is equivalent to a density uniform in  $(\alpha/(\alpha + \beta), (\alpha + \beta)^{-1/2})$ .

2. Observations  $x_1, x_2, \dots, x_n$  are independently distributed given parameters  $\theta_1, \theta_2, \dots, \theta_n$  according to the Poisson distribution  $p(x_i|\theta) = \theta_i^{x_i} \exp(-\theta_i)/x_i$ . The prior distribution for  $\theta$  is constructed hierarchically. First, the  $\theta_i$ 's are assumed to be independently identically distributed given a hyperparameter  $\phi$  according to the exponential distribution  $p(\theta_i|\phi) = \phi \exp(-\phi\theta_i)$  for  $\theta_i \geq 0$  and then  $\phi$  is given the improper uniform prior  $p(\phi) \propto 1$  for  $\phi \geq 0$ . Provided that  $n\bar{x} > 1$ , prove that the posterior distribution of  $z = 1/(1 + \phi)$  has the beta form

$$p(z|x) \propto z^{n\bar{x}-2}(1-z)^n.$$

Thereby show that the posterior means of the  $\theta_i$  are shrunk by a factor  $(n\bar{x} - 1)/(n\bar{x} + n)$  relative to the usual classical procedure which estimates each of the  $\theta_i$  by  $x_i$ .

What happens if  $n\bar{x} \leq 1$ ?

3. Carry out the Bayesian analysis for known overall mean developed in Section 8.2 mentioned earlier (a) with the loss function replaced by a weighted mean

$$\mathcal{L}(\theta, \hat{\theta}) = \sum_{i=1}^r w_i(\theta_i - \hat{\theta}_i)^2,$$

and (b) with it replaced by

$$\mathcal{L}(\theta, \hat{\theta}) = \sum_{i=1}^r |\theta_i - \hat{\theta}_i|.$$

4. Compare the effect of the Efron–Morris estimator on the baseball data in Section 8.3 with the effect of a James–Stein estimator which shrinks the values of  $\pi_i$  towards  $\pi_0 = 0.25$  or equivalently shrinks the values of  $X_i$  towards  $\mu = 2\sqrt{n} \sin^{-1} \sqrt{\pi_0}$ .

5. The *Helmert transformation* is defined by the matrix

$$A = \begin{pmatrix} r^{-1/2} & 2^{-1/2} & 6^{-1/2} & 12^{-1/2} & \dots & \{r(r-1)\}^{-1/2} \\ r^{-1/2} & -2^{-1/2} & 6^{-1/2} & 12^{-1/2} & \dots & \{r(r-1)\}^{-1/2} \\ r^{-1/2} & 0 & -2 \times 6^{-1/2} & 12^{-1/2} & \dots & \{r(r-1)\}^{-1/2} \\ r^{-1/2} & 0 & 0 & -3 \times 12^{-1/2} & \dots & \{r(r-1)\}^{-1/2} \\ r^{-1/2} & 0 & 0 & 0 & \dots & \{r(r-1)\}^{-1/2} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ r^{-1/2} & 0 & 0 & 0 & \dots & -(r-1)^{1/2}r^{-1/2} \end{pmatrix},$$

so that the element  $a_{ij}$  in row  $i$ , column  $j$  is

$$a_{ij} = \begin{cases} r^{-1/2} & (j=1) \\ \{j(j-1)\}^{-1/2} & (i < j) \\ 0 & (i > j > 1) \\ -(j-1)^{1/2}j^{-1/2} & (i = j > 1). \end{cases}$$

It is also useful to write  $\alpha_j$  for the (column) vector which consists of the  $j$ th column of the matrix  $A$ . Show that if the variates  $X_i$  are independently  $N(\theta_i, 1)$ , then the variates  $W_j = \alpha_j^T(X - \mu) = \sum_i a_{ij}(X_i - \mu_i)$  are independently normally distributed with unit variance and such that  $EW_j = 0$  for  $j > 1$  and  $W^T W = \sum_j W_j^2 = \sum_i (X_i - \mu_i)^2 = (X - \mu)^T(X - \mu)$ .

By taking  $a_{ij} \propto \theta_j - \mu_j$  for  $i > j$ ,  $a_{ij} = 0$  for  $i < j$  and  $a_{jj}$  such that  $\sum_j a_{ij} = 0$ , extend this result to the general case and show that  $EW_1 \propto \gamma = \sum_i (\theta_i - \mu_i)^2$ . Deduce that the distribution of a non-central chi-squared variate depends only of  $r$  and  $\gamma$ .

**6.** Show that  $R(\theta, \hat{\theta}^{JS+}) < R(\theta, \hat{\theta}^{JS})$ , where

$$\hat{\theta}^{JS+} = \mu + \max \left[ \left( 1 - \frac{r-2}{S_1} \right), 0 \right] (X - \mu)$$

(Lehmann 1983, Section 4.6, Theorem 6.2).

**7.** Writing

$$\hat{\theta} = (A^T A)^{-1} A^T x, \quad \hat{\theta}_k = (A^T A + kI)^{-1} A^T x$$

for the least-squares and ridge regression estimators for regression coefficients  $\theta$ , show that

$$\hat{\theta} - \hat{\theta}_k = k(A^T A)^{-1} \hat{\theta}_k$$

and that the bias of  $\hat{\theta}_k$  is

$$b(k) = \{(A^T A + kI)^{-1} A^T A - I\}\theta$$

while its variance-covariance matrix is

$$\sqrt{\hat{\theta}_k} = \phi(A^T A + kI)^{-1} A^T A (A^T A + kI)^{-1}.$$

Deduce expressions for the sum  $\mathcal{G}(k)$  of the squares of the biases and for the sum  $\mathcal{F}(k)$  of the variances of the regression coefficients, and hence show

that the mean square error is

$$MSE_k = E(\hat{\theta}_k - \theta)^T (\hat{\theta}_k - \theta) = \mathcal{F}(k) + \mathcal{G}(k).$$

Assuming that  $\mathcal{F}(k)$  is continuous and monotonic decreasing with  $\mathcal{F}'(0) = 0$  and that  $\mathcal{G}(k)$  is continuous and monotonic increasing with  $\mathcal{G}(k) = \mathcal{G}'(k) = 0$ , deduce that there always exists a  $k$  such that  $MSE_k < MSE_0$  (Theobald, 1974).

**8.** Show that the matrix  $H$  in Section 8.6 satisfies  $B^T H^{-1} B = 0$  and that if  $B$  is square and non-singular then  $H^{-1}$  vanishes.

**9.** Consider the following particular case of the two way layout. Suppose that eight plots are harvested on four of which one variety has been sown, while a different variety has been sown on the other four. Of the four plots with each variety, two different fertilizers have been used on two each. The yield will be normally distributed with a mean  $\theta$  dependent on the fertiliser and the variety and with variance  $\phi$ . It is supposed a priori that the mean for plots yields sown with the two different varieties are independently normally distributed with mean  $\alpha$  and variance  $\psi_\alpha$ , while the effect of the two different fertilizers will add an amount which is independently normally distributed with mean  $\beta$  and variance  $\psi_\beta$ . This fits into the situation described in Section 8.6 with  $\Phi$  being  $\phi$  times an  $8 \times 8$  identity matrix and

$$A = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 \end{pmatrix}; \quad B = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}; \quad \Psi = \begin{pmatrix} \psi_\alpha & 0 & 0 & 0 \\ 0 & \psi_\alpha & 0 & 0 \\ 0 & 0 & \psi_\beta & 0 \\ 0 & 0 & 0 & \psi_\beta \end{pmatrix}.$$

Find the matrix  $K^{-1}$  needed to find the posterior of  $\theta$ .

**10.** Generalize the theory developed in Section 8.6 to deal with the case where  $x \sim N(A\theta, \Phi)$  and  $\theta \sim N(B\mu, \Psi)$  and knowledge of  $\mu$  is vague to deal with the case where  $\mu \sim N(C\nu, K)$  (Lindley and Smith, 1972).

**11.** Find the elements of the variance–covariance matrix  $\Sigma$  for the one way model in the case where  $n_i=n$  for all  $i$ .

# 9

## The Gibbs sampler and other numerical methods

### 9.1 Introduction to numerical methods

#### 9.1.1 Monte Carlo methods

Bayesian statistics proceeds smoothly and easily as long as we stick to well-known distributions and use conjugate priors. But in real life, it is often difficult to model the situations which arise in practice in that way; instead, we often arrive at a posterior  $p(\theta|x) \propto p(\theta)l(\theta|x)$

but have no easy way of finding the constant multiple and have to have recourse to some numerical technique. This is even more true when we come to evaluate the marginal density of a single component  $\theta_i$  of  $\theta$ . A simple example was given earlier in Section 3.9 on ‘The circular normal distribution’, but the purpose of this chapter is to discuss more advanced numerical techniques and the Gibbs sampler in particular.

There are, of course, many techniques available for numerical integration, good discussions of which can be found in, for example, Davis and Rabinowitz (1984), Evans (1993) or Evans and Swartz (2000). Most of the methods we shall discuss are related to the idea of ‘Monte Carlo’ integration as a method of finding an expectation. In the simplest version of this, we write

$$\int_a^b f(x) p(x) dx \cong \frac{1}{n} \sum_{i=1}^n f(x_i),$$

where the points  $x_1, x_2, \dots, x_n$  are chosen as independent ‘pseudo-random’ numbers with density  $p(x)$  on the interval  $(a, b)$ , which in the simplest case is the uniform distribution  $U(a, b)$ . Although the results of this technique get better in higher dimensions, ‘In all [dimensions], crude Monte Carlo exists as a last resort, especially for integrals defined over nonstandard regions and for integrands of low-order continuity. It is usually quite reliable but very expensive

and not very accurate' (Davis and Rabinowitz, op. cit., Section 5.10). We shall not, however, use crude Monte Carlo as such, but rather certain refinements of it.

An extension of crude Monte Carlo is provided by *importance sampling*. This method is useful when we want to find a parameter  $\theta$  which is defined as the expectation of a function  $f(x)$  with respect to a density  $q(x)$  but we cannot easily generate random variables with that density although we *can* generate variates  $x_i$  with a density  $p(x)$  which is such that  $p(x)$  roughly approximates  $|f(x)|q(x)$  over the range of integration. Then

$$\theta = \int_a^b f(x) q(x) dx = \int_a^b f(x) \left( \frac{q(x)}{p(x)} \right) p(x) dx \cong \frac{1}{n} \sum_{i=1}^n \frac{f(x_i) q(x_i)}{p(x_i)}.$$

The function  $p(x)$  is called an *importance function*.

### 9.1.2 Markov chains

Another key notion is that of a Markov chain, which can be thought of as a model for a system which moves randomly through series of 'states' without having any memory of where it has been – where it jumps to next depends solely on where it is now. Another way of putting this is to say that given the present, the past and the future are independent. We thus have a probability density, called the *transition probability* density representing the chance of taking the state  $y$  at time  $t$  given that its state at time  $t-1$  is  $x$ . In the cases, we are most interested in, this density will not depend on  $t$  and so takes the form  $p(y|x)$ . If the probability density of its state at time 0 is  $p^{(0)}(x)$ , then clearly the density of its

state at time 1 is given by the generalized addition law as

$$p^{(1)}(y) = \sum_x p^{(0)}(x) p(y|x).$$

A similar result with the sum replaced by an integral holds when the *state space*, that is, the set of possible states, is continuous. Iterating this process we can clearly find the distribution of the state at any time  $t$  in terms of  $p^{(0)}(x)$  and  $p(y|x)$ . The key point for our purposes is that in many cases this density converges to a limit  $p(y)$  which does not depend on  $p^{(0)}(x)$  but rather is uniquely determined by  $p(y|x)$ . The limiting distribution is known as the *stationary distribution* or the *invariant distribution*. Properties of Markov chains which can take only a finite or countable number of values are discussed in Feller (1968, Volume 1; Chapter XV) and an indication of the general theory is given in Breiman (1968, Section 7.4). A more detailed account can be found in Meyn and Tweedie (1993) and a more applied one in Norris (1997).

It transpires that some densities which we are interested in turn up as stationary distributions for Markov chains.

## 9.2 The EM algorithm

### 9.2.1 The idea of the *EM* algorithm

A useful numerical technique which finds the posterior mode, that is, the value at which  $p(\eta|x)$  or equivalently  $\log p(\eta|x)$  is a maximum, but does not provide full information on the posterior distribution is the *EM algorithm*. We can exemplify this by an example on genetic linkage due to Rao (1973, Section 5g) quoted by Dempster *et al.* (1977), by Gelfand and Smith (1990) and by Tanner (1996, Section 4.1). We have observations  $x = (x_1, x_2, x_3, x_4)$  with cell probabilities  $(\frac{1}{2} + \frac{1}{4}\eta, \frac{1}{4}(1-\eta), \frac{1}{4}(1-\eta), \frac{1}{4}\eta)$

and we want to estimate  $\eta$ . The likelihood is then

$$(\frac{1}{2} + \frac{1}{4}\eta)^{x_1} (\frac{1}{4}(1-\eta))^{x_2} (\frac{1}{4}(1-\eta))^{x_3} (\frac{1}{4}\eta)^{x_4} \propto (2 + \eta)^{x_1} (1 - \eta)^{x_2+x_3} \eta^{x_4}$$

What we do is to *augment* the data  $x$  by adding further data  $z$  to produce augmented data  $y$ . It should be noted that to a considerable extent the distinction between parameters of a model and augmentations to the data is artificial (i.e. fictitious augmentations to the data can be regarded as parameters of a model). In the example under consideration, the augmentation consists simply in splitting the first cell into two cells with probabilities  $\frac{1}{2}$  and  $\frac{1}{4}\eta$ . The advantage of this is that the likelihood then takes the much simpler form  $(\frac{1}{2})^{y_0} (\frac{1}{4}\eta)^{y_1} (\frac{1}{4}(1-\eta))^{y_2} (\frac{1}{4}(1-\eta))^{y_3} (\frac{1}{4}\eta)^{y_4} \propto \eta^{y_1+y_4} (1 - \eta)^{y_2+y_3}$

and if we take the standard reference prior  $\text{Be}(0, 0)$ , then the posterior is a beta distribution

$$p(\eta|y) \propto \eta^{y_1+y_4-1} (1 - \eta)^{y_2+y_3-1}.$$

Not much in life is free, and in this case we have to pay for the greater simplicity of the model by estimating the split of  $x_1$  into  $y_0$  and  $y_1$ . The *EM* algorithm for finding the posterior mode is an iterative method starting from some plausible guess  $\eta^{(0)}$  for the value of  $\eta$ . At stage  $t$ , we suppose that the current guess is  $\eta^{(t)}$ . Each stage has two steps. In the first, the *E*-step (expectation step), we compute  $Q(\eta, \eta^{(t)}) = E_{\eta^{(t)}} \log p(\eta|y)$

that is, the expectation of the log-likelihood function, the expectation being computed at  $\eta = \eta^{(t)}$ , so that  $Q(\eta, \eta^{(t)}) = \int \log\{p(\eta|y)\} p(y|\eta^{(t)}, x) dy$ .

In the second step, the *M*-step (the maximization step), we find that value  $\eta^{(t+1)}$  of  $\eta$  which maximizes  $Q(\eta, \eta^{(t)})$ . In this particular example as  $y_i=x_i$  for  $i > 1$

$$\begin{aligned} Q(\eta, \eta^{(t)}) &= \mathbb{E}\{(y_1 + y_4 - 1) \log \eta + (y_2 + y_3 - 1) \log(1 - \eta) \mid \eta^{(t)}, \mathbf{x}\} \\ &= (\mathbb{E}\{y_1 \mid \eta^{(t)}, \mathbf{x}\} + x_4 - 1) \log \eta + (x_2 + x_3 - 1) \log(1 - \eta). \end{aligned}$$

For the  $M$ -step, we note that

$$\frac{\partial Q(\eta, \eta^{(t)})}{\partial \eta} = \frac{\mathbb{E}\{y_1 \mid \eta^{(t)}, \mathbf{x}\} + x_4 - 1}{\eta} - \frac{x_2 + x_3 - 1}{1 - \eta},$$

and equating this to zero we deduce that

$$\eta^{(t+1)} = \frac{\mathbb{E}\{y_1 \mid \eta^{(t)}, \mathbf{x}\} + x_4 - 1}{\mathbb{E}\{y_1 \mid \eta^{(t)}, \mathbf{x}\} + x_2 + x_3 + x_4 - 2}.$$

Since  $y_1$  has a binomial  $B(x_1, \pi)$  distribution with

$$\pi = \frac{\frac{1}{4}\eta^{(t)}}{\frac{1}{2} + \frac{1}{4}\eta^{(t)}} = \frac{\eta^{(t)}}{\eta^{(t)} + 2},$$

so that

$$\mathbb{E}\{y_1 \mid \eta^{(t)}, \mathbf{y}\} = x_1 \eta^{(t)} / (\eta^{(t)} + 2)$$

the iteration becomes

$$\begin{aligned} \eta^{(t+1)} &= \frac{x_1 \eta^{(t)} / (\eta^{(t)} + 2) + x_4 - 1}{x_1 \eta^{(t)} / (\eta^{(t)} + 2) + x_2 + x_3 + x_4 - 2} \\ &= \frac{\eta^{(t)}(x_1 + x_4 - 1) + 2(x_4 - 1)}{\eta^{(t)}(x_1 + x_2 + x_3 + x_4 - 2) + 2(x_2 + x_3 + x_4 - 2)}. \end{aligned}$$

The values actually observed were  $x_1 = 125$ ,  $x_2 = 18$ ,  $x_3 = 20$ ,  $x_4 = 34$ . We can then estimate  $\eta$  by iteration starting from, for example,  $\eta^{(0)} = 0.5$  and using

$$\eta^{(t+1)} = \frac{158\eta^{(t)} + 66}{195\eta^{(t)} + 140}.$$

In fact, the iteration will converge to the positive root of  $195\eta^2 - 18\eta - 66 = 0$  which is 0.630.

## 9.2.2 Why the $EM$ algorithm works

We will first show that  $\log p(\eta^{(t)} \mid \mathbf{x})$  increases with  $t$  and, presuming that  $\eta^{(t)}$  converges to some limit  $\hat{\eta}$ , then show that  $\partial \log p(\eta \mid \mathbf{y}) / \partial \eta = 0$  at  $\eta = \hat{\eta}$ , so that  $\hat{\eta}$  will normally be the posterior mode.

From the obvious equation  $p(\eta) = p(\eta \mid \mathbf{y})p(\mathbf{y}) / p(\mathbf{y} \mid \eta)$  considered conditional on  $\mathbf{x}$ , it follows, on noting that  $\mathbf{y}$  includes  $\mathbf{x}$  and hence  $p(\eta \mid \mathbf{y}, \mathbf{x}) = p(\eta \mid \mathbf{y})$ , that  $\log p(\eta \mid \mathbf{x}) = \log p(\eta \mid \mathbf{y}) - \log p(\mathbf{y} \mid \eta, \mathbf{x}) + \log p(\mathbf{y} \mid \mathbf{x})$ .

Multiplying both sides by the density  $p(\mathbf{y} \mid \eta^*, \mathbf{x})$  and integrating (noting that the left hand side does not depend on  $\mathbf{y}$ ), we get for any fixed  $\eta^*$

$$\begin{aligned} \log p(\eta \mid \mathbf{x}) &= \int \log\{p(\eta \mid \mathbf{y})\} p(\mathbf{y} \mid \eta^*, \mathbf{x}) d\mathbf{y} - \int \log\{p(\mathbf{y} \mid \eta, \mathbf{x})\} p(\mathbf{y} \mid \eta^*, \mathbf{x}) d\mathbf{y} \\ &\quad + \int \log\{p(\mathbf{y} \mid \mathbf{x})\} p(\mathbf{y} \mid \eta^*, \mathbf{x}) d\mathbf{y} \\ &= Q(\eta, \eta^*) - H(\eta, \eta^*) + K(\eta^*), \text{ say} \end{aligned}$$

(we note that  $K$  clearly does not depend on  $\eta$ ). Taking  $\eta^* = \eta^{(t)}$ , we get

$$\begin{aligned}\log p(\eta^{(t+1)}|x) - \log p(\eta^{(t)}|x) &= [Q(\eta^{(t+1)}, \eta^{(t)}) - Q(\eta^{(t)}, \eta^{(t)})] \\ &\quad - [H(\eta^{(t+1)}, \eta^{(t)}) - H(\eta^{(t)}, \eta^{(t)})].\end{aligned}$$

Now  $Q(\eta^{(t+1)}, \eta^{(t)})Q(\eta^{(t)}, \eta^{(t)})$  because of the way in which we chose  $\eta^{(t+1)}$ . Moreover,

$$\begin{aligned}-[H(\eta^{(t+1)}, \eta^{(t)}) - H(\eta^{(t)}, \eta^{(t)})] &= \int \log \left\{ \frac{p(y|\eta^{(t)}, x)}{p(y|\eta^{(t+1)}, x)} \right\} p(y|\eta^{(t)}, x) dy \\ &= \int \phi(g(y)) p(y|\eta^{(t+1)}, x) dy,\end{aligned}$$

where

$$\phi(u) = u \log u \quad \text{and} \quad g(y) = \frac{p(y|\eta^{(t)}, x)}{p(y|\eta^{(t+1)}, x)}.$$

Because (as is easily seen by differentiation)  $\phi(u) - u + 1$  takes a minimum value of 0 when  $u=1$ , we see that  $\phi(t) \geq t - 1$  for all  $t$ . Consequently,  $-[H(\eta^{(t+1)}, \eta^{(t)}) - H(\eta^{(t)}, \eta^{(t)})] \geq \int (g(y) - 1) p(y|\eta^{(t+1)}, x) dy$

$$\begin{aligned}&= \int p(y|\eta^{(t)}, x) dy - \int p(y|\eta^{(t+1)}, x) dy \\ &= 1 - 1 = 0,\end{aligned}$$

so that

$$\log p(\eta^{(t+1)}|x) \geq \log p(\eta^{(t)}|x).$$

We now show that the limit is a stationary point. As

$$\left. \frac{\partial Q(\eta, \eta^{(t)})}{\partial \eta} \right|_{\eta=\eta^{(t+1)}} = 0,$$

we see that provided  $Q$  is a reasonably smooth function

$$\left. \frac{\partial Q(\eta, \hat{\eta})}{\partial \eta} \right|_{\eta=\hat{\eta}} = 0.$$

Further

$$\left. \frac{\partial H(\eta, \eta^*)}{\partial \eta} \right|_{\eta=\eta^*} = \int \left\{ \frac{\partial p(y|\eta, x)/\partial \eta}{p(y|\eta, x)} \right\} p(y|\eta^*, x) dy$$

and in particular taking  $\eta = \eta^* = \hat{\eta}$

$$\left. \frac{\partial H(\eta, \hat{\eta})}{\partial \eta} \right|_{\eta=\hat{\eta}} = \int \frac{\partial p(y|\eta, x)}{\partial \eta} dy = \frac{\partial}{\partial \eta} \int p(y|\eta, x) dy = \frac{\partial}{\partial \eta} 1 = 0.$$

Since  $\log p(\eta|x) = Q(\eta, \eta^*) - H(\eta, \eta^*) + K(\eta^*)$  for any fixed  $\eta^*$ , and in particular

$$\left. \frac{\partial \log p(\eta|x)}{\partial \eta} \right|_{\eta=\hat{\eta}} = 0,$$

so that  $\hat{\eta}$  is a stationary point, which will usually be the posterior mode.

More information about the *EM* algorithm can be found in Dempster *et al.*

(1977) or Tanner (1996).

### 9.2.3 Semi-conjugate prior with a normal likelihood

Suppose that we have independent normally distributed observations  $x_1, x_2, \dots, x_n \sim N(\theta, \phi)$ . We use the notation  $\bar{x} = \sum x_i/n$  and  $S_x = \sum (x_i - \bar{x})^2$ .

In Section 2.13, we noted that the conjugate joint prior  $p(\theta, \phi)$  for the normal mean and variance is *not* a product of a function of  $\theta$  and  $\phi$ , so that  $\theta$  and  $\phi$  are *not* independent a priori. Nevertheless, an assumption of prior independence would seem appropriate for many problems, and so we are sometimes led to consider a situation in which  $\theta \sim N(\theta_0, \phi_0)$  and  $\phi \sim S_0 \chi_v^{-2}$

independently of one another. Such a prior is described as *semi-conjugate* by Gelman *et al.* (2004, Section 3.4) and as *conditionally conjugate* by O'Hagan (2004, Section 6.33). The latter term arises because conditional on knowledge of  $\theta$ , it is a conjugate prior for  $\phi$  and vice versa.

With such a prior, we know that, conditional on knowledge of  $\phi$ , the posterior of  $\theta$  is  $\theta | \phi \sim N(\theta_1, \phi_1)$ ,

where

$$\phi_1 = \{\phi_0^{-1} + (\phi/n)^{-1}\}^{-1} \quad \text{and} \quad \theta_1 = \phi_1 \{\theta_0/\phi_0 + \bar{x}/(\phi/n)\}$$

(cf. Section 2.3) whereas, conditional on knowledge of  $\theta$ , the posterior of  $\phi$  is  $\phi | \theta \sim (S_0 + S) \chi_{v+n}^{-2}$ ,

where

$$S = \sum (x_i - \theta)^2 = S_x + n(\bar{x} - \theta)^2$$

(cf. Section 2.7).

In such a case, we can use the *EM* algorithm to estimate the posterior mode of  $\theta$ , effectively augmenting the data with the variance  $\phi$ . We begin by observing that (ignoring terms which do not depend on  $\theta$ ) the log posterior density is  $\log p(\theta | \phi, x) = -\frac{1}{2}(\theta - \theta_0)^2/\phi_0 - \frac{1}{2} \sum (x_i - \theta)^2/\phi$ .

To carry out the *E*-step we first note that since

$$\phi \sim (S_0 + S) \chi_{v+n}^{-2}$$

and the expectation of a chi-squared variate is equal to its number of degrees of freedom, we know that

$$Q(\theta, \theta^{(t)}) = -\frac{1}{2}(\theta - \theta_0)^2/\phi_0 - \frac{1}{2}n_1 \sum (x_i - \theta)^2/S_1,$$

where  $S_1 = S_0 + S = S_0 + S_x + n(\bar{x} - \theta^{(t)})^2$  and  $n_1 = v + n$ . For the *M*-step we must find that value  $\theta^{(t+1)}$  which maximizes  $Q(\theta, \theta^{(t)})$ . But  $Q(\theta, \theta^{(t)})$  is of the form of the log of the posterior density we would have got for  $\theta$  had we had the

same prior ( $N(\theta_0, \phi_0)$ ) and observations  $x_i \sim N(\theta, S_1/n_1)$  with mean  $\theta$  and *known* variance  $S_1/n_1$ . Now it follows from Section 2.3 on ‘Several normal observations with a normal prior’ that the posterior mean of  $\theta$  in this case is at

$$\theta^{(t+1)} = \frac{\theta_0/\phi_0 + n\bar{x}/(S_1/n_1)}{1/\phi_0 + n/(S_1/n_1)}$$

(the right-hand side depends on  $\theta^{(t)}$  through  $S_1$ ), and so we see that this is the value of  $\theta$  giving the required maximum.

To illustrate this case, we return to the data on wheat yield considered in the example towards the end of Section 2.13 in which  $n=12$ ,  $\bar{x} = 119$  and  $S=13\,045$ . We will take the prior for the variance considered in that Section, namely,  $\phi \sim S_0 \chi^2_\nu$  with  $S_0=2700$  and  $\nu = 11$ . For the mean, we will take a prior which is  $N(\theta_0, \phi_0)$  where  $\theta_0 = 110$  and  $\phi_0 = S_0/n_0(\nu - 2) = 2700/(15 \times 9) = 20$ , which approximates well to the values for the marginal distribution in that Section (according to which  $(\theta - \theta_0)/\sqrt{(S_0/n_0)} \sim \nu$ ). The resulting *marginal* densities are nearly the same, but because we now assume prior independence the *joint* distribution is different.

It seems that a reasonable starting point for the iteration is  $\theta^{(0)} = 110$ . Then we get  $\theta^{(1)} = 112.234$ ,  $\theta^{(2)} = 112.277$ , and  $\theta^{(t)} = 112.278$  thereafter.

### 9.2.4 The *EM* algorithm for the hierarchical normal model

The *EM* algorithm is well suited to the analysis of hierarchical models and we can see this by examining the important case of the hierarchical normal model. Although the formulae we shall derive do not look pretty, they are very easy to use. Suppose that  $x_{ij} \sim N(\theta_i, \phi)$  ( $i = 1, 2, \dots, r$ ;  $j = 1, 2, \dots, n_i$ )

with  $N = \sum n_i$ , where

$$\theta_i \sim N(\mu, \psi)$$

as in Section 8.5 and that we wish to estimate the hyperparameters  $\mu, \phi$  and  $\psi$ . In this case  $\eta = (\mu, \phi, \psi)$  takes the place of  $\eta$  in the example on genetic linkage, while we augment the data by  $z = \theta$  to give augmented data  $y = (x, \theta)$ . We can use a reference prior for  $\phi$ , but (as mentioned in Section 8.5) this will not work for  $\psi$ . Accordingly, we will adopt the prior  $p(\mu, \phi, \psi) \propto 1/\phi$  ( $0 < \mu < \infty$ ;  $\phi > 0$ ;  $\psi > 0$ ).

It is then easily seen that (up to an additive constant)

$$\begin{aligned}\log p(\mu, \phi, \psi | \mathbf{x}, \boldsymbol{\theta}) = & -\frac{1}{2}(N+2)\log\phi - \frac{1}{2}\sum_i\sum_j(x_{ij} - \theta_i)^2/\phi \\ & -\frac{1}{2}r\log\psi - \frac{1}{2}\sum_i(\theta_i - \mu)^2/\psi.\end{aligned}$$

To carry out the *E*-step, observe that conditional on the value of  $\eta^{(t)} = (\mu^{(t)}, \phi^{(t)}, \psi^{(t)})$  the parameters  $\theta_i \sim N(\theta_i^{(t)}, \mathcal{V}_i^{(t)})$ , where  
 $\mathcal{V}_i^{(t)} = \{1/\psi^{(t)} + n_i/\phi^{(t)}\}^{-1}$   
 $\theta_i^{(t)} = \mathcal{V}_i^{(t)} \{\mu^{(t)}/\psi^{(t)} + n_i x_{i*}/\phi^{(t)}\}$

using the notation introduced in Section 6.5 for averaging over a suffix (cf. Section 2.3). We can now see that

$$\mathbb{E}\{(x_{ij} - \theta_i)^2 | \eta^{(t)}\} = \mathcal{V}_i^{(t)} + (x_{ij} - \theta_i^{(t)})^2$$

and similarly

$$\mathbb{E}\{(\theta_i - \mu)^2 | \eta^{(t)}\} = \mathcal{V}_i^{(t)} + (\mu - \theta_i^{(t)})^2,$$

so that

$$\begin{aligned}Q(\eta, \eta^{(t)}) = & -\frac{1}{2}(N+2)\log\phi - \frac{1}{2}\sum_{i=1}^r\sum_{j=1}^{n_i}\left\{\mathcal{V}_i^{(t)} + (x_{ij} - \theta_i^{(t)})^2\right\}/\phi \\ & -\frac{1}{2}r\log\psi - \frac{1}{2}\sum_{i=1}^r\left\{\mathcal{V}_i^{(t)} + (\mu - \theta_i^{(t)})^2\right\}/\psi.\end{aligned}$$

It is now clear that the *M*-step gives

$$\begin{aligned}\mu^{(t+1)} &= \frac{1}{r}\sum_{i=1}^r\theta_i^{(t)} \\ \phi^{(t+1)} &= \frac{1}{N+2}\sum_{i=1}^r\sum_{j=1}^{n_i}\left\{\mathcal{V}_i^{(t)} + (x_{ij} - \theta_i^{(t)})^2\right\} \\ \psi^{(t+1)} &= \frac{1}{r}\sum_{i=1}^r\left\{\mathcal{V}_i^{(t)} + (\mu - \theta_i^{(t)})^2\right\}.\end{aligned}$$

### 9.2.5 A particular case of the hierarchical normal model

Gelman *et al.* (2004, Table 11.2) quote the following data from Box *et al.* (1978, Table 6.1) on the coagulation time in seconds for blood drawn from  $N=24$  animals randomly allocated to four different diets:

$i$	$n_i$	$x_{i*}$	$\sum(x_{ij} - x_{i*})^2$
1	62	60	63
	59		
2	63	67	71
	64	65	66
3	68	66	71
	67	68	68
4	56	62	60
	61	63	64
	63	59	8
	61		61
			48

The data has been slightly adjusted, so that the averages come out to be whole numbers. It should perhaps be noted that the prior adopted in this and the

preceding subsections differs slightly from that adopted by Gelman *et al.* The within-groups sum of squares is  $\sum \sum (x_{ij} - \bar{x}_{i\cdot})^2 = 112$  and the overall mean is  $\bar{x}_{..} = 64$ . It, therefore, seems reasonable to start iterations from  $\mu^{(0)} = 64$  and  $\phi^{(0)} = 112/23 = 4.870$ . As for  $\psi^{(0)}$ , we can take  $\{(61 - 64)^2 + (66 - 64)^2 + (68 - 64)^2 + (61 - 64)^2\} / 3 = 38/3 = 12.667$ .

This results in

$$\begin{aligned}\mathcal{V}_1^{(0)} &= \{1/\psi^{(0)} + n_1/\phi^{(0)}\}^{-1} = 1.111 \\ \theta_1^{(0)} &= \mathcal{V}_1^{(0)} \{\mu^{(0)}/\psi^{(0)} + n_1 \bar{x}_{1\cdot}/\phi^{(0)}\} = 61.263.\end{aligned}$$

Similarly, we get  $\mathcal{V}_2^{(0)} = 0.762$ ,  $\theta_2^{(0)} = 65.880$ ,  $\mathcal{V}_3^{(0)} = 0.762$ ,  $\theta_3^{(0)} = 67.759$ ,  $\mathcal{V}_4^{(0)} = 0.581$  and  $\theta_4^{(0)} = 61.138$ .

We can now feed these values in and commence the iteration, getting  $\mu^{(1)} = 64.037$ ,  $\phi^{(1)} = 9.912$  and  $\psi^{(1)} = 21.808$ . Continuing the iteration, we get rapid convergence to  $\theta_1 = 61.388$ ,  $\theta_2 = 65.822$ ,  $\theta_3 = 67.643$ ,  $\theta_4 = 61.207$ ,  $\mu = 64.015$ ,  $\phi = 5.090$  and  $\psi = 8.614$ .

## 9.3 Data augmentation by Monte Carlo

### 9.3.1 The genetic linkage example revisited

We can illustrate this technique with the example on genetic linkage we considered in connection with the *EM* algorithm. Recall that we found that likelihood of the augmented data was  $p(\mathbf{y}|\eta) \propto \eta^{y_1+y_4}(1-\eta)^{y_2+y_3}$ .

In this method, we suppose that at each stage we have a ‘current’ distribution for  $\eta$ , which initially is the prior distribution. At all stages, this has to be a proper distribution, so we may as well take our prior as the uniform distribution  $\text{Be}(1, 1)$ , which in any case differs little from the reference prior  $\text{Be}(0, 0)$ . At the  $t$ th stage in the *imputation step*, we pick  $m$  possible values  $\eta^{(1)}, \dots, \eta^{(m)}$  of  $\eta$  by some (pseudo-) random mechanism with the current density, and then for each of these values of  $\eta$ , we generate a value for the augmented data  $y$ , which in the particular example simply means picking a value  $y_1^{(i)}$  with a binomial distribution of index  $x_1$  and parameter  $\eta^{(i)} / (\eta^{(i)} + 2)$ . Since we had a  $\text{Be}(1, 1)$  prior, this gives a posterior  $p(\eta|y^{(i)}) \sim \text{Be}(y_1^{(i)} + x_4, x_2 + x_3)$ .

In the *posterior step* we now take the new ‘current’ distribution of  $\eta$  as a mixture of the  $m$  beta distributions so generated, all values of  $i$  being treated as equally likely (cf. the end of Section 2.10). We could then construct an estimate for the

posterior  $p(\eta|x)$  from a histogram of the values of  $\eta$  found at each iteration, but a better and smoother estimate results from taking the ‘current’ distribution at the end of the iteration.

It is worth noting that,

In practice, it is inefficient to take  $m$  large during the first few iterations when the posterior distribution is far from the true distribution. Rather, it is suggested that  $m$  initially be small and then increased with successive iterations (Tanner and Wong, 1987).

### 9.3.2 Use of $\text{R}$

A number of examples in this chapter and the next are set out as programs in  $\text{R}$ . The  $\text{R}$  project for statistical computing is described on the web page

<http://www.r-project.org/>

and useful books covering it are Dalgaard (2008), Krause and Olsen (2000), Fox (2002) and Venables and Ripley (2002). A very good book on Bayesian statistics with examples in  $\text{R}$  is Albert (2009). At a lower level, Gill (2002) is useful (although it may be noted that his highest posterior densities (HPDs) are only approximately highest density regions). Another book with much useful material is Robert and Casella (2010). For the ordinary user,  $\text{R}$  is virtually indistinguishable from S-plus, but has the advantage that it is free.

Even if you do not know  $\text{R}$ , these programs should be reasonably easy to follow once you realize that  $<-$  (a form of arrow) is an assignment operator. In fact, programs are provided on the web site associated with the book for all the numerical examples in this book, making use of the programs for finding HDRs and various functions associated with Behrens’ distribution which can be found in Appendix C.

### 9.3.3 The genetic linkage example in $\text{R}$

A program in  $\text{R}$  for the genetic linkage example is as follows:

```
niter <- 50
mininitial <- 20
etadivs <- 1000
mag <- 10
scale <- 8
x <- c(125, 18, 20, 34)
```

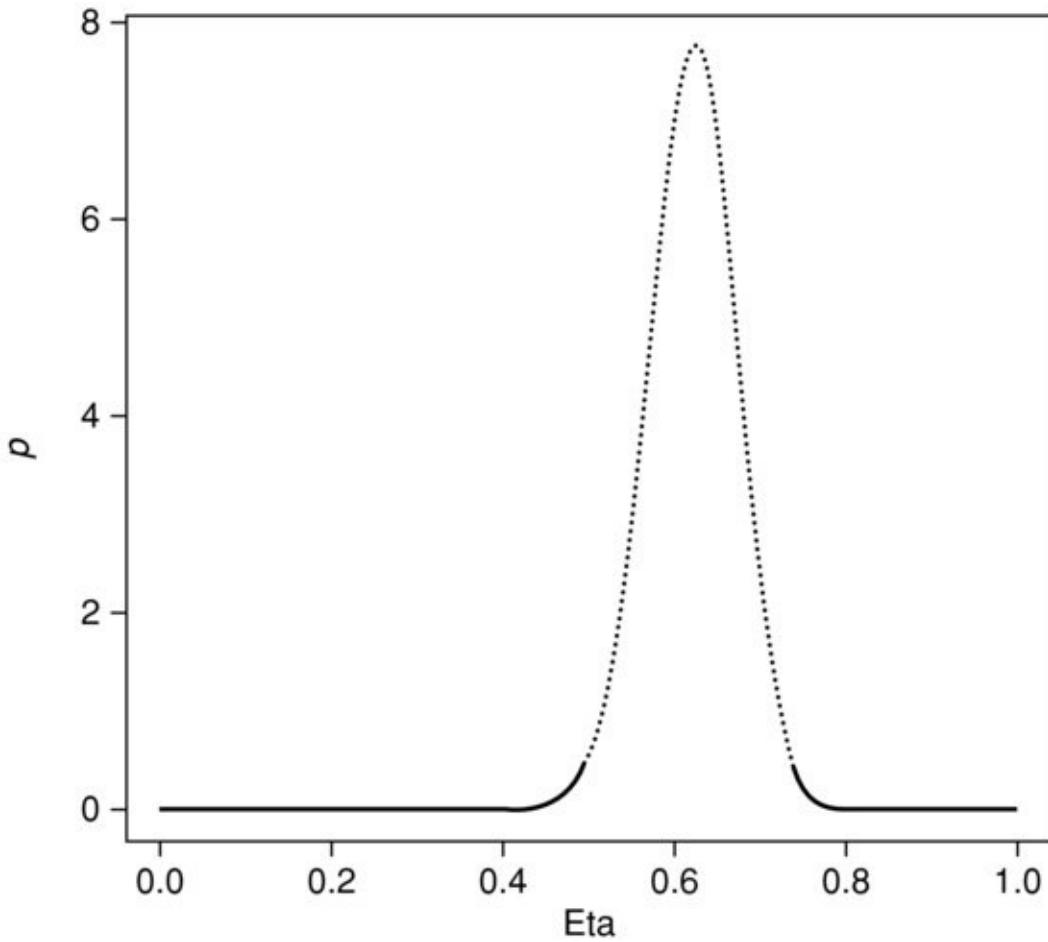
```

pagethrow < - 12
m <- minitial
eta <- runif(m) # random from U(0,1)
y <- rbinom(m,x[1],eta/(eta+2)) # random binomial
for (t in 1:niter)
{
  mold <- m
  if (t > 30)
    m <- 200
  if (t > 40)
    x[m] <- 1000
  i0 <- floor(runif(m,0,mold))
  eta <- rbeta(m,y[i0]+x[4]+1,x[2]+x[3]+1) # random beta
  y <- rbinom(m,x[1],eta/(eta+2))
}
p <- rep(0,etadivs) # vector of etadivs zeroes
for (etastep in 1:(etadivs-1)){
  eta <- etastep/etadivs
  term <- exp((y+x[4])*log(eta) + (x[2]+x[3])*log(1-eta)
  +lgamma(y+x[2]+x[3]+x[4]+2)-lgamma(y+x[4]+1)
  -lgamma(x[2]+x[3]+1)) # lgamma is log gamma fn
  p[etastep] <- p[etastep] + sum(term)/m
}
plot(1:etadivs/etadivs,p,pch=".",xlab="eta")

```

The resulting plot is shown as [Figure 9.1](#). The function  $p(\eta)$  is thus evaluated for  $\eta = 0, 1/n, 2/n, \dots, 1$ , and can now be used to estimate the posterior density of  $\eta$ . A simulation in R with  $T=50$ ,  $m=m(t)=20$  for  $t \leq 30$ ,  $m=200$  for  $30 < t \leq 40$ ,  $m=1000$  for  $t > 40$ , and  $n=1000$  showed a posterior mode at  $\eta = 0.627$ , which is close to the value found earlier using the *EM* algorithm. However, it should be noted that this method gives an approximation to the whole posterior distribution as opposed to the *EM* algorithm which will only give the mode.

[Figure 9.1](#) Plot of  $p(\eta)$  for the linkage example.



Programs and algorithms for generating pseudo-random numbers with many common distributions can be found in Press *et al.* (1986–1993) or Kennedy and Gentle (1980); it may help to note that discrete uniform variates are easily constructed by seeing which of  $m$  equal intervals a  $U(0, 1)$  variate falls in, and a  $\text{Be}(\alpha, \beta)$  variate  $w$  can be constructed as  $u/(u+v)$ , where  $u \sim G(\alpha)$  and  $v \sim G(\beta)$ .

The data augmentation technique was originated by Tanner and Wong (1987) and is discussed in detail in Tanner (1996).

### 9.3.4 Other possible uses for data augmentation

Suppose that  $x_1, x_2, \dots, x_n$  are independently  $N(\theta, \phi)$ , where  $\theta$  is unknown but  $\phi$  is known but that instead of a normal prior for  $\theta$  you have a prior

$$p(\theta) \propto \{vs^2 + (\theta - \mu)^2\}^{-(v+1)/2},$$

so that

$$t = (\theta - \mu)/s \sim t_v$$

has a Student's t distribution (where  $\mu$  and  $s$  are known). The posterior for  $\theta$  in

this case is of a complicated form, but if we augment the data with a parameter  $\psi$  such that  $\psi \sim \chi^2_\nu$  and  $\theta | \psi \sim N(\mu, s^2\psi)$ ,

so that the unconditional distribution of  $\theta$  is as above (cf. Section 2.12;  $n$  in that section is here replaced by unity, so we do not have  $\nu = n - 1$ ), then the conditional distribution of  $\theta$  given  $\phi$  and  $x$  is normal and we can now use the algorithm to find  $p(\theta|x)$ .

Further examples can be found in Tanner (1996).

## 9.4 The Gibbs sampler

### 9.4.1 Chained data augmentation

We will now restrict our attention to cases where  $m=1$  and the augmented data  $y$  consists of the original data  $x$  augmented by a single scalar  $z$  (as in the linkage example). The algorithm can then be expressed as follows: Start from a value  $\eta^{(0)}$  generated from the prior distribution for  $\eta$  and then iterate as follows:

(a<sub>1</sub>) Choose  $\eta^{(i+1)}$  of  $\eta$  from the density  $p(\eta | z^{(i)}, x)$ ; (a<sub>2</sub>) Choose  $z^{(i+1)}$  of  $z$  from the density  $p(z | \eta^{(i+1)}, x)$ .

(There is, of course, a symmetry between  $\eta$  and  $z$  and the notation is used simply because it arose in connection with the first example we considered in connection with the data augmentation algorithm.) This version of the algorithm can be referred to as *chained data augmentation*, since it is easy to see that the distribution of the next pair of values  $(\eta, z)$  given the values up to now depends only on the present pair and so these pairs move as a Markov chain. It is a particular case of a numerical method we shall refer to as the Gibbs sampler. As a result of the properties of Markov chains, after a reasonably large number  $T$  iterations the resulting values of  $\eta$  and  $z$  have a joint density which is close to  $p(\eta, z | x)$ , irrespective of how the chain started.

Successive observations of the pair  $(\eta, z)$  will not, in general, be independent, so in order to obtain a set of observations which are, to all intents and purposes, independently identically distributed (i.i.d.), we can run the aforementioned process through  $T$  successive iterations, retaining only the final value obtained, on  $k$  different replications.

We shall illustrate the procedure by the following example due to Casella and George (1992). We suppose that  $\pi$  and  $y$  have the following joint distribution:

$$p(y, \pi) = \binom{n}{y} \pi^{y+\alpha-1} (1-\pi)^{n-y+\beta-1} \quad (x = 0, 1, \dots, n; 0 \leq y \leq 1)$$

and that we are interested in the marginal distribution of  $y$ . Rather than integrating with respect to  $\pi$  (as we did at the end of Section 3.1), which would show that  $y$  has a beta-binomial distribution, we proceed to find the required

$$y | \pi \sim B(n, \pi)$$

distribution from the two conditional distributions:  $\pi | y \sim Be(y + \alpha, n - y + \beta)$ .

This is a simple case in which there is no observed data  $x$ . We need to initialize the process somewhere, and so we may as well begin with a value of  $\pi$  which is chosen from a  $U(0, 1)$  distribution. Again the procedure is probably best expressed as an R program, which in this case constructs a histogram as well as the better and smoother estimate resulting from taking the ‘current’ distribution at the end of each iteration:

```

nr <- 50
m <- 500
k <- 10
n <- 16
alpha <- 2.0
beta <- 4.0
lambda <- 16.0
maxn <- 24
h <- rep(0, n+1)
for (i in 1:m)
{
  pi <- runif(1)
  for (j in 1:k)
  {
    y <- rbinom(1, n, pi)
    newalpha <- y + alpha
    newbeta <- n - y + beta
    pi <- rbeta(1, newalpha, newbeta)
  }
  for (t in 0:n)
  {
    if (t == y)
      h[t+1] <- h[t+1] + 1
  }
}

```

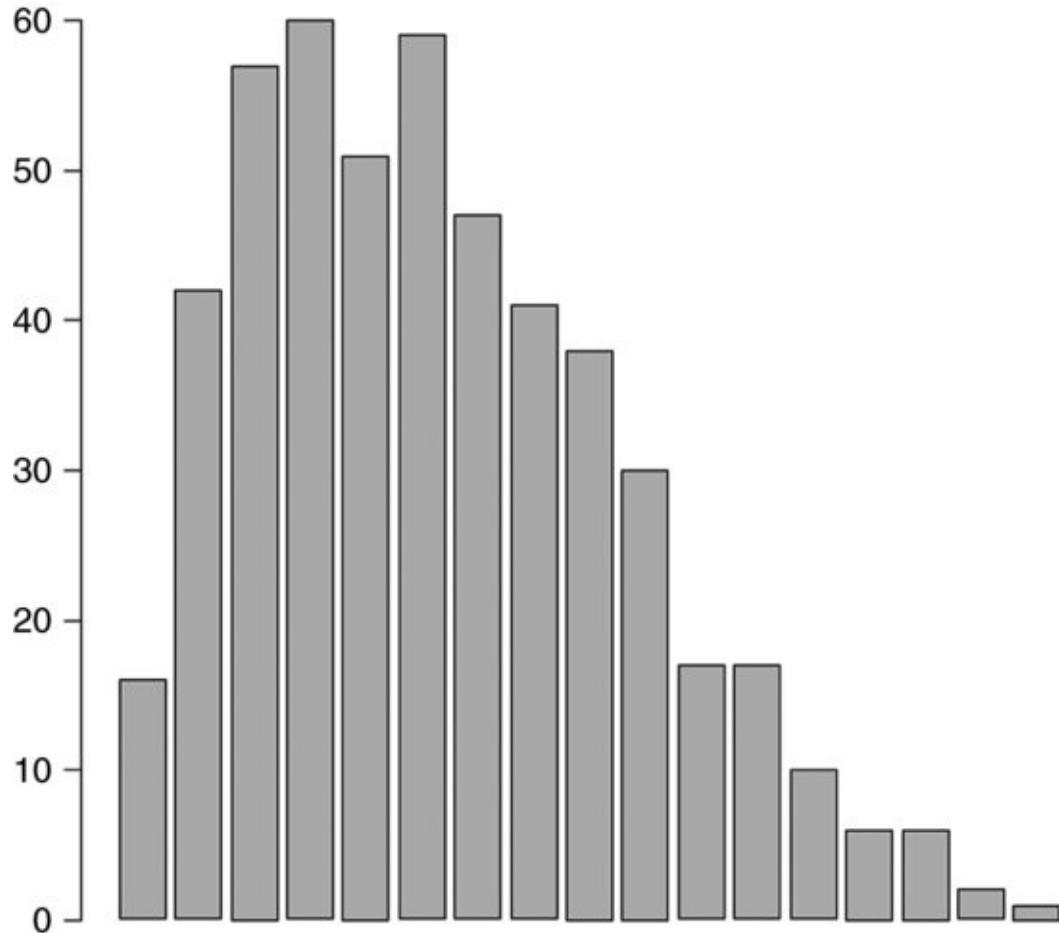
```

        }
barplot(h)

```

The resultant histogram is given as [Figure 9.2](#). References for the generation of ‘pseudo-random’ numbers can be found in the previous section. An implementation in R with  $T=10$  iterations replicated  $m=500$  times for the case  $n=16$ ,  $\alpha = 2$  and  $\beta = 4$  will, as Casella and George remark, give a very good approximation to the beta-binomial density.

[Figure 9.2](#) Barplot of values of  $y$  between 0 and 16.



## 9.4.2 An example with observed data

Gaver and O’Muircheartaigh (1987) quote the data later on the rates of loss of feedwater flow for a collection of nuclear power systems. In the following table, a small set of data representing failure of pumps in several systems of the nuclear plant Farley 1 is given. The apparent variation in failure rates has several sources.

Number	$y_i$	of	pump	failures	in	$t_i$	thousand	hours
System ( $i$ )	$y_i$		$t_i$		$r_i = y_i/t_i$			
1	5		94.320		0.05301103			
2	1		15.720		0.06361323			
3	5		62.880		0.07951654			
4	14		125.760		0.11132316			
5	3		5.240		0.57251908			
6	19		31.440		0.60432570			
7	1		1.048		0.95419847			
8	1		1.048		0.95419847			
9	4		2.096		1.90839695			
10	22		10.480		2.09923664			

It should seem plausible that the number of failures in any fixed time interval should have a Poisson distribution and that the mean of this distribution should be proportional to the length of the interval, with a constant of proportionality varying from pump to pump. It seems sensible to assume that these constants come from the conjugate family, the multiples of chi-squared (cf. Section 3.4), so that  $y_i | \theta_i \sim P(\theta_i t_i)$ ,  $\theta_i \sim S_0^{-1} \chi_{v'}^2$ .

For the moment, we shall regard it as known that  $v = 1.4$ . We seek the marginal distributions  $p(\theta_i | y)$ , but unfortunately these do not have a closed form. We need  $\theta = (\theta_1, \theta_2, \dots, \theta_k)$ ,

to write  $\theta_{-i} = (\theta_1, \dots, \theta_{i-1}, \theta_{i+1}, \dots, \theta_k)$ .

Now

$$p(\theta, y, S_0) = \prod p(y_i, \theta_i | S_0) p(S_0) = \prod p(y_i | \theta_i) p(\theta_i | S_0) p(S_0)$$

from which it easily follows that

$$\begin{aligned} p(\theta_i | \theta_{-i}, y, S_0) &= p(\theta, S_0, y) / p(\theta_{-i}, y, S_0) \\ &\propto p(y_i | \theta_i) p(\theta_i | S_0) \\ &= \frac{(\theta_i t_i)^{y_i}}{y_i!} \exp(-\theta_i t_i) \frac{S_0^{v/2}}{2^{v/2} \Gamma(v/2)} \theta_i^{v/2-1} \exp(-\frac{1}{2} S_0 \theta_i) \\ &\propto \theta_i^{(v+2y_i)/2-1} \exp\left\{-\frac{1}{2}\theta_i(S_0 + 2t_i)\right\}. \end{aligned}$$

It is then clear from Appendix A.2 that

$$\theta_i | \theta_{-i}, S_0, y \sim S_1^{-1} \chi_{v'}^2,$$

where

$$S_1 = S_0 + 2t_i, \quad v' = v + 2y_i.$$

As for  $S_0$ , we find that if we take a  $U_0^{-1} \chi_\rho^2$  prior for  $S_0$ , then

$$\begin{aligned}
p(S_0 | \theta, y) &\propto \left[ \prod p(\theta_i | S_0) \right] p(S_0) \\
&= \left[ \prod \frac{S_0^{\nu/2}}{2^{\nu/2} \Gamma(\nu/2)} \theta_i^{\nu/2-1} \exp(-\frac{1}{2} S_0 \theta_i) \right] \frac{U_0^{\rho/2}}{2^{\rho/2} \Gamma(\rho/2)} S_0^{\rho/2-1} \exp(-\frac{1}{2} U_0 S) \\
&\propto S_0^{(\rho+k\nu)/2} \exp\{-\frac{1}{2}(U_0 + \sum \theta_i)S_0\}.
\end{aligned}$$

This is clearly a chi-squared distribution, so that

$$S_0 | \theta, y \sim U_1^{-1} \chi_{\rho'}^2$$

with

$$U_1 = U_0 + \sum \theta_i, \quad \rho' = \rho + k\nu.$$

In accordance with the general principle of the Gibbs sampler, we now take a value of  $S_0$  and then generate values of the  $\theta_i$ , then use those values to generate a value of  $S_0$ , then use this new value to generate new values of the  $\theta_i$ , and so on.

An R program to carry out this process is as follows:

```

N <- 10000
burnin <- 1000
k <- 10
y <- c(5, 1, 5, 14, 3, 19, 1, 1, 4, 22)
t <- c(94.320, 15.720, 62.880, 125.760, 5.240,
      31.440, 1.048, 1.048, 2.096, 10.480)
r <- y/t
U <- 1.0
rho <- 0.2
nu <- 1.4
S <- rep(NA, N)
S[1] <- 2.0
theta <- matrix(NA, nrow=N, ncol=k)
theta[1, ] <- rep(1.0, k)
for (j in 2:N) {
  for (i in 1:k) {
    theta[j, i] <- (S[j-1]+2*t[i])^(-1)*rchisq(1, nu+2*y[i])
  }
  S[j] <- (U+sum(theta[j,]))^(-1)*rchisq(1, rho+k*nu)
}
Strunc <- S[burnin:N]
thetatrunc <- theta[burnin:N,]
thetamean <- apply(thetatrunc, 2, mean)

```

```

thetasd <- apply(thetaTrunc, 2, sd)
thetaStats <- cbind(thetaMean, thetasd)
colnames(thetaStats) <- c("mean", "sd")
cat("\nPump means estimated as\n")
print(thetaStats)
cat("\nS has mean", mean(Strunc),
    "and s.d.", sd(Strunc), "\n")

```

The results of a run of  $N=10\,000$  (of which the first 1000 are ignored) are as follows:

System ( $i$ )	Mean	s.d.
1	0.05989735	0.02507362
2	0.10256774	0.07870071
3	0.08914099	0.03705638
4	0.11561089	0.03005465
5	0.60907691	0.31762446
6	0.60667401	0.13747133
7	0.89859718	0.72278527
8	0.89560013	0.71677873
9	1.58454532	0.75715648
10	1.99107727	0.42022407

It turns out that  $S_0$  has mean 1.849713 and s.d. 0.7906609, but this is less important.

### 9.4.3 More on the semi-conjugate prior with a normal likelihood

Recall that we said in Section 9.2 that in the case where we have observations  $x_1, x_2, \dots, x_n \sim N(\theta, \phi)$ , we say that we have a semi-conjugate or conditionally conjugate prior for  $\theta$  and  $\phi$  (which are both supposed unknown) if  $\theta \sim N(\theta_0, \phi_0)$  and  $\phi \sim S_0 \chi_v^{-2}$  independently of one another.

Conditional conjugacy allows the use of the procedure described in Section 9.3, since conditional on knowledge of  $\phi$  we know that the posterior of  $\theta$  is  $\theta | \phi \sim N(\theta_1, \phi_1)$ , where

$$\phi_1 = \{\phi_0^{-1} + (\phi/n)^{-1}\} \quad \text{and} \quad \theta_1 = \phi_1 \{\theta_0/\phi_0 + \bar{x}/(\phi/n)\}$$

(cf. Section 2.3). Then, conditional on knowledge of  $\theta$  we know that the posterior of  $\phi$  is  $\phi | \theta \sim (S_0 + S) \chi_{\nu+n}^{-2}$ ,

where

$$S = \sum (x_i - \theta)^2 = S_x + n(\bar{x} - \theta)^2$$

(cf. Section 2.7).

To illustrate this case, we return to the data on wheat yield considered in the example towards the end of Section 2.13 in which  $n=12$ ,  $\bar{x} = 119$  and  $S=13\,045$  which we also considered in connection with the *EM* algorithm. We will take the same prior we used in connection with the *EM* algorithm, so that we will take  $\phi \sim S_0 \chi_\nu^2$  with  $S_0=2700$  and  $\nu = 11$  and (independently of  $\phi$ )  $\theta \sim N(\theta_0, \phi_0)$ , where  $\theta = 110$  and  $\phi_0 = S_0/n_0(\nu - 2) = 2700/(15 \times 9) = 20$ .

We can express an R program as a procedure for finding estimates of the mean and variance of the posterior distributions of  $\theta$  from the values of  $\theta$  generated. It is easy to modify this to find other information about the posterior distribution of  $\theta$  or of  $\phi$ .

```

iter <- 10 # Number of iterations of the EM algorithm
m <- 500 # Number of replications
t <- 10 # Number of iterations
n <- 12
xbar <- 119
sxx <- 13045
s0 <- 2700
nu0 <- 11
n0 <- 15
theta0 <- 110
phi0 <- s0/(n0*(nu0-2))
thetabar <- 0
phibar <- 0
phi <- sxx/(n-1) # Initialize
thetafinal <- rep(0,m)
phifinal <- rep(0,m)
for (j in 1:m) # Replicate m times
{
  for (s in 1:t) # Iterate t times
  {
    
```

```

phi1 <- 1/((1/phi0)+(n/phi))
theta1 <- phi1*((theta0/phi0)+(n*xbar/phi))
#  $\theta | \phi \sim N(\theta_1, \phi_1)$ 
theta <- theta1+sqrt(phi1)*rnorm(1)
#  $S_1 = S_0 + \sum(x(i) - \theta)^2$ 
s1 <- s0+sxx+n*(xbar-theta)*(xbar-theta)
#  $\phi | \theta \sim S_1 * \chi_{\nu_1}^{-2}$ 
phi <- s1/rchisq(1, nu0+n)
}
thetafinal[j] <- theta
phifinal[j] <- phi
}
thetabar <- mean(thetafinal)
thetavar <- var(thetafinal)
cat("Posterior for theta has mean", thetabar,
    "and variance", thetavar, "\n")

```

This procedure can be implemented using algorithms in Kennedy and Gentle (1980); an implementation in R with  $k=500$  replications of  $T=10$  iterations resulted in  $E\theta = 112.2031$  and  $V\theta = 14.357$ , so that the standard deviation is 3.789. (Note that there is no reason to think that the posterior of  $\theta$  is symmetrical in this case, so the mean as found here need not be exactly the same as the median which we found by the *EM* algorithm.) If we are prepared to modify the prior very slightly, so that  $\nu + n$  is even, it can be implemented very easily with algorithms in Press *et al.* (1986–1993).

Extensions of this methodology to deal with the general linear model are described by O'Hagan (1994, Sections 9.45 et seq.).

#### 9.4.4 The Gibbs sampler as an extension of chained data augmentation

With the chained data augmentation algorithm, we had two stages in which we estimated alternately two parameters  $\eta$  and  $z$ . The Gibbs sampler can be regarded as a multivariate extension of the chained data augmentation algorithm in which we estimate  $r$  parameters  $\theta_1, \theta_2, \dots, \theta_r$ . To use it, we take a starting point  $(\theta_1^{(0)}, \theta_2^{(0)}, \dots, \theta_r^{(0)})$  and then iterate as follows:

(a<sub>1</sub>) Choose a value  $\theta_1^{(t+1)}$  of  $\theta_1$  from  $p(\theta_1 | \theta_2^{(t)}, \theta_3^{(t)}, \dots, \theta_r^{(t)}, x)$ ; (a<sub>2</sub>) Choose a value  $\theta_2^{(t+1)}$  of  $\theta_2$  from  $p(\theta_2 | \theta_1^{(t+1)}, \theta_3^{(t)}, \dots, \theta_r^{(t)}, x)$ ; :

(a<sub>r</sub>) Choose a value  $\theta_r^{(t+1)}$  of  $\theta_r$  from  $p(\theta_r | \theta_1^{(t+1)}, \theta_2^{(t+1)}, \dots, \theta_{r-1}^{(t+1)}, x)$ .

Then values of  $\theta^{(t)} = (\theta_1^{(t)}, \dots, \theta_r^{(t)})$  move as a Markov chain, so that once they get to a particular position, in thinking of where to go next they have no memory of where they have previously been. As we remarked earlier, in many cases this will mean that values taken after a large number of iterations should have a distribution which does not depend on their starting distribution. There are complicated questions relating to the circumstances in which convergence to an equilibrium distribution occurs and, assuming that it does, the speed at which equilibrium is approached, which we will return to later.

The Gibbs sampler became widely known from the work of Geman and Geman (1984), although its roots can be traced back to Metropolis *et al.* (1953) and its use in this context was proposed by Turchin (1971). Its name arises from connections with the Gibbs distribution in statistical mechanics which are explained in Geman and Geman (op. cit.) or Ó Ruanaidh and Fitzgerald (1996, Chapter 4).

## 9.4.5 An application to change-point analysis

It is difficult to appreciate the Gibbs sampler without following through a realistic example in detail, but many real uses of the method involve distributions not dealt with in this book as well as other complications. We shall now consider in detail the case of a Poisson process with a change point which was first introduced in Section 8.1 on ‘The idea of a hierarchical model’. There are no undue complications in this example, although it necessarily takes a certain amount of work to deal with a problem involving so many parameters.

Recall that we supposed that  $x_i \sim P(\lambda)$  for  $i = 1, 2, \dots, k$  while  $x_i \sim P(\mu)$  for  $i = k + 1, \dots, n$  and that we took independent priors for the parameters  $\theta = (\lambda, \mu, k)$  such that

**a.**  $k \sim UD(1, n)$ , that is,  $k$  has a discrete uniform distribution on  $[1, n]$ ; **b.**  $\lambda = U/\gamma$ , where  $U$  has an exponential distribution of mean 1 (or equivalently a one-parameter gamma distribution with parameter 1, so that  $2U \sim \chi_2^2$ ); **c.**  $\mu = V/\delta$  where  $V$  is independent of  $U$  and has the same distribution.

Finally, we supposed that the hyperparameters  $\gamma$  and  $\delta$  had prior distributions

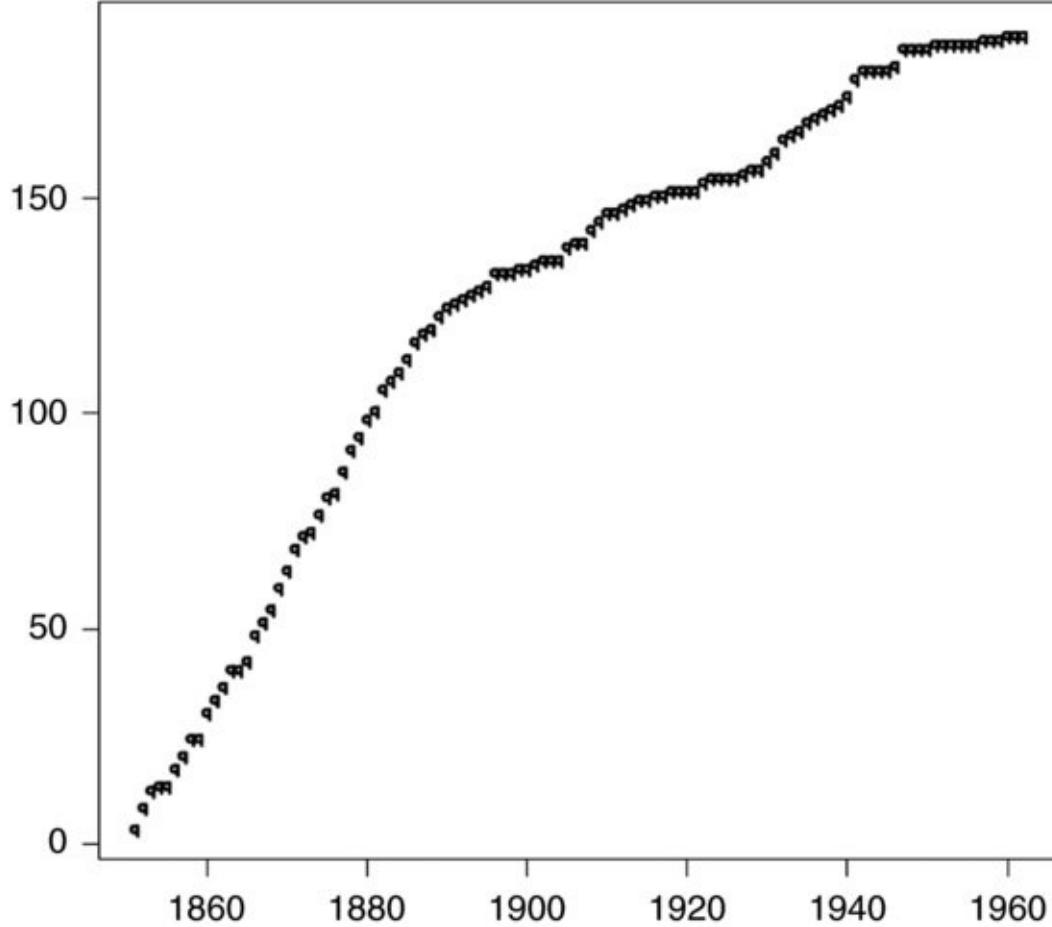
which were (independently of one another) multiples of chi-squared. In what follows, we shall take  $p(\gamma) \propto \gamma^{-1} \exp(-\gamma)$  and  $p(\delta) \propto \delta^{-1} \exp(-\delta)$ . These priors are improper and it would be better to take  $p(\gamma) \propto \gamma^{-1/2} \exp(-\gamma)$  and  $p(\delta) \propto \delta^{-1/2} \exp(-\delta)$  as other authors have done. On the other hand, our choice enables us to restrict attention to gamma distributions with integer parameters and many computer systems make it easier to generate values from such distributions.

The actual data on the numbers of coal-mining disasters in the  $n=112$  years from 1851 to 1962 inclusive were as follows:

4	5	4	1	0	4	3	4	0	6	3	3	4	0	2	6	3	3	5	4
5	3	1	4	4	1	5	5	3	4	2	5	2	2	3	4	2	1	3	2
1	1	1	1	1	3	0	0	1	0	1	1	0	0	3	1	0	3	2	2
0	1	1	1	0	1	0	1	0	0	0	2	1	0	0	0	1	1	0	2
2	3	1	1	2	1	1	1	2	4	2	0	0	0	1	4	0	0	0	0
1	0	0	0	0	0	1	0	0	1	0	0								

The plot of the cumulative number of disasters against time given as [Figure 9.3](#) appears to fall into two approximately straight sections, suggesting that disasters occurred at a roughly constant rate until a ‘change-point’  $k$  when the average rate suddenly decreased. It, therefore, seems reasonable to try to fit the aforementioned model. If we are interested in the distribution of  $k$ , it rapidly appears that direct integration of the joint distribution presents a formidable task.

[Figure 9.3](#) Cumulative number of disasters.



We can, however, proceed to find the conditional distribution of each of the parameters  $\lambda$ ,  $\mu$  and  $k$  and the hyperparameters  $\gamma$  and  $\delta$  given all of the others and the data  $x$ . For this purpose, we shall find it helpful to write

$$S(i) = \sum_{j=1}^i x_j \quad \text{and} \quad S = S(n).$$

From the facts that we have a prior  $\lambda = \frac{1}{2}\gamma^{-1}\chi_2^2$  and data  $x_i \sim P(\lambda)$  for  $i=1, 2, \dots, k$ , it follows as in Section 3.4 on ‘The Poisson distribution’ that

$$\lambda | \mu, \gamma, \delta, k, x \sim \frac{1}{2}\gamma_1^{-1}\chi_\alpha^2,$$

where

$$\gamma_1 = \gamma + k, \quad \alpha = 2 + 2S(k)$$

or equivalently

$$\lambda | \mu, \gamma, \delta, k, x \sim G(1 + S(k)) / (\gamma + k)$$

has a two-parameter gamma distribution (see Appendix A). Similarly,

$$\mu | \lambda, \gamma, \delta, k, x \sim \frac{1}{2}\delta_1^{-1}\chi_\beta^2,$$

where

$$\delta_1 = \delta + (n - k), \quad \beta = 2 + 2(S - S(k))$$

or equivalently

$$\mu | \lambda, \gamma, \delta, k, x \sim G(1 + S - S(k)) / (\delta + n - k).$$

Next, we note that we have a prior  $p(\gamma) = \gamma^{-1} \exp(-\gamma)$  and (since  $\lambda = \gamma^{-1} U$ ) a likelihood for  $\gamma$  given  $\lambda$  of  $\gamma \exp(-\lambda\gamma)$ . It follows just as in Section 2.7 on ‘Normal variance’ that the posterior is  $p(\gamma | \lambda, \mu, \delta, k, x) \propto \exp(-(\lambda + 1)/\gamma)$  which we can express as

$$\gamma | \lambda, \mu, \delta, k, x \sim \frac{1}{2}(\lambda + 1)^{-1} \chi_2^2$$

or equivalently

$$\gamma | \lambda, \mu, \delta, k, x \sim (\lambda + 1)^{-1} G(1).$$

Similarly,

$$\delta | \lambda, \mu, \gamma, k, x \sim \frac{1}{2}(\mu + 1)^{-1} \chi_2^2$$

or equivalently

$$\delta | \lambda, \mu, \gamma, k, x \sim (\mu + 1)^{-1} G(1).$$

The distribution of  $k$  is not from any well-known parametric family. However, we can easily see that

$$p(x | \lambda, \mu, k) = \prod_{i=1}^k \frac{\lambda^{x_i}}{x_i!} \exp(-\lambda) \prod_{i=k+1}^n \frac{\mu^{x_i}}{x_i!} \exp(-\mu),$$

so that (after dividing by  $p(x | \lambda, \mu, n)$ ) we can write the log-likelihood as

$$L(k | \lambda, \mu, \gamma, x) = \log l(k | \lambda, \mu, \gamma, x) = k(\mu - \lambda) + S(k)(\log \lambda - \log \mu).$$

Since, the prior for  $k$  is  $UD(1, n)$ , it follows that

$$p(k | \lambda, \mu, \gamma, \delta, x) = \frac{l(k | \lambda, \mu, \gamma, \delta, x)}{\sum_{k=1}^n l(k | \lambda, \mu, \gamma, \delta, x)} \quad (k = 1, 2, \dots, n).$$

Once we have these conditional distributions, we can proceed using the Gibbs sampler. The starting point of each iteration is not of great importance in this case, so we use plausible guesses of  $\gamma = 1$ ,  $\delta = 1$  and  $k$  an integer somewhere near the middle of the range (it might seem better to pick  $k$  randomly between 1 and  $n$ , but there are computational difficulties about this approach and in any case we expect from inspection of the data that  $k$  is somewhere in the middle). A further indication of how to proceed is given in the R program later, which uses only uniformly distributed random variates and one parameter gamma variates with integer parameters; the resulting analysis is slightly different from that in Carlin *et al.* (1992), but our restriction to integer parameters for the gamma variates makes the computation slightly simpler. We approximate the posterior density of  $k$  by  $Pr(k)$  which is found as the average over  $m$  replications of the current distribution  $p(k)$  for  $k$  after  $T$  iterations. We also find the mean of this posterior distribution.

```

m <- 2 # Number of replications
t <- 15 # Number of iterations
startyear <- 1851 # First year data available
x <- c(
  4,5,4,1,0,4,3,4,0,6,3,3,4,0,2,6,3,3,5,4,5,3,1,4,4,
  1,5,5,3,4,2,5,2,2,3,4,2,1,3,2,1,1,1,1,1,3,0,0,1,0,
  1,1,0,0,3,1,0,3,2,2,0,1,1,1,0,1,0,1,0,0,0,2,1,0,0,
  0,1,1,0,2,2,3,1,1,2,1,1,1,2,4,2,0,0,0,1,4,0,0,0,
  1,0,0,0,0,1,0,0,1,0,0)
n <- length(x) # Number of years of data available
S <- cumsum(x)
endyear <- startyear+n-1
plot(startyear:endyear,S)
pp <- rep(0,n)
L <- rep(0,n)
pp <- rep(0,n)
for (j in 1:m) # Replicate m times
{
  k <- 1+floor(n*runif(1)) # k random in [1,n]
  gamma <- 1
  delta <- 1 # Initialize gamma=delta=1
  for (s in 1:t) # Iterate t times
  {
    # Sample  $\lambda|x, \mu, \gamma, \delta, k$ 
    lambda <- rgamma(1, 1+S[k]) / (gamma+k)
    # Sample  $\mu|x, \lambda, \gamma, \delta, k$ 
    mu <- rgamma(1, 1+S[n]-S[k])
    mu <- mu/(delta+n-k)
    # Sample  $\gamma|x, \lambda, \mu, \delta, k$ 
    gamma <- rgamma(1, 1)/(1+lambda)
    # Sample  $\delta|x, \lambda, \mu, \gamma, k$ 
  }
}

```

```

delta <- rgamma(1,1)/(1+mu)
# Find L(k|λ, μ, γ, δ) for k = 0 to n-1
for (k in 1:n)
{
  L[k] <- exp((mu-lambda)*k+
    (log(lambda)-log(mu))*S[k])
}
# Find p(k|x, λ, μ, γ, δ)
p <- L/sum(L)
cumprob <- cumsum(p)
# Pick U at random between 0 and 1
U <- runif(1)
# Sample k|x, λ, μ, γ, δ
for (i in 1:n)
  if ((cumprob[i]<U) (U<=cumprob[i+1])) k <- i
}# End iteration
pp <- pp + p/m
}# End replication
# Find posterior density and mean of k
year <- startyear:endyear
meandate <- sum((year+0.5)*pp)
# Print out results
cat("Mean is",meandate,"\n")
barplot(pp,names.arg=startyear:endyear)

```

A run in R with  $m=100$  replicates of  $t=15$  iterations resulted in a posterior mean for  $k$  of 1889.6947, that is, 11 September 1889. A slightly different answer results from the use of slightly more realistic priors and the use of gamma variates with non-integer parameters. It is possible that the change is associated with a tightening of safety regulations as a result of the Coal Mines Regulation Act (50 & 51 Vict. c. 58 [1887]) which came into force on 1 January 1888.

## 9.4.6 Other uses of the Gibbs sampler

The Gibbs sampler can be used in a remarkable number of cases, particularly in connection with hierarchical models, a small subset of which are indicated later. Further examples can be found in Bernardo *et al.* (1992 and 1996), Tanner (1996), Gatsonis *et al.* (1993 and 1995), Gelman *et al.* (2004), Carlin and Louis (2000) and Gilks *et al.* (1996).

### 9.4.6.1 Radiocarbon dating

Buck *et al.* (1996, Sections 1.1.1, 7.5.3 and 8.6.2) describe an interesting application to archaeological dating. Suppose that the dates of organic material from three different samples  $s$  are measured as  $x_1$ ,  $x_2$  and  $x_3$  and it is believed that,

to a reasonable approximation,  $x_i \sim N(\theta_i, \phi_i)$ , where  $\theta_i$  is the age of the sample and  $\phi_i$  can be regarded as known. It can be taken as known that the ages are positive and less than some large constant  $k$ , and the time order of the three samples it is also regarded as known on archaeological grounds, so that

$$\theta_1 < \theta_2 < \theta_3. \text{ Thus, } p(\boldsymbol{\theta}) = p(\theta_1, \theta_2, \theta_3) = \begin{cases} c & (0 < \theta_1 < \theta_2 < \theta_3 < k) \\ 0 & (\text{otherwise}), \end{cases}$$

so that

$$p(\boldsymbol{\theta} | \mathbf{x}) \propto \begin{cases} \exp\left\{-\frac{1}{2} \sum (x_i - \theta_i)^2 / \phi_i\right\} & (0 < \theta_1 < \theta_2 < \theta_3 < k) \\ 0 & (\text{otherwise}). \end{cases}$$

It is tricky to integrate this density to find the densities of the three parameters, but the *conditional* densities are easy to deal with. We have, for example,

$$p(\theta_1 | \theta_2, \theta_3, \mathbf{x}) \propto \begin{cases} \exp\left\{-\frac{1}{2} \sum (x_i - \theta_i)^2 / \phi_i\right\} & (0 < \theta_1 < \theta_2) \\ 0 & (\text{otherwise}) \end{cases}$$

and we can easily sample from this distribution by taking a value  $\theta_1 \sim N(x_1, \phi_1)$  but, if a value outside the interval  $(0, \theta_2)$  occurs, rejecting it and trying again. Knowing this and the other conditional densities, we can approximate the posteriors using the Gibbs sampler.

In fact, the above description is simplified because the mean of measurements made if the true age is  $\theta$  is  $\mu(\theta)$  where  $\mu(\theta)$  is a *calibration function* which is generally non-linear for a variety of reasons described in Buck *et al.* (op. cit., Section 9.2), although it is reasonably well approximated by a known piecewise linear function, but the method is essentially the same.

### 9.4.6.2 Vaccination against Hepatitis B

The example in Section 8.1 on ‘The idea of a hierarchical model’ about vaccination can be easily seen to be well adapted to the use of the Gibbs sampler since all the required conditional densities are easy to sample from. The details are discussed in Gilks *et al.* (1996).

### 9.4.6.3 The hierarchical normal model

We last considered the hierarchical normal model in Subsection 9.2.4 in connection with the *EM* algorithm, and there we found that the joint posterior

$$\log p(\mu, \phi, \psi | \boldsymbol{\theta}, \mathbf{x}) = -\frac{1}{2}(N + 2)\log \phi - \frac{1}{2} \sum_i \sum_j (x_{ij} - \theta_i)^2 / \phi - \frac{1}{2}r \log \psi - \frac{1}{2} \sum_i (\theta_i - \mu)^2 / \psi$$

takes the form

and this is also the form of  $\log p(\theta, \mu, \phi, \psi | \mathbf{x})$  (assuming uniform reference priors

for the  $\theta_i$  ), so that, treating the other parameters as constants, it is easily deduced that  $p(\theta | \mu, \phi, \psi, x) \propto \exp(-\frac{1}{2}S/\phi - \frac{1}{2}S'/\psi)$ ,

where

$$S = \sum_i \sum_j (x_{ij} - \theta_i)^2 \quad \text{and} \quad S' = \sum_i (\theta_i - \mu)^2$$

from which it is easily deduced that

$$\theta_i | \mu, \phi, \psi, x \sim N(\widehat{\theta}_i, \psi_i),$$

where

$$\psi_i = \{1/\psi + n_i/\phi\}^{-1} \quad \text{and} \quad \widehat{\theta}_i = \psi_i \{ \mu/\psi + n_i x_{i*}/\phi \}.$$

Similarly,

$$\phi | \mu, \psi, \theta, x \sim S \chi_N^{-2}$$

$$\mu | \phi, \psi, \theta, x \sim N(\bar{\theta}, \psi/r)$$

$$\psi | \mu, \phi, \theta, x \sim S' \chi_{r-2}^{-2}.$$

We now see that these conditional densities are suitable for the application of the Gibbs sampler.

#### 9.4.6.4 The general hierarchical model

Suppose observations  $p(x|\theta)$  have a distribution depending on parameters  $\theta$  which themselves have distributions depending on hyperparameters  $\mu$ . In turn, we suppose that the distribution of the hyperparameters  $\mu$  depends on hyperhyperparameters  $\nu$ . Then  $p(x, \theta, \mu, \nu) = p(x | \theta, \mu, \nu) p(\theta | \mu, \nu) p(\mu | \nu) p(\nu)$

and the joint density of any subset can be found by appropriate integrations. Using the formula for conditional probability, it is then easily seen after some

$$p(x | \theta, \mu, \nu) = p(x | \theta)$$

$$p(\theta | \mu, \nu, x) = p(\theta | \mu, x)$$

$$p(\mu | \theta, \nu, x) = p(\mu | \theta, \nu)$$

cancellations that  $p(\nu | \theta, \mu, x) = p(\nu | \mu)$ .

In cases where  $\nu$  is known it can simply be omitted from the above formulae. In cases where there are more stages these formulae are easily extended (see Gelfand and Smith, 1990). These results were implicitly made use of in the example on the change point in a Poisson process.

It should be clear that in cases where we assume conjugate densities at each stage the conditional densities required for Gibbs sampling are easily found.

#### 9.4.7 More about convergence

While it is true, as stated earlier, that processes which move as a Markov chain

are frequently such that after a large number of steps the distribution of the state they then find themselves in is more or less independent of the way they started off, this is *not* invariably the case, and even when it is the case, the word ‘large’ can conceal a number of assumptions.

An instructive example arises from considering a simple case in which there is no data  $x$ . Suppose we have parameters  $\theta = (\theta_1, \theta_2)$  such that

$$p(\theta) = \begin{cases} c & (0 \leq \theta_1 \leq 1; 0 \leq \theta_2 \leq 1); \\ 1 - c & (2 \leq \theta_1 \leq 3; 2 \leq \theta_2 \leq 3); \\ 0 & (\text{otherwise}) \end{cases}$$

for  $0 < c < 1$  (cf. O’Hagan, 1994, Example 8.8). Now observe that if  $0 \leq \theta_1 \leq 1$ , then  $\theta_2 | \theta_1 \sim U(0, 1)$ , while if  $2 \leq \theta_1 \leq 3$ , then  $\theta_2 | \theta_1 \sim U(2, 3)$ , and similarly for the distribution of  $\theta_1$  given  $\theta_2$ . It should then be clear that a Gibbs sampler cannot converge to an equilibrium distribution, since the distribution after *any* number of stages will depend on the way the sampling started. This is also an example in which the conditional distributions do not determine the joint distribution, which is most clearly seen from the fact that  $c$  does not appear in the conditional distributions.

The trouble with the aforementioned example is that the chain is *reducible* in that the set of possible states  $\theta$  can be partitioned into components such that if the chain starts in one component, then it stays in that component thereafter. Another possible difficulty arises if the chain is *periodic*, so that if it is in a particular set of states at one stage it will only return to that set of states after a fixed number of stages. For there to be satisfactory convergence, it is necessary that the movement of the states should be irreducible and aperiodic. Luckily, these conditions are very often (but not always) met in cases where we would like to apply the Gibbs sampler. For further information, see Smith and Roberts (1993).

The question as to how many stages are necessary before  $\theta^{(t)}$  has a distribution which is sufficiently close to the equilibrium distribution  $p(\theta)$  whatever the starting point  $\theta^{(0)}$  is a difficult one and one which has not so far been addressed in this book. Basically, the rate of approach to equilibrium is determined by the degree of correlation between successive values of  $\theta^{(t)}$  and can be very slow if there is a high correlation.

Another artificial example quoted by O’Hagan (1994, Example 8.11) arises if  $\theta = (\theta_1, \theta_2)$  where  $\theta_1$  and  $\theta_2$  take only the values 0 and 1 and have joint density

$$p(\theta_1 = 0, \theta_2 = 0) = p(\theta_1 = 1, \theta_2 = 1) = \frac{1}{2}\pi$$

$$p(\theta_1 = 1, \theta_2 = 0) = p(\theta_1 = 0, \theta_2 = 1) = \frac{1}{2}(1 - \pi)$$

for some  $0 < \pi < 1$ , so that

$$p(\theta_2 = \theta_1 | \theta_1) = \pi \quad \text{and} \quad p(\theta_2 \neq \theta_1 | \theta_1) = 1 - \pi$$

and similarly for  $\theta_1$  given  $\theta_2$ . It is not hard to see that if we operate a Gibbs sampler with these conditional distributions then

$$p(\theta_2^{(t)} = 1 | \theta_1^{(t-1)} = 1) = \pi^2 + (1 - \pi)^2 \quad \text{and} \quad p(\theta_2^{(t)} = 1 | \theta_1^{(t-1)} = 0) = 2\pi(1 - \pi)$$

from which it is easy to deduce that  $\pi_t = p(\theta^{(t)} = 1)$  satisfies the difference equation  $\pi_t = \alpha\pi_{t-1} + \beta$ , where  $\alpha = \pi^2 + (1 - \pi)^2 - 2\pi(1 - \pi) = (2\pi - 1)^2$  and  $\beta = 2\pi(1 - \pi)$ . It is easily checked (using the fact that  $1 - \alpha = 2\beta$ ) that the solution of this equation for a given value of  $\pi_0$  is  $\pi_t = \alpha^t\pi_0 + \frac{1}{2}(1 - \alpha^t)$

and clearly  $\pi_t \rightarrow \frac{1}{2}$  as  $t \rightarrow \infty$ , but if  $\pi$  is close to 0 or 1, then  $\alpha$  is close to 1 and convergence to this limit can be very slow. If, for example,  $\pi = 0.999$ , so that  $\alpha = (2\pi - 1)^2 = 0.996\,004$  then  $p_{128} = 0.2$  if  $\pi_0 = 0$  but  $p_{128} = 0.8$  if  $\pi_0 = 1$ . A further problem is that the series may have appeared to have converged even though it has not – if the process starts from  $(0, 0)$ , then the number of iterations until  $\theta^{(t)}$  is not equal to  $(0, 0)$  is geometrically distributed with mean  $(1 - \pi^2)^{-1} = 500$ .

Problems of convergence for Markov chain Monte Carlo methods such as the are very difficult and by no means fully solved, and there is not room here for a full discussion of the problems involved. A fuller discussion can be found in O'Hagan (1994, Section 8.65 et seq.), Gilks *et al.* (1996, especially Chapter 7), Robert (1995), Raftery and Lewis (1996) and Cowles and Carlin (1996). Evidently, it is necessary to look at distributions obtained at various stages and see whether it appears that convergence to an equilibrium distribution is taking place, but even in cases where this is apparently the case, one should be aware that this could be an illusion. In the examples in this chapter, we have stressed the derivation of the algorithms rather than diagnostics as to whether we really have achieved equilibrium.

## 9.5 Rejection sampling

### 9.5.1 Description

An important method which does not make use of Markov chain methods but which helps to introduce the Metropolis–Hastings algorithm is *rejection sampling* or *acceptance-rejection sampling*. This is a method for use in connection with a density  $p(\theta) = f(\theta)/K$  in the case where the normalizing constant  $K$  is quite possibly unknown, which, as remarked at the beginning of

this chapter, is a typical situation occurring in connection with posterior distributions in Bayesian statistics. To use this method, we need to assume that there is a *candidate density*  $h(\theta)$  from which we can simulate samples and a constant  $c$  such that  $f(\theta) \leq ch(\theta)$ . Then, to obtain a random variable  $\tilde{\theta}$  with density  $p(\theta)$  we proceed as follows:

1. Generate a variate  $Y$  from the density  $h(\theta)$ ;
2. Generate a value  $U \sim U(0, 1)$  which is uniformly distributed on  $(0, 1)$ ;
3. Then if  $Uf(Y)/ch(Y) < 1$  we define  $\tilde{\theta} = Y$ ; otherwise go back to step 1.

In the discrete case we can argue as follows. In  $N$  trials we get the value  $\theta$  on  $Nh(\theta)$  occasions, of which we retain it  $Nf(\theta)/c$  times, which is proportional to  $f(\theta)$ . Similar considerations apply in the continuous case; a formal proof can be found in Ripley (1987, Section 3.2) or Gamerman and Lopes (2011, Section 5.1).

## 9.5.2 Example

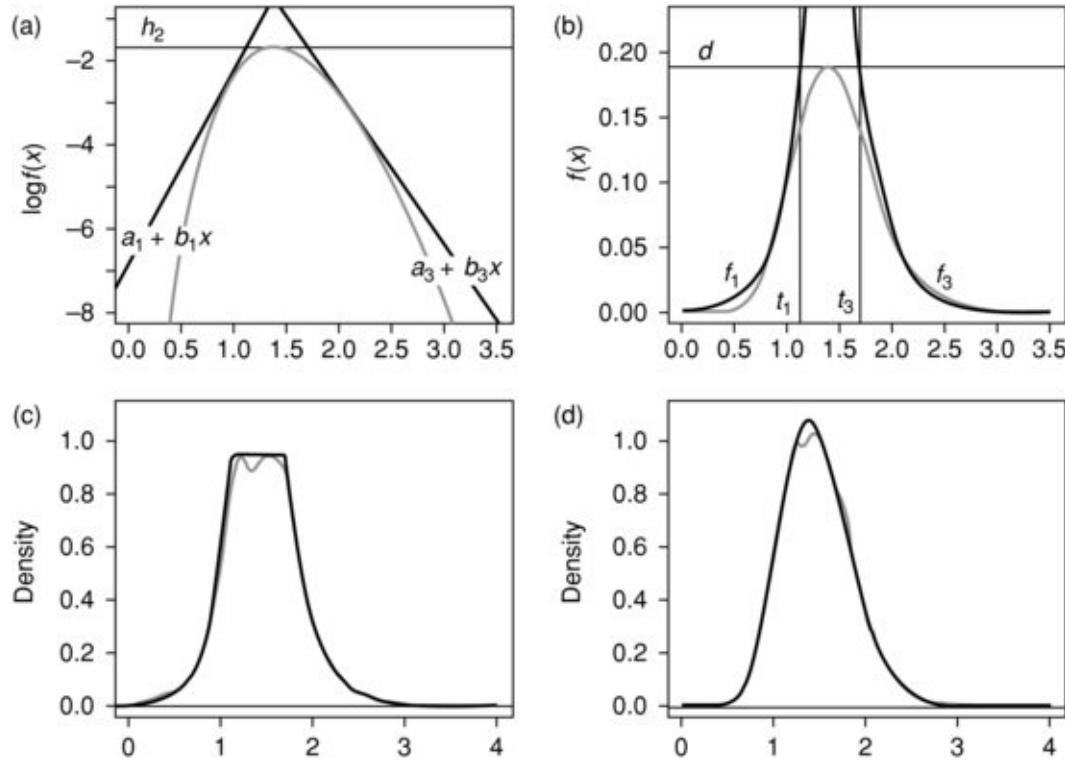
As a simple example, we can use this method to generate values  $X$  from a beta distribution with density  $f(x) = x^{\alpha-1}(1-x)^{\beta-1}/K$ , where  $K = B(\alpha, \beta)$  in the case where  $\alpha \geq 1$  and  $\beta \geq 1$  (in fact, we know that in this case  $K = B(\alpha, \beta)$  and  $X \sim Be(\alpha, \beta)$ ). We simply note that the density has a maximum at  $\text{mode}(X) = (\alpha - 1)/(\alpha + \beta - 2)$  (see Appendix A), so that we can take  $c = (\alpha - 1)^{\alpha-1}(\beta - 1)^{\beta-1}/(\alpha + \beta - 2)^{\alpha+\beta-2}$  and  $h(x)=1$ , that is,  $Y \sim U(0, 1)$  as well as  $U \sim U(0, 1)$ .

## 9.5.3 Rejection sampling for log-concave distributions

It is often found that inference problems simplify when a distribution is log-concave, that is when the logarithm of the density function is concave (see Walther, 2009). In such a case, we can fit a piecewise linear curve over the density as depicted in [Figure 9.4a](#). In that particular case, there are three pieces to the curve, but in general there could be more. Exponentiating, we find an upper bound for the density itself which consists of curves of the form  $\exp(a+bx)$ , as depicted in [Figure 9.4b](#). As these curves are easily integrated, the proportions of the area under the upper bound which fall under each of the curves are easily calculated. We can then generate a random variable with a

density proportional to the upper bound in two stages. We first choose a constituent curve  $\exp(a+bx)$  between  $x=t$  and  $x = t'$  with a probability proportional to

**Figure 9.4** Rejection sampling: (a) log density, (b) un-normalized density, (c) empirical density before rejection and (d) empirical density of v.



$$\int_t^{t'} \exp(a + bx) dx = b^{-1} (\exp(a + bt') - \exp(a + bt)) = q, \text{ say.}$$

After this, a point between  $t$  and  $t'$  is chosen as a random variable with a density proportional to  $\exp(a+bx)$ .

For the latter step, we note that such a random variable has distribution function

$$F(x) = \frac{1}{q} \int_t^x \exp(a + b\xi) d\xi = \frac{1}{qb} (\exp(a + bx) - \exp(a + bt)).$$

The inverse function of this is easily shown to be

$$x = F^{-1}(v) = \frac{1}{b} \{\log (qbe^{-a}v + e^{bt})\}.$$

If then  $V$  is uniformly distributed over  $[0, 1]$  and we write  $X=F^{-1}(V)$ , it follows that  $P(X \leq x) = P(F^{-1}(V) \leq x) = P(F(F^{-1}(V)) \leq F(x)) = P(V \leq F(x)) = F(x)$ ,

so that  $X$  has the distribution function  $F(x)$ .

[Figure 9.4c](#) show that simulation using this method produces an empirical distribution which fits closely to the upper envelope of curves we have chosen. We can now use the method of rejection sampling as described earlier to find a sample from the original log-concave density. The empirical distribution thus produced is seen to fit closely to this density in [Figure 9.4d](#).

The choice of the upper envelope can be made to evolve as the sampling proceeds; for details, see Gilks and Wild (1992).

## 9.5.4 A practical example

In Subsection 9.4.2 headed ‘An example with observed data’ which concerned pump failures, we assumed that the parameter  $v$  was known to be equal to 1.4. It would clearly be better to treat  $v$  as a parameter to be estimated in the same way that we estimated  $S$ . Now

$$p(v | \theta, y, S_0) = p(v, y | \theta, S_0) / p(y | \theta, S_0) = p(v | \theta, S_0) p(y | \theta, v, S_0) / p(y | \theta, S_0) \\ = p(v | \theta, S_0)$$

since the distribution of  $y$  given  $\theta$  does not depend on  $v$  or  $S_0$ . Consequently  $p(v | \theta, y, S_0) = p(v, \theta | S_0) / p(\theta | S_0) \propto p(v, \theta | S_0) = p(\theta | v, S_0) p(v)$ .

Unfortunately, the latter expression is not proportional to any standard family for any choice of hyperprior  $p(\nu)$ . If, however,  $p(v) = \mu^{-1} \exp(-v/\mu)$  has the form of an exponential distribution of mean  $\mu$ , we find that

$$p(v | \theta, y, S_0) \propto p(\theta | v, S_0) p(v) \\ = \left[ \prod \frac{S_0^{v/2}}{2^{v/2} \Gamma(v/2)} \theta_i^{v/2-1} \exp(-\frac{1}{2} S_0 \theta_i) \right] \mu^{-1} \exp(-v/\mu),$$

so that

$$\log p(\theta | v, S_0) p(v) = (kv/2) \log(S_0) - (kv/2) \log(2) - k \log\{\Gamma(v/2)\} \\ + (v/2 - 1) \sum (\log(\theta)) - v/\mu + \text{constant}.$$

The logarithm of this density is convex. A proof of this follows, but the result can be taken for granted. We observe that the first derivative of the density is  $(k/2) \log(S) - (k/2) \log(2) - (k/2)\psi(v/2) + \frac{1}{2} \sum \log(\theta) - 1/\mu$ ,

where

$$\psi(z) = \frac{d}{dz} \log \Gamma(z) = \frac{\Gamma'(z)}{\Gamma(z)}$$

(cf. Appendix A.7), while the second derivative is

$$-(k/4)\psi_1(v/2),$$

where  $\psi_1(z) = \psi'(z)$  is the so-called trigamma function. Since the trigamma function satisfies

$$\psi_1(z) = \sum_{n=1}^{\infty} \frac{1}{(z+n)^2}$$

(see Whittaker and Watson, 1927, Section 12.16), it is clear that this second derivative is negative for all  $z > 0$ , from which convexity follows.

We now find a piecewise linear function which is an upper bound of the density in the manner described earlier. In this case, it suffices to approximate by a three-piece function, as illustrated in [Figure 9.4a](#). This function is constructed by taking a horizontal line tangent at the zero  $n_2$  of the derivative of the log-density, where this achieves its maximum  $h_2$  and then taking tangents at two nearby points  $n_1$  and  $n_3$ . The joins between the lines occur at the points  $t_1$  and  $t_3$  where these three curves intersect. Exponentiating, this gives an upper bound for the density itself, consisting of a number of functions of the form  $\exp(a+bx)$  (the central one being a constant), as illustrated in [Figure 9.4b](#).

In this case, the procedure is very efficient and some 87–88% of candidate points are accepted. The easiest way to understand the details of this procedure is to consider the R program:

```
windows(record=T)
par(mfrow=c(2,2))
N <- 5000
mu <- 1
S <- 1.850
theta <- c(0.0626, 0.1181, 0.0937, 0.1176, 0.6115,
          0.6130, 0.8664, 0.8661, 1.4958, 1.9416)
k <- length(theta)
logf <- function(nu) {
  k*(nu/2)*log(S)-k*(nu/2)*log(2)-k*log(gamma(nu/2))+(
    (nu/2-1)*sum(log(theta))-nu/mu
  )
}
dlogf <- function(nu) {
  (k/2)*log(S)-(k/2)*log(2)-0.5*k*digamma(nu/2)+(
    0.5*sum(log(theta))-1/mu
  )
}
scale <- 3/2
f <- function(nu) exp(logf(nu))
n2 <- uniroot(dlogf, lower=0.01, upper=100)\$root
h2 <- logf(n2)
```

```

n1 <- n2/scale
h1 <- logf(n1)
b1 <- dlogf(n1)
a1 <- h1-n1*b1
n3 <- n2*scale
h3 <- logf(n3)
b3 <- dlogf(n3)
a3 <- h3-n3*b3
d <- exp(h2)
# First plot; Log density
curve(logf, 0.01, 3.5, ylim=c(-8, -1), xlab="a. Log density")
abline(h=h2)
abline(a1, b1)
abline(a3, b3)
text(0.25, h2+0.3, expression(italic(h)[2]))
text(0.3, -6,
     expression(italic(a)[1]+italic(b)[1]*italic(x)))
text(3.25, -7,
     expression(italic(a)[3]+italic(b)[3]*italic(x)))
# End of first plot
f1 <- function(x) exp(a1+b1*x)
f3 <- function(x) exp(a3+b3*x)
t1 <- (h2-a1)/b1
t3 <- (h2-a3)/b3
# Second plot; Un-normalized density
curve(f, 0.01, 3.5, ylim=c(0, 0.225),
       xlab="b. Un-normalized Density")
abline(h=d)
par(new=T)
curve(f1, 0.01, 3.5, ylim=c(0, 0.225), ylab="", xlab="")
par(new=T)
curve(f3, 0.01, 3.5, ylim=c(0, 0.225), ylab="", xlab="")
abline(v=t1)
abline(v=t3)
text(t1-0.1, 0.005, expression(italic(t)[1]))
text(t3-0.1, 0.005, expression(italic(t)[3]))

```

```

text(0.25,0.2,expression(italic(d)))
text(0.5,0.025,expression(italic(f)[1]))
text(2.5,0.025,expression(italic(f)[3]))
# End of second plot
q1 <- exp(a1)*(exp(b1*t1)-1)/b1
q2 <- exp(h2)*(t3-t1)
q3 <- -exp(a3+b3*t3)/b3
const <- q1 + q2 + q3
p <- c(q1,q2,q3)/const
cat("p =",p,"\\n")
case <- sample(1:3,N,replace=T,prob=p)
v <- runif(N)
w <- (case==1)*(1/b1)*log(q1*b1*exp(-a1)*v+1) +
      (case==2)*(t1+v*(t3-t1))+
      *(1/b3)*(log(q3*b3*exp(-a3)*v+exp(b3*t3)))
dq <- d/const
f1q <- function(x) f1(x)/const
f3q <- function(x) f3(x)/const
h <- function(x) {
  dq +
  (x<t1)*(f1q(x)-dq) +
  (x>t3)*(f3q(x)-dq)
}
# Third plot; Empirical density
plot(density(w),xlim=c(0.01,4),ylim=c(0,1.1),
      xlab="c. Empirical density before rejection",
      main="")
par(new=T)
curve(h,0.01,4,ylim=c(0,1.1),ylab="",xlab="")
# End of third plot
u <- runif(N)
nu <- w
nu[u>f(nu)/(const*h(nu))] <- NA
nu <- nu[!is.na(nu)]
cat("Acceptances",100*length(nu)/length(w),"\\n")
int <- integrate(f,0.001,100)\$value

```

```

fnorm <- function(x) f(x)/int
# Fourth plot
plot(density(nu), xlim=c(0.01,4), ylim=c(0,1.1),
      xlab=bquote(paste("d. Empirical density of ", nu)),
      main="")
par(new=T)
curve(fnorm, 0.01, 4, ylim=c(0,1.1), ylab="", xlab="")
# End of fourth plot}

```

It is possible to combine this procedure with the method used earlier to estimate the  $\theta_i$  and  $S_o$ , and an R program to do this can be found on the web. We shall not discuss this matter further here since it is in this case considerably simpler to use WinBUGS as noted in Section 9.7.

## 9.6 The Metropolis–Hastings algorithm

### 9.6.1 Finding an invariant distribution

This whole section is largely based on the clear expository article by Chib and Greenberg (1995). A useful general review which covers Bayesian integration problems more generally can be found in Evans and Swartz (1995–1996; 2000).

A topic of major importance when studying the theory of Markov chains is the determination of conditions under which there exists a stationary distribution (sometimes called an invariant distribution), that is, a distribution  $\pi(\theta)$  such that  $\pi(\phi) = \sum_{\theta} \pi(\theta) p(\phi|\theta)$ .

In the case where the state space is continuous, the sum is, of course, replaced by an integral.

In Markov Chain Monte Carlo methods, often abbreviated as MCMC methods, the opposite problem is encountered. In cases where such methods are to be used, we have a target distribution  $\pi(\theta)$  in mind from which we want to take samples, and we seek transition probabilities  $p(\phi|\theta)$  for which this target density is the invariant density. If we can find such probabilities, then we can start a Markov chain at an arbitrary starting point and let it run for a long time, after which the probability of being in state  $\theta$  will be given by the target density  $\pi(\theta)$ .

It turns out to be useful to let  $r(\theta)$  be the probability that the chain remains in

the same state and then to write  $p(\phi|\theta) = p^*(\phi|\theta) + r(\theta)\delta(\phi|\theta)$ , where  $p^*(\theta|\theta) = 0$  while

$$\delta(\theta|\theta) = 1 \quad \text{and} \quad \delta(\phi|\theta) = 0 \quad \text{for } \phi \neq \theta.$$

Note that because the chain must go *somewhere* from its present position  $\theta$  we

$$1 = \sum_{\phi} p(\phi|\theta) = \sum_{\phi} p^*(\phi|\theta) + r(\theta)$$

have for all  $\theta$ . If the equation

$$\pi(\theta) p^*(\phi|\theta) = \pi(\phi) p^*(\theta|\phi),$$

is satisfied, we say that the probabilities  $p^*(\phi|\theta)$  satisfy *time reversibility* or *detailed balance*. This condition means that the probability of starting at  $\theta$  and ending at  $\phi$  when the initial probabilities are given by  $\pi(\theta)$  is the same as that of starting at  $\phi$  and ending at  $\theta$ . It turns out that when it is satisfied then  $p(\phi|\theta)$  gives a set of transition probabilities with  $\pi(\theta)$  as an invariant density. For

$$\begin{aligned} \sum_{\theta} \pi(\theta) p(\phi|\theta) &= \sum_{\theta} \pi(\theta) p^*(\phi|\theta) + \sum_{\theta} \pi(\theta) r(\theta) \delta(\phi|\theta) \\ &= \sum_{\theta} \pi(\phi) p^*(\theta|\phi) + \pi(\phi) r(\phi) \\ &= \pi(\phi) \left\{ \sum_{\theta} p^*(\theta|\phi) \right\} + \pi(\phi) r(\phi) \\ &= \pi(\phi) \{1 - r(\phi)\} + \pi(\phi) r(\phi) \\ &= \pi(\phi). \end{aligned}$$

The result follows very similarly in cases where the state space is continuous.

The aforementioned proof shows that all we need to find the required transition probabilities is to find probabilities which are time reversible.

## 9.6.2 The Metropolis–Hastings algorithm

The Metropolis–Hastings algorithm begins by drawing from a candidate density as in rejection sampling, but, because we are considering Markov chains, the density depends on the current state of the process. We denote the candidate density by  $q(\phi|\theta)$  and we suppose that  $\sum_{\phi} q(\phi|\theta) = 1$ . If it turns out that the density  $q(y|x)$  is itself time reversible, then we need look no further. If, however, we find that  $\pi(\theta)q(\phi|\theta) > \pi(\phi)q(\theta|\phi)$

then it appears that the process moves from  $\theta$  to  $\phi$  too often and from  $\phi$  to  $\theta$  too rarely. We can reduce the number of moves from  $\theta$  to  $\phi$  by introducing a probability  $\alpha(\phi|\theta)$ , called the *probability of moving*, that the move is made. In order to achieve time reversibility, we take  $\alpha(\phi|\theta)$  to be such that the detailed balance equation  $\pi(\theta)q(\phi|\theta)\alpha(\phi|\theta) = \pi(\phi)q(\theta|\phi)$

holds and consequently

$$\alpha(\phi|\theta) = \frac{\pi(\phi)q(\theta|\phi)}{\pi(\theta)q(\phi|\theta)}.$$

We do not want to reduce the number of moves from  $\phi$  to  $\theta$  in such a case, so we take  $\alpha(\theta|\phi) = 1$ , and similarly  $\alpha(\phi|\theta) = 1$  in the case where the inequality is reversed and we have  $\pi(\theta)q(\phi|\theta) < \pi(\phi)q(\theta|\phi)$ .

It is clear that a general formula is

$$\alpha(\phi|\theta) = \min\left[\frac{\pi(\phi)q(\theta|\phi)}{\pi(\theta)q(\phi|\theta)}, 1\right],$$

so that the probability of going from state  $\theta$  to state  $\phi$  is  $p^*(\phi|\theta) = q(\phi|\theta)\alpha(\phi|\theta)$ , while the probability that the chain remains in its present state  $\theta$  is  $r(\theta) = 1 - \sum_{\phi} q(\phi|\theta)\alpha(\phi|\theta)$ .

The matrix of transition probabilities is thus given by

$$p(\phi|\theta) = p^*(\phi|\theta) + r(\theta)\delta(\phi|\theta) = q(\phi|\theta)\alpha(\phi|\theta) + \left(1 - \sum_{\phi} q(\phi|\theta)\alpha(\phi|\theta)\right)\delta(\phi|\theta).$$

The aforementioned argument assumes that the state space is discrete, but this is just to make the formulae easier to take in – if the state space is continuous the same argument works with suitable adjustments, such as replacing sums by integrals.

Note that it suffices to know the target density  $\pi(\theta)$  up to a constant multiple, because it appears both in the numerator and in the denominator of the expression for  $\alpha(\phi|\theta)$ . Further, if the candidate-generating density  $q(\phi|\theta)$  is symmetric, so that  $q(\phi|\theta) = q(\theta|\phi)$ , the  $\alpha(\phi|\theta)$  reduces to  $\alpha(\phi|\theta) = \min\left[\frac{\pi(\phi)}{\pi(\theta)}, 1\right]$ .

This is the form which Metropolis *et al.* (1953) originally proposed, while the general form was proposed by Hastings (1970).

We can summarize the Metropolis–Hastings algorithm as follows:

1. Sample a *candidate point*  $\theta^*$  from a *proposal distribution*<sup>1</sup>  $q(\theta^*|\theta^{(t-1)})$ .
2. Calculate

$$\alpha = \min\left\{\frac{p(\theta^*)q(\theta^{(t-1)}|\theta^*)}{p(\theta^{(t-1)})q(\theta^*|\theta^{(t-1)})}, 1\right\};$$

3. Generate a value  $U \sim U(0, 1)$  which is uniformly distributed on  $(0, 1)$ ;

4. Then if  $U \leq \alpha$  we define  $\theta^{(t)} = \theta^*$ ; otherwise we define  $\theta^{(t)} = \theta^{(t-1)}$

5. Return the sequence  $\{\theta^{(1)}, \theta^{(2)}, \dots, \theta^{(n)}\}$ .

Of course we ignore the values of  $\theta^{(i)}$  until the chain has converged to

equilibrium. Unfortunately, it is a difficult matter to assess how many observations need to be ignored. One suggestion is that chains should be started from various widely separated starting points and the variation within and between the sampled draws compared; see Gelman and Rubin (1992). We shall not go further into this question; a useful reference is Gamerman and Lopes (2011, Section 5.4).

### 9.6.3 Choice of a candidate density

We need to specify a candidate density before a Markov chain to implement the Metropolis–Hastings algorithm. Quite a number of suggestions have been made, among them:

- *Random walk chains* in which  $q(\phi|\theta) = q_1(\phi - \theta)$ , so that  $\phi$  takes the form  $\theta + \psi$  where  $\psi$  has the density  $q_1(\psi)$ . The density  $q_1(\psi)$  is usually symmetric, so that  $q_1(\psi) = q_1(-\psi)$ , and in such cases the formula for the probability of moving reduces to  $\alpha(\phi|\theta) = \min\left[\frac{\pi(\phi)}{\pi(\theta)}, 1\right]$ .
- *Independence chains* in which  $q(\phi|\theta) = q_2(\phi)$  does not depend on the current state  $\theta$ . Although it sounds at first as if the transition probability does not depend on the current state  $\theta$ , in fact it does through the probability of moving which is  $\alpha(\phi|\theta) = \min\left[\frac{\pi(\phi)/q_2(\phi)}{\pi(\theta)/q_2(\theta)}, 1\right]$ .

In cases where we are trying to sample from a posterior distribution, one possibility is to take  $q_2(\theta)$  to be the prior, so that the ratio becomes a ratio of likelihoods  $\alpha(\phi|\theta) = \min\left[\frac{l(\phi)}{l(\theta)}, 1\right]$ .

Note that there is no need to find the constant of proportionality in the posterior distribution.

- *Autoregressive chains* in which  $\phi = a + b(\theta - a) + \lambda$ , where  $\lambda$  is random with density  $q_3$ . It follows that  $q(\phi|\theta) = q_3(\phi - a - b(\theta - a))$ . Setting  $b=-1$  produces chains that are reflected about the point  $a$  and ensures negative correlation between successive elements of the chain.
- A Metropolis–Hastings acceptance–rejection algorithm due to Tierney (1994) is described by Chib and Greenberg (1995). We will not go into this algorithm here.

### 9.6.4 Example

Following Chib and Greenberg (1995), we illustrate the method by considering ways of simulating the bivariate normal distribution  $N(\mu, \Sigma)$ , where we take  $\mu = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$ ,  $\Sigma = \begin{pmatrix} 1 & 0.9 \\ 0.9 & 1 \end{pmatrix}$ .

If we define

$$P = \begin{pmatrix} 1 & 0.9 \\ 0 & \sqrt{1 - 0.9^2} \end{pmatrix},$$

so that  $P^T P = \Sigma$  (giving the Choleski factorization of  $\Sigma$ ), it is easily seen that if we take

$$\phi = \mu + PZ,$$

where the components of  $Z$  are independent standard normal variates, then  $\phi$  has the required  $N(\mu, \Sigma)$  distribution, so that the distribution is easily simulated without using Markov chain Monte Carlo methods. Nevertheless, a simple illustration of the use of such methods is provided by this distribution. One possible set-up is provided by a random walk distribution  $\phi = \theta + Z$ , where the components  $Z_1$  and  $Z_2$  of  $Z$  are taken as independent uniformly distributed random variables with  $Z_1 \sim U(-0.75, 0.75)$  and  $Z_2 \sim U(-1, 1)$  and taking the

$$\alpha(\phi|\theta) = \min \left\{ \frac{\exp[-\frac{1}{2}(\phi - \mu)^T \Sigma^{-1}(\phi - \mu)]}{\exp[-\frac{1}{2}(\theta - \mu)^T \Sigma^{-1}(\theta - \mu)]}, 1 \right\}.$$

probability of a move as

We can program this in R as follows

```

n <- 600
mu <- c(1, 2)
n <- 600
Sigma <- matrix(c(1, 0.9, 0.9, 1), nr=2)
InvSigma <- solve(Sigma)
lik <- function(theta){opencurle
  as.numeric(exp(-(theta-mu)^%*%InvSigma%*%(theta-mu)/2))
\closecurle
alpha <- function(theta, phi) min(lik(phi)/lik(theta), 1)
theta1 <- matrix(0, n, 2)
theta1[1, ] <- mu
k1 <- 0
for (i in 2:n){opencurle
  theta <- theta1[i-1, ]
  Z <- c(0.75, 1) - c(1.5*runif(1), 2*runif(1))
  phi <- theta + Z
  alpha(theta, phi)
  if (runif(1) < alpha(theta, phi)) {
    theta1[i, ] <- phi
  } else {
    theta1[i, ] <- theta
  }
}
theta1
}
```

```

k <- rbinom(1,1,alpha(theta,phi))
k1 <- k1 + k
theta1[i,] <- theta + k*(phi-theta)
\closecurle
plot(theta1,xlim=c(-4,6),ylim=c(-3,7))}
```

An autoregressive chain can be constructed simply by replacing the line

```
phi <- theta + Z
```

by the line

```
phi <- mu - (theta - mu) + Z
```

and otherwise proceeding as before.

## 9.6.5 More realistic examples

A few further examples of the use of the Metropolis–Hastings algorithm areas follows:

- Carlin and Louis (2000, Example 5.7) consider the data on a number of flour beetles exposed to varying amounts of gaseous carbon disulphide ( $\text{CS}_2$ ) and attempt to fit a generalization (suggested by Prentice, 1976) of the logistic model we shall consider in Section 9.8.
- Chib and Greenberg (1994) consider an application to inference for regression models with  $\text{ARMA}(p, q)$  errors.
- Geman and Geman (1984) showed how the Metropolis algorithm can be used in image restoration, a topic later dealt with by Besag (1986).
- The algorithm can also be used to solve the ‘travelling salesman problem’. A nice web reference for this is

<http://hermetic.nofadz.com/misc/ts3/ts3demo.htm>

## 9.6.6 Gibbs as a special case of Metropolis–Hastings

Gibbs sampling can be thought of as a special case of the Metropolis–Hastings algorithm in which all jumps from  $\theta^{(t-1)}$  to  $\theta^*$  are such that  $\theta^{(t-1)}$  and  $\theta^*$  differ only in one component. Moreover, if  $j$  is the component currently being updated and we write  $\theta_{-j}^{(t-1)} = (\theta_1^{(t)}, \dots, \theta_{j-1}^{(t)}, \theta_{j+1}^{(t)}, \dots, \theta_r^{(t)})$  for the vector  $\theta^{(t-1)}$  omitting the  $j$ th component, then, since we choose a value  $\theta_j^{(t)}$  of  $\theta_j$  from the density

$p(\theta_j^{(t)} | \theta_{-j}^{(t-1)}, x)$ , it follows from the usual rules for conditional probability that  $p(\theta^*)/q(\theta^* | \theta^{(t-1)}) = p(\theta^{(t-1)})/q(\theta^{(t-1)} | \theta^{(t-1)}) = p(\theta_{-j}^{(t-1)} | x)$ .

We can thus conclude that

$$\alpha = \min \left\{ \frac{p(\theta^*) q(\theta^{(t-1)} | \theta^*)}{p(\theta^{(t-1)}) q(\theta^* | \theta^{(t-1)})}, 1 \right\}$$

is identically equal to unity. In this way, the Gibbs sampler emerges as a particular case of the Metropolis–Hastings algorithm in which *every* jump is accepted.

### 9.6.7 Metropolis within Gibbs

In the Gibbs sampler as we originally described it, the process always jumps in accordance with the transition probability  $p(\theta_j | \theta_{-j}^{(t-1)}, x)$ . A modification, referred to as Metropolis within Gibbs or the Metropolized Gibbs sampler, is to sample a candidate point  $\theta_j^*$  *different* from  $\theta_j^{(t)}$  with probability  $\frac{p(\theta_j^* | \theta_{-j}^{(t-1)}, x)}{1 - p(\theta_j^{(t-1)} | \theta_{-j}^{(t-1)}, x)}$ .

The value  $\theta_j^{(t-1)}$  is then replaced by  $\theta_j^*$  with probability  $\alpha = \min \left\{ \frac{1 - p(\theta_j^{(t-1)} | \theta^{(t-1)})}{1 - p(\theta_j^* | \theta^{(t-1)})}, 1 \right\}$ .

This modification is statistically more efficient, although slightly more complicated, than the usual Gibbs sampler; see Liu (1996, 2001). It is employed under certain circumstances by WinBUGS, which is described later.

## 9.7 Introduction to WinBUGS and OpenBUGS

### 9.7.1 Information about WinBUGS and OpenBUGS

The package BUGS (Bayesian Inference Using Gibbs Sampling) grew from a statistical research project at the MRC Biostatistics Unit in Cambridge, but now is developed jointly with the Imperial College School of Medicine at St Mary's,

London. It is now most widely used in its Windows form, which is known as WinBUGS, and it has become the most popular means for numerical investigations of Bayesian inference. The software and documentation can be freely copied provided that it is noted that it is copyright © MRC Biostatistics Unit 1995.

We will only give a very brief overview of its use here, but with this introduction it is hoped that you will be able to get started in using WinBUGS employing its online help system. Much valuable information, including information on how to download the package (which is currently free) can be obtained from the website

<http://www.mrc-bsu.cam.ac.uk/bugs/>

More recently a program called OpenBUGS has been developed which from the user's point of view is more or less indistinguishable from WinBUGS while having the slight advantage that there is no need to bother with a key. It is freely available from

<http://www.openbugs.info/w/>

OpenBUGS is copyright © Free Software Foundation 1989, 1991.

Information installing running OpenBUGS or WinBUGS from a Mac can be found on the website associated with this book.

## 9.7.2 Distributions in WinBUGS and OpenBUGS

WinBUGS and OpenBUGS use a simple syntax which (luckily but not coincidentally) is very similar to that of R to describe statistical models. Most of the standard distributions listed in the Appendix A are available in WinBUGS and OpenBUGS (there are exceptions such as Behrens' distribution and there are some such as the F distribution where a transformation has to be employed and others such as the hypergeometric distribution where a certain amount of cunning has to be employed).

The parameterization used by WinBUGS and OpenBUGS is not quite the same as that used in this book. We have denoted a normal distribution by  $N(\mu, \phi)$  where  $\phi$  is the variance, but WinBUGS and OpenBUGS use `dnorm(mu, tau)` where tau is the precision, that is, the reciprocal of the variance. Similarly, we have written  $G(\alpha, \beta)$  for the two-parameter gamma distribution with density proportional to  $x^{\alpha-1} \exp(-x/\beta)$  whereas WinBUGS and OpenBUGS use

`dgamma(alpha,delta)`, where  $\delta$  is the reciprocal of  $\beta$ . It should also be noted that WinBUGS will *not* deal with improper priors, although, of course, we can use proper priors that are in practice very close to improper priors, so that for a location parameter  $\theta$  we can use a prior `dnorm(0, 0.0001)` which is close to a uniform prior over the real line, or for a scale parameter  $\psi$  we can use a prior `dgamma(0.0001, 0.0001)` which is close to  $p(\psi) \propto 1/\psi$  over the positive reals.

### 9.7.3 A simple example using WinBUGS

Let us begin by considering the example due to Casella and George which we considered in Section 9.4. It will be seen that the code bears a close resemblance to the code in that section. To run this example, double click on the WinBUGS icon and either click `File > New` or press `CTRL` and `N` simultaneously. A new window will appear into which the code can be typed as follows:

```
model;
{
  y ~ dbin(pi,n)
  pi ~ dbeta(newalpha,newbeta)
  newalpha <- y + alpha
  newbeta <- n - y + beta
}
```

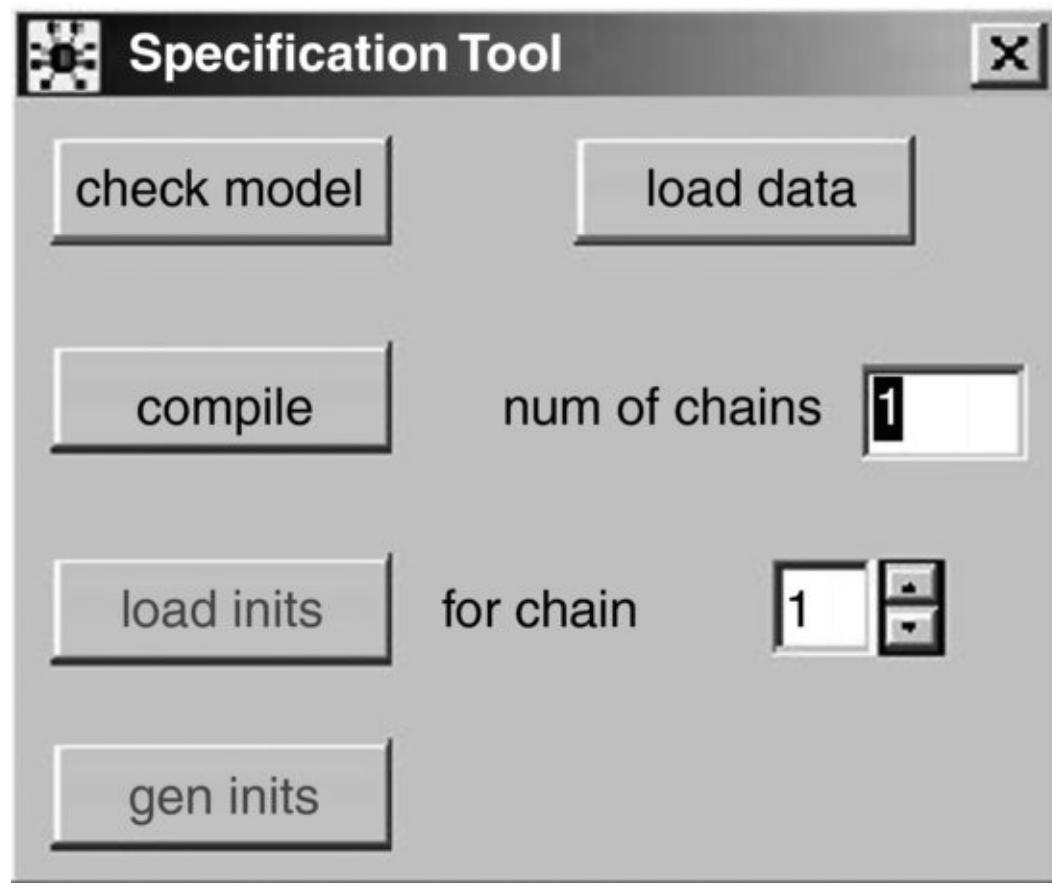
Alternatively, you can store the code on a file and open that by `File > Open` or equivalently by press `CTRL` and `O` simultaneously. Naturally, it is also necessary to include the data, so you should leave a line or two and then add the following:

```
data;
list(n=16,alpha=2,beta=4)
```

The next thing to do is to go `Model > Specification` (or equivalently press `ALT` and then `M`). A window (referred to as the *Specification Tool*) depicted in [Figure 9.5](#) will then appear. Click on check `model` and, if a model has been correctly specified, the message `model is syntactically correct` will appear in an almost unnoticeable place at the bottom of the WinBUGS window. If the model has *not* been correctly specified, then the computer emits a beep and a rather cryptic error message appears in the same place at the bottom of the WinBUGS window (in fact, the difficulty of interpreting these error messages is one of the few problems encountered with WinBUGS). Assuming that the model *is* correct, the next thing to do is to load the data, which is done by selecting the line with the word `list` in it below the word `data`; (or at least the first part of that line)

and clicking on load and to compile the program by clicking on compile. In many cases, it is necessary to specify initial values, which we can do in very much the same way we specify the data by adding a couple of lines giving reasonable starting points. In this case, we might put

**Figure 9.5** Specification Tool.

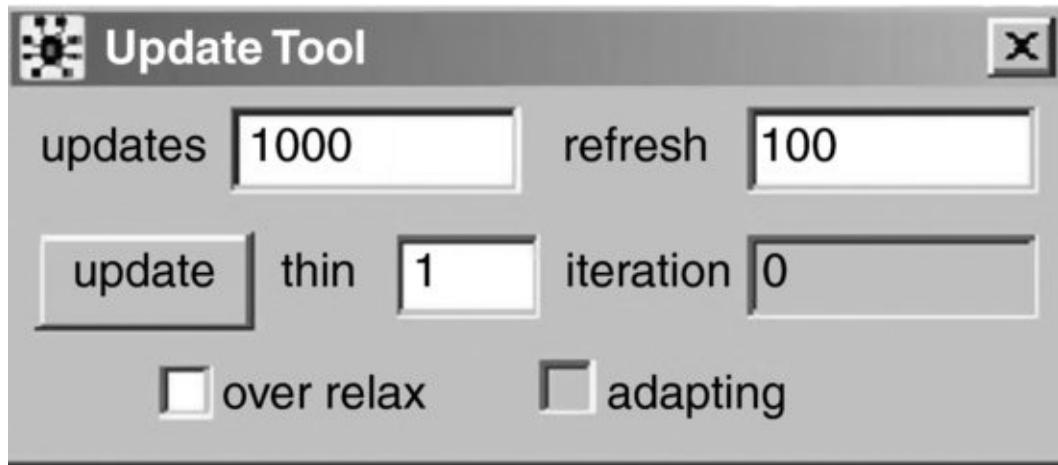


```
inits;  
list(pi=0.5, y=8)
```

but in the example we are considering (as well as in many others) WinBUGS can be left to choose initial values itself by clicking gen inits. The Specification Tool can be closed down at this point.

You then start the Markov chain going by going Model > Update (or equivalently ALT, then M, then u) and at which point a window (referred to as the *Update Tool*) as depicted in [Figure 9.6](#) should appear. Clicking on update will cause the chain to go through a large number of states each time. In a simple example, the default number of states is likely (but not certain) to be enough. After you have done this, you should not close down this tool, because you will need it again later.

**Figure 9.6** Update Tool.



The updating tool will cause the Markov chain to progress through a series of states, but will not store any information about them, so once you have gone through an initial ‘burn-in’ so that (you hope) the chain has converged to equilibrium, you need to state which parameters you are interested in and what you want to know about them. You can do this by going `Inference > Samples` or by pressing ALT and then n (at least provided you no longer have text selected on your data input window). A window (referred to as the *Sample Monitor Tool*) depicted in [Figure 9.7](#) will then appear. In the example, we are considering, you should enter y against node. The word set will then be highlighted and you should click on it. The space by node will again be blank, allowing you to enter further parameters in cases where you are interested in knowing about several parameters.

**Figure 9.7** Sample Monitor Tool.



You should now cause the Markov chain to run through more states by using the Update Tool again, perhaps clicking update several times or altering the number against update to ensure that the averaging is taken over a large number of states. Then go back to the Sample Monitor Tool. You can select one of the parameters you have stored information about by clicking on the downward arrow to the right of node or by typing its name in the window there. Various statistics about the parameter or parameters of interest are then available. You can, for example, get a bar chart rather like the barplot obtained by R by entering y against node and then clicking on density.

## 9.7.4 The pump failure example revisited

WinBUGS is well suited to dealing with the example we considered in Sections 9.4 and 9.5 of pump failure data. A suitable program is as follows:

```

model
{
  for (i in 1:k) {
    a[i] ~ dchisqr(nu);
    theta[i] <- a[i]/S0;
    lambda[i] <- theta[i]*t[i]
    Y[i] ~ dpois(lambda[i]);
  }
  nu ~ dexp(1.0);
  S ~ dchisqr(1.0);
}
data;
list(k=10, Y=c(5,1,5,14,3,19,1,1,4,22),
      t=c(94.320,15.720,62.880,125.760, 5.240,
           31.440,1.048,1.048,2.096,10.480))
inits;
list(a=c(1,1,1,1,1,1,1,1,1),nu=2,S0=2)

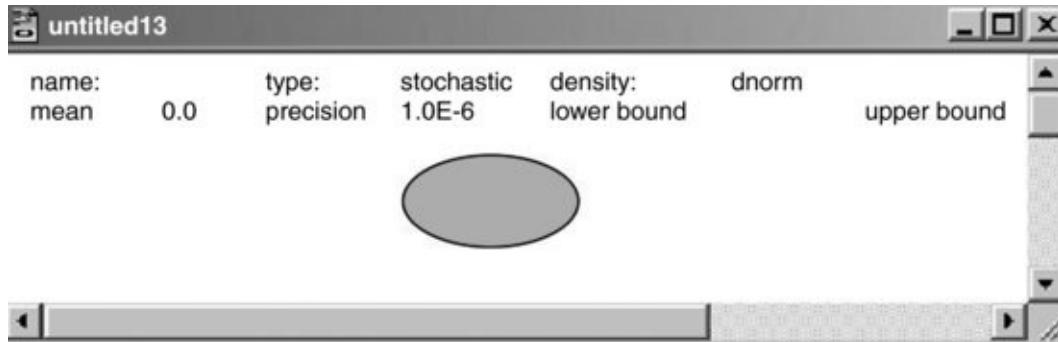
```

## 9.7.5 DoodleBUGS

The inventors of WinBUGS recommend a graphical method of building up models. To start a Doodle go Doodle > New or press ALT and then D (followed by pressing RETURN twice). You should then see a blank window. Once you click

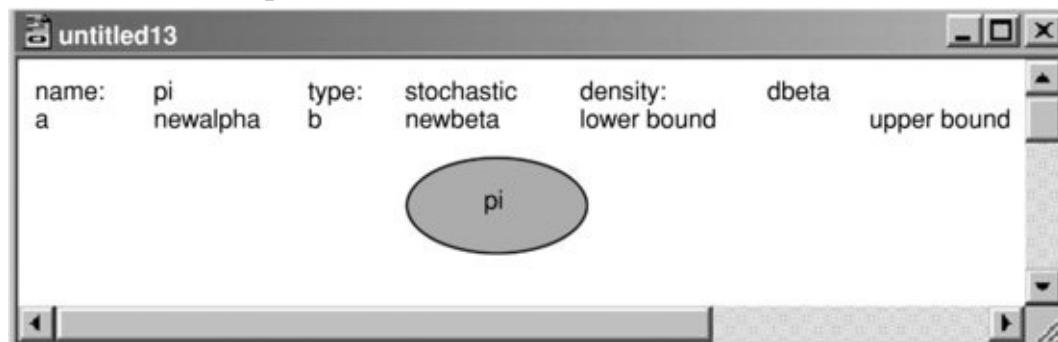
anywhere on this window, you should see a picture like the one in [Figure 9.8](#).

[Figure 9.8](#) Doodle Screen.



You can then adjust the parameters, so for example you can get a representation of the node  $\pi$  in our simple example by altering the screen (beginning by clicking on density and altering that from the menu offered to  $d\beta$ ) to look as in [Figure 9.9](#).

[Figure 9.9](#) Doodle for  $\pi$ .

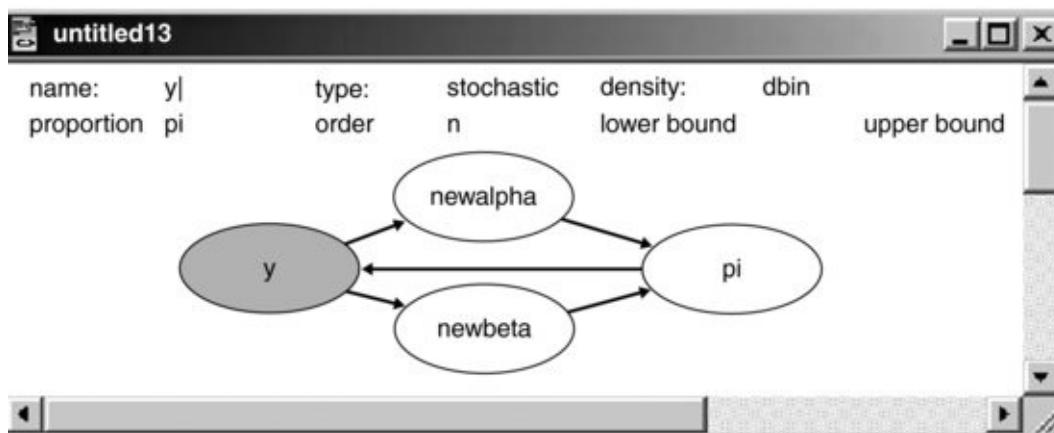


Clicking somewhere else on the window starts another node, until eventually you have nodes for  $\pi$  and  $y$  (which have type set to stochastic) stochastic nodes) and for  $newalpha$  and  $newbeta$  (which have type set to logical and value set to  $y + alpha$  and  $n - y + beta$ , respectively). If you click on a previously formed node, then it is highlighted and you can alter its parameters. The fact that the parameters of one node are determined by the value of another is indicated by edges. Such an edge can be created when one node is highlighted by holding down CTRL and clicking on another node. The dependence of a node of type logical on another node will be indicated by a double arrow, while dependence on a stochastic node will be indicated by a single arrow. Nodes can be removed by pressing Delete while they are highlighted and edges can be removed as they were created (by ensuring that the node towards which the arrow comes is highlighted and then holding down CTRL and clicking on the

node from which the arrow comes).

With the simple example we have been considering, the end result looks as indicated in [Figure 9.10](#) (if the node for  $y$  is highlighted). You can then go Doodle > New or press ALT, then D, and then w or w to produce the code for the model as given earlier.

**Figure 9.10** Doodle for the Casella–George example.



Working with DoodleBUGS is likely to be slower than writing the code directly, but if you do work this way you are less likely to end up with mystifying messages in the bottom left-hand corner of the WinBUGS window.

Of course, there are many more things you need to know if you want to work seriously with WinBUGS. The point of this subsection is to give you the self-confidence to get started with very simple examples, from which point you should be able to find out more for yourself. The program has a comprehensive online help system which is reasonably easy to follow. There is also a useful web resource that has been provided by Gillian Raab and can be found (among other places) at

[http://www-users.york.ac.uk/\\$\sim\\$pm11/bayes/winbugsinfo/raab.htm](http://www-users.york.ac.uk/$\sim$pm11/bayes/winbugsinfo/raab.htm)

## 9.7.6 coda

The acronym coda stands for Convergence Diagnostic and Output Analysis software, and is a menu-driven set of R functions which serves as an output processor for the BUGS software. It may also be used in conjunction with Markov chain Monte Carlo output from a user's own programs, providing the output is formatted appropriately (see the coda manual for details). Using coda it is possible to compute convergence diagnostics and statistical and graphical summaries for the samples produced by the Gibbs sampler.

The inventors of WinBUGS remark, ‘Beware: MCMC sampling can be dangerous!’ and there is no doubt that they are right. You need to be extremely careful about assuming convergence, especially when using complex models, and for this reason it is well worth making use of the facilities which coda provides, although we shall not describe them here.

## 9.7.7 R2WinBUGS and R2OpenBUGS

It is now possible to run WinBUGS or OpenBugs directly from R. To do this, it may be necessary, particularly in the case of WinBUGS, to run R as administrator, which is done by right clicking the icon for R and choosing Run as administrator. Suppose we wish to run the program considered in Sections 9.4 and 9.7 concerning pumps in a nuclear plant. We first need a file called, for example, pumpsmodel.txt which contains the description of the model, in this case

```
model;
{
  y ~ dbin(pi,n)
  pi ~ dbeta(newalpha,newbeta)
  newalpha <- y + alpha
  newbeta <- n - y + beta
}
```

We can then type into R

```
library(R2WinBUGS)
windows(record=T)
data <- list(k=10, Y=c(5,1,5,14,3,19,1,1,4,22),
             t=c(94.320,15.720,62.880,125.760,5.240,
                 31.440,1.048,1.048,2.096,10.480))
inits <- function() {
  list(a=c(1,1,1,1,1,1,1,1,1,1), nu=2, S=2)
}
pumps.sim <- bugs(data, inits,
                   model.file="pumpsmodel.txt",
                   parameters=c("theta","S"),
                   n.chains=3,n.iter=20000)
print(pumps.sim,digits=3)
```

```
plot(pumps.sim)
```

The results of a run with WinBUGS were as follows:

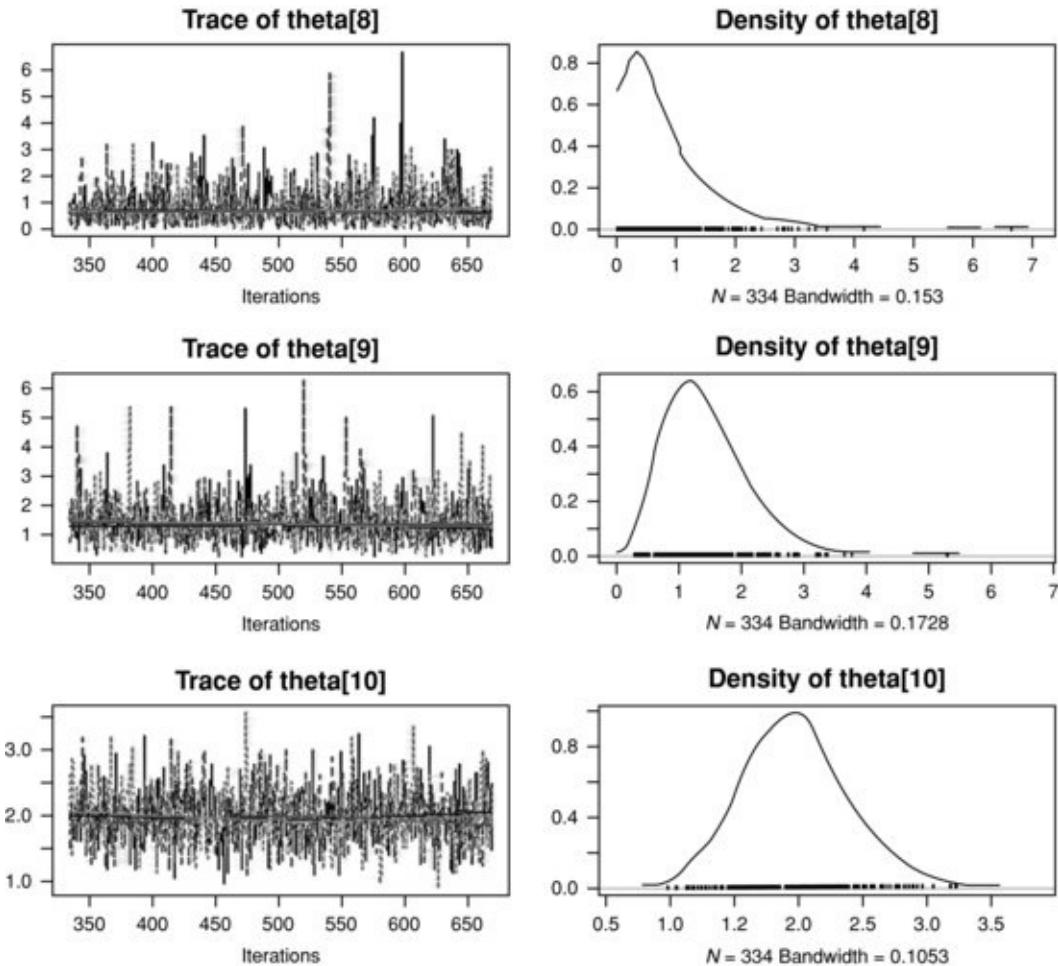
System ( $i$ )	mean	s.d.
1	0.058	0.025
2	0.102	0.079
3	0.088	0.038
4	0.116	0.032
5	0.602	0.325
6	0.610	0.138
7	0.829	0.664
8	0.827	0.705
9	1.472	0.727
10	1.978	0.409

It turns out that  $S_0$  has mean 2.269 and s.d. 1.151, but this is less important. We note that the results for the  $\theta_i$  are very similar to those we found earlier in Section 9.4, although those for  $S$  are not quite so close.

If we modify the call to bugs by adding in codaPkg=T, it is possible to investigate matters further using the package coda. A simple example of this is as follows:

```
.....
pumps.sim <- bugs(data, inits,
  model.file="pumpsmodel.txt",
  parameters=c("theta","S"),
  n.chains=3,n.iter=20000,
  codaPkg=T)
codaobject <- read.bugs(pumps.sim)
plot(codaobject)
```

**Figure 9.11** Plot of codaobject.



Part of the resulting plot can be seen in [Figure 9.11](#). If you type `coda.menu()`, then R will respond

CODA startup menu

- 1: Read BUGS output files
- 2: Use an mcmc object
- 3: Quit

the response 2 followed by choosing `codaobject` as the name of saved object will allow further detailed analysis.

If you are running OpenBUGS you should use `library (R2OpenBUGS)` instead of `library(R2WinBUGS)`. The package R2OpenBUGS will doubtless soon be available from the same source (CRAN) as other user-contributed packages for R, but for the time being can be found at <http://openbugs.info/w/UserContributedCode>.

The manual is also available on the web site associated with this book.

## 9.8 Generalized linear models

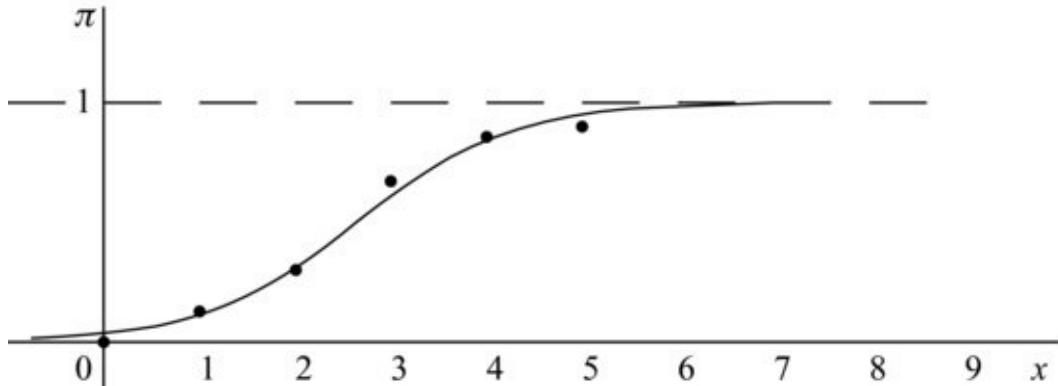
### 9.8.1 Logistic regression

We shall illustrate this technique by considering some data from an experiment carried out by R. Norrell quoted by Weisberg (1985, Section 12.2). The matter of interest in this experiment was the effect of small electrical currents on farm animals, with the eventual goal of understanding the effects of high-voltage powerlines on livestock. The experiment was carried out with seven cows, and six shock intensities, 0, 1, 2, 3, 4 and 5 milliamps (shocks on the order of 15 milliamps are painful for many humans; see Dalziel *et al.* (1941)). Each cow was given 30 shocks, five at each intensity, in random order. The entire experiment was then repeated, so each cow received a total of 60 shocks. For each shock the response, mouth movement, was either present or absent. The data as quoted give the total number of responses, out of 70 trials, at each shock level. We ignore cow differences and differences between blocks (experiments).

Current (milliamps) $x$	Number of responses $y$	Number of trials $n$	Proportion of responses $p$
0	0	70	0.000
1	9	70	0.129
2	21	70	0.300
3	47	70	0.671
4	60	70	0.857
5	63	70	0.900

Our intention is to fit a smooth curve like the one shown in [Figure 9.12](#) through these points in the same way that a straight line is fitted in ordinary regression. The procedure here has to be different because proportions, which estimate probabilities, have to lie between 0 and 1.

[Figure 9.12](#) Logistic curve  $\text{logit}\pi = -3.301 + 1.246x$ .



We define

$$\text{logit } \pi = \log \left( \frac{\pi}{1 - \pi} \right),$$

and then observe that a suitable family of curves is given by

$$\text{logit } \pi = \beta_0 + \beta_1 x$$

or equivalently

$$\pi = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}.$$

We have a binomial likelihood

$$l(\beta_0, \beta_1 | y) = \prod \pi_i^{y_i} (1 - \pi_i)^{n_i - y_i}$$

(ignoring the binomial coefficients) with corresponding log-likelihood

$$L(\beta_0, \beta_1 | y) = \sum y_i \ln \pi_i + \sum (n_i - y_i) \ln (1 - \pi_i),$$

where the binomial probabilities  $\pi_i$  are related to the values of the explanatory variable  $x_i$  by the above logistic equation. We then consider how to choose values of  $\beta_0$  and  $\beta_1$ , so that the chosen curve best fits the data. The obvious general principle used by classical statisticians is that of maximum likelihood, and a simple iteration shows that with the data in the example, the maximum likelihood values are  $\hat{\beta}_0 = -3.301$  and  $\hat{\beta}_1 = 1.246$ .

There are nowadays many books which deal with logistic regression, usually from a classical standpoint, for example Kleinbaum (1994) and Hosmer and Lemeshow (1989).

In this book, we have tried to show how most of the standard statistical models can be fitted into a Bayesian framework, and logistic regression is no exception. All we need to do is to give  $\beta_0$  and  $\beta_1$  suitable prior distributions, noting that while we have no idea about the sign of  $\beta_0$ , we would expect  $\beta_1$  to be positive. Unfortunately, we cannot produce a model with conjugate priors, but we are able to use Gibbs sampling, and a suitable program for use in WinBUGS is as follows.

```

model;
{
  for (i in 1:6) {
    y[i] ~ dbin(pi[i],n[i])
    logit(pi[i]) < - beta0 + beta1*x[i]
  }
  beta0 ~ dnorm(0,0.001)
  beta1 ~ dgamma(0.001,0.001)
}
data;
list(y=c(0,9,21,47,60,63),
      n=c(70,70,70,70,70,70),
      x=c(0,1,2,3,4,5))
inits;
list(beta0=0,beta1=1)

```

A run of 5000 after a burn-in of 1000 resulted in a posterior mean of  $-3.314$  for  $\beta_0$  and of  $1.251$  for  $\beta_1$ . Unsurprisingly, since we have chosen relatively uninformative priors, these are close to the maximum likelihood estimates.

Of course, a Bayesian analysis allows us to incorporate any prior information, we may have by taking different priors for  $\beta_0$  and  $\beta_1$ .

## 9.8.2 A general framework

We note that in Section 9.7, the parameter  $\pi$  of the binomial distribution is not itself linearly related to the explanatory variable  $x$ , but instead we have a relationship of the form  $g(\pi) = \beta_0 + \beta_1 x$ ,

where in the case of logistic regression the function  $g$ , which is known as the *link function* is the logit function.

This pattern can be seen in various other models. For example, Gelman *et al.* (2004, Section 16.5) model the number  $y_{ep}$  of persons of ethnicity  $e$  stopped by the police in precinct  $p$  by defining  $n_{ep}$  as the number of arrests by the Division of Criminal Justice Services (DCJS) of New York State for that ethnic group in that precinct and then supposing that  $y_{ep}$  has a Poisson distribution  $y_{ep} \sim P(\lambda_{ep})$ , where

$$\log(\lambda_{ep}) = \log(n_{ep}) + \alpha_e + \beta_p + \varepsilon_{ep},$$

where the coefficients  $\alpha_e$  control for the ethnic group and the  $\beta_p$  adjust for

variation among precincts. Writing  $g(\lambda) = \log(\lambda)$  and taking explanatory variates such as  $\delta_{e3}$  where  $\delta_{33} = 1$  and  $\delta_{e3} = 0$  for  $e \neq 3$

this is a similar relationship to the one we encountered with logistic regression.

Similarly, a standard model often used for the  $2 \times 2$  table we encountered in 5.6 is to suppose that the entries in the four cells of the table each have Poisson distributions. More precisely, we suppose that the mean number in the top left-hand corner has some mean  $\mu$ , that numbers in the second column have a mean  $\alpha$  times bigger than corresponding values in the first column, and that numbers in the second row have means  $\beta$  times bigger than those in the first row (where, of course,  $\alpha$  and  $\beta$  can be either greater or less than unity). It follows that the means in the four cells are  $\mu$ ,  $\mu\alpha$ ,  $\mu\beta$  and  $\mu\alpha\beta$  respectively. In this case we can regard the explanatory variables as the row and column in which any entry occurs and the Poisson distributions we encounter have means  $\lambda_{ij}$  satisfying the equation  $\lambda_{ij} = \log(\mu) + \log(\alpha)\delta_{i1} + \log(\beta)\delta_{j2}$ ,

where

$$\delta_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}.$$

Further discussion of generalized linear models can be found in Dobson (2002) or McCullagh and Nelder (1989), and in a Bayesian context in Gelman *et al.* (2004, Chapter 16) and Dey *et al.* (2000).

## 9.9 Exercises on Chapter 9

1. Find the value of  $\int_0^1 e^x dx$  by crude Monte Carlo integration using a sample size of  $n=10$  values from a uniform distribution  $U(0, 1)$  taken from tables of random numbers [use, e.g. groups of random digits from Lindley and Scott (1995, Table 27) or Neave (1978, Table 8.1)]. Repeat the experiment ten times and compute the overall mean and the sample standard deviation of the values you obtain. What is the theoretical value of the population standard deviation and how does the value you obtained compare with it?

2. Suppose that, in a Markov chain with just two states, the probabilities of going from state  $i$  to state  $j$  in one time unit are given by the entries of the matrix

$$A = \begin{pmatrix} 1/4 & 3/4 \\ 1/2 & 1/2 \end{pmatrix}$$

in which  $i$  represents the row and  $j$  the column. Show that the probability of getting from state  $i$  to state  $j$  in  $t$  time units is given by the  $t$ th power of the matrix  $A$  and that

$$A^t = \begin{pmatrix} 2/5 & 3/5 \\ 2/5 & 3/5 \end{pmatrix} + (-\frac{1}{4})^t \begin{pmatrix} 3/5 & -3/5 \\ -2/5 & 2/5 \end{pmatrix}.$$

Deduce that, irrespective of the state the chain started in, after a long time it will be in the first state with probability  $2/5$  and in the second state with probability  $3/5$ .

3. Smith (1969, Section 21.10) quotes an example on genetic linkage in which we have observations  $x = (x_1, x_2, x_3, x_4)$  with cell probabilities  $(\frac{1}{4} + \frac{1}{4}\eta, \frac{1}{4}\eta, \frac{1}{4}(1 - \eta), \frac{1}{4}(1 - \eta) + \frac{1}{4})$ .

The values quoted are  $x_1=461$ ,  $x_2=130$ ,  $x_3=161$  and  $x_4=515$ . Divide  $x_1$  into  $y_0$  and  $y_1$  and  $x_4$  into  $y_4$  and  $y_5$  to produce augmented data  $y = (y_0, y_1, y_2, y_3, y_4, y_5)$  and use the *EM* algorithm to estimate  $\eta$ .

4. Dempster *et al.* (1977) define a generalized *EM* algorithm (abbreviated as a *GEM* algorithm) as one in which  $Q(\theta^{(t+1)}, \theta^{(t)}) \geq Q(\theta^{(t)}, \theta^{(t)})$ . Give reasons for believing that *GEM* algorithms converge to the posterior mode.

5. In question 16 in Chapter 2, we supposed that the results of a certain test were known, on the basis of general theory, to be normally distributed about the same mean  $\mu$  with the same variance  $\phi$ , neither of which is known. In that

question, we went on to suppose that your prior beliefs about  $(\mu, \phi)$  could be represented by a normal/chi-squared distribution with

$$v_0 = 4, \quad S_0 = 350, \quad n_0 = 1 \quad \text{and} \quad \theta_0 = 85.$$

Find a semi-conjugate prior which has marginal distributions that are close to the marginal distributions of the normal/chi-squared prior but is such that the mean and variance are independent a priori. Now suppose as previously that 100 observations are obtained from the population with mean 89 and sample variance  $s^2=30$ . Find the posterior distribution of  $(\mu, \phi)$ . Compare the posterior mean obtained by the *EM* algorithm with that obtained from the fully conjugate prior.

**6.** A textile company weaves a fabric on a large number of looms. Four looms selected at random from those available, and four observations of the tensile strength of fabric woven on each of these looms are available (there is no significance to the order of the observations from each of the looms), and the resulting data are as follows:

Loom	Observations			
1	98	97	99	96
2	91	90	93	92
3	96	95	97	95
4	95	96	99	98

Estimate the means for each of the looms, the overall mean, the variance of observations from the same loom, and the variance of means from different looms in the population.

**7.** Write computer programs in C++ equivalent to the programs in R in this chapter.

**8.** Use the data augmentation algorithm to estimate the posterior density of the parameter  $\eta$  in the linkage model in question 3.

**9.** Suppose that  $y | \pi \sim B(n, \pi)$  and  $\pi | y \sim Be(y + \alpha, n - y + \beta)$ , where  $n$  is a Poisson variable of mean  $\lambda$  as opposed to being fixed as in Section 9.4. Use the Gibbs sampler (chained data augmentation) to find the unconditional distribution of  $n$  in the case where  $\lambda = 16$ ,  $\alpha = 2$  and  $\beta = 4$  (cf. Casella and George, 1992).

**10.** Find the mean and variance of the posterior distribution of  $\theta$  for the data in question 5 mentioned earlier using the prior you derived in answer to that question by means of the Gibbs sampler (chained data augmentation).

**11.** The following data represent the weights of  $r=30$  young rats measured weekly for  $n=5$  weeks as quoted by Gelfand *et al.* (1990), Tanner (1996,

Table 1.3 and Section 6.2.1), Carlin and Louis (2000, Example 5.6):

Rat\week	1	2	3	4	5	Rat\week	1	2	3	4	5
1	151	199	246	283	320	16	160	207	248	288	324
2	145	199	249	293	354	17	142	187	234	280	316
3	147	214	263	312	328	18	156	203	243	283	317
4	155	200	237	272	297	19	157	212	259	307	336
5	135	188	230	280	323	20	152	203	246	286	321
6	159	210	252	298	331	21	154	205	253	298	334
7	141	189	231	275	305	22	139	190	225	267	302
8	159	201	248	297	338	23	146	191	229	272	302
9	177	236	285	340	376	24	157	211	250	285	323
10	134	182	220	260	296	25	132	185	237	286	331
11	160	208	261	313	352	26	160	207	257	303	345
12	143	188	220	273	314	27	169	216	261	295	333
13	154	200	244	289	325	28	157	205	248	289	316
14	171	221	270	326	358	29	137	180	219	258	291
15	163	216	242	281	312	30	153	200	244	286	324

The weight of the  $i$ th rat in week  $j$  is denoted  $x_{ij}$  and we suppose that weight growth is linear, that is,

$$x_{ij} \sim N(\alpha_i + \beta_i j, \phi),$$

but that the slope and intercept vary from rat to rat. We further suppose that  $\alpha_i$  and  $\beta_i$  have a bivariate normal distribution, so that

$$\theta_i = \begin{pmatrix} \alpha_i \\ \beta_i \end{pmatrix} \sim N(\theta_0, \Sigma), \quad \text{where } \theta_0 = \begin{pmatrix} \alpha_0 \\ \beta_0 \end{pmatrix},$$

and thus we have a random effects model. At the third stage, we suppose that

$$C\phi \sim S_0 \chi_v^2$$

$$\theta_0 \sim N(\eta, C)$$

$$\Sigma^{-1} \sim W((\rho R)^{-1}, \rho),$$

where we have used the notation  $W(\Omega, \nu)$  for the *Wishart distribution* for a random  $k \times k$  symmetric positive definite matrix  $V$ , which has density

$$p(V | \nu, \Omega) \propto \frac{|V|^{(\nu-k-1)/2}}{|\Omega|^{\nu/2}} \exp \left[ -\frac{1}{2} \text{Trace}(\Omega^{-1}V) \right].$$

Methods of sampling from this distribution are described in Odell and Feiveson (1966), Kennedy and Gentle (1990, Section 6.5.10) and Gelfand *et al.* (1990). [This example was omitted from the main text because we have avoided use of the Wishart distribution elsewhere in the book. A slightly simpler model in which  $\Sigma$  is assumed to be diagonal is to be found as the example ‘Rats’ distributed with WinBUGS.]

Explain in detail how you would use the Gibbs sampler to estimate the posterior distributions of  $\alpha_0$  and  $\beta_0$ , and if possible carry out this procedure.

- 12.** Use the Metropolis–Hastings algorithm to estimate the posterior density of the parameter  $\eta$  in the linkage model in Sections 9.2 and 9.3 using candidate values generated from a uniform distribution on  $(0, 1)$  [cf. Tanner (1996, Section 6.5.2)].
- 13.** Write a WinBUGS program to analyze the data on wheat yield considered towards the end of Section 2.13 and in Section 9.3.
- 14.** In bioassays, the response may vary with a covariate termed the *dose*. A typical example involving a binary response is given in the following table, where  $R$  is the number of beetles killed after 5 hours of exposure to gaseous carbon disulphide at various concentrations (data from Bliss, 1935, quoted by Dobson, 2002, Example 7.3.1).

Dose $x_i$ ( $\log_{10} \text{CS}_2 \text{mg l}^{-2}$ )	Number of insects, $n_i$	Number killed, $r_i$
1.6907	59	6
1.7242	60	13
1.7552	62	18
1.7842	56	28
1.8113	63	52
1.8369	59	53
1.8610	62	61
1.8839	60	60

Fit a logistic regression model and plot the proportion killed against dose and the fitted line.

<sup>1</sup> Sometimes called a *jumping distribution*.

# 10

## Some approximate methods

### 10.1 Bayesian importance sampling

Importance sampling was first mentioned towards the start of Section 9.1. We said then that it is useful when we want to find a parameter  $\theta$  which is defined as the expectation of a function  $f(x)$  with respect to a density  $q(x)$  but we cannot easily generate random variables with that density although we *can* generate variates  $x_i$  with a density  $p(x)$  which is such that  $p(x)$  roughly approximates  $|f(x)|q(x)$  over the range of integration. Then

$$\theta = \int_a^b f(x) q(x) dx = \int_a^b f(x) \left( \frac{q(x)}{p(x)} \right) p(x) dx \cong \frac{1}{n} \sum_{i=1}^n \frac{f(x_i) q(x_i)}{p(x_i)}.$$

The function  $p(x)$  is called an *importance function*. In the words of Wikipedia, ‘Importance sampling is a variance reduction technique that can be used in the Monte Carlo method. The idea behind importance sampling is that certain values of the input random variables in a simulation have more impact on the parameter being estimated than others. If these ‘important’ values are emphasized by sampling more frequently, then the estimator variance can be reduced.’

A case where this technique is easily seen to be valuable occurs in evaluating tail areas of the normal density function (although this function is, of course, well-enough tabulated for it to be unnecessary in this case). If, for example, we wanted to find the probability  $\theta = P(Z > 4) = \Phi(-4)$  that a standard normal variate was greater than 4, so that

$$q(x) = (2\pi)^{-1/2} \exp(-x^2/2) = \phi(x) \quad f(x) = \begin{cases} 1 & (x > 4) \\ 0 & (x \leq 4) \end{cases}$$

a naïve Monte Carlo method would be to find the proportion of values of such a variate in a large sample that turned out to be greater than 4. However, even with a sample as large as  $n=10\ 000\ 000$  a trial found just 299 such values, implying that the required integral came to 0.000 030 0, whereas a reference to tables shows that the correct value of  $\theta$  is 0.000 031 67, which is, therefore, underestimated by  $5\frac{1}{2}\%$ . In fact, the number we observe will have a binomial

$B(n, \theta)$  distribution and our estimate will have a standard deviation of  $\sqrt{\theta(1-\theta)/n} = 0.000\ 001\ 78$ .

We can improve on this slightly by noting that  $\theta = \frac{1}{2}P(|Z| > 4)$ , and therefore the fact that in the same trial 619 values were found to exceed 4 in absolute value leads to an estimate of 0.000 030 95, this time an underestimate by about  $\frac{1}{4}\%$ . This time the standard deviation can easily be shown to be  $\frac{1}{4}\sqrt{n(2\theta)(1-2\theta)} = 0.000\ 001\ 26$ . There are various other methods by which slight improvements can be made (see Ripley, 1987, Section 5.1).

A considerably improved estimate can be obtained by importance sampling with  $p(x)$  being the density function of an exponential distribution  $E(1)$  of mean 1 (see Appendix A.4) truncated to the range  $(4, \infty)$ , so that

$$p(x) = e^{-x} / \int_4^\infty e^{-x} dx = e^{4-x} \quad (x > 4).$$

It is easily seen that we can generate a random sample from this density by simply generating values from the usual exponential distribution  $E(1)$  of mean 1 and adding 4 to these values. By using this truncated distribution, we avoid needing to generate values which will never be used. We can take then an importance sampling estimator

$$\frac{1}{n} \sum_{i=1}^n \frac{f(x_i)q(x_i)}{p(x_i)} = \sum_{i=1}^n (2\pi)^{-\frac{1}{2}} \exp\left[-\frac{1}{2}x_i^2 + x_i - 4\right].$$

Using this method, a sample of size only 1000 resulted in an estimate of 0.000 031 5 to three decimal places, which is within 1% of the correct value.

With any importance sampler each observed value  $x$  has density  $p(x)$  and so  $w(x)=f(x)q(x)/p(x)$  has mean

$$\mathbb{E}w(x) = \mathbb{E}\left(\frac{f(x)q(x)}{p(x)}\right) = \int \left(\frac{f(x)q(x)}{p(x)}\right) dx = \int f(x)q(x) dx = \theta$$

and variance

$$\begin{aligned} \mathcal{V}w(x) &= \mathbb{E}\left(\frac{f(x)q(x)}{p(x)} - \theta\right)^2 = \int \left(\frac{f(x)q(x)}{p(x)} - \theta\right)^2 p(x) dx \\ &= \int \left(\frac{f(x)q(x)}{p(x)}\right)^2 p(x) dx - \theta^2. \end{aligned}$$

If  $f(x) \neq 0$  for all  $x$  and we take

$$p(x) = \frac{f(x)q(x)}{\int f(\xi)q(\xi) d\xi},$$

then  $w(x)$  is constant and so has variance 0 which is necessarily a minimum. This choice, however, is impossible because to make it we would need to know the

value of  $\theta$ , which is what we are trying to find. Nevertheless, if  $fq/p$  is roughly constant then we will obtain satisfactory results. Very poor results occur when  $fq/p$  is small with a high probability but can be very large with a low probability, as, for example, when  $fq$  has very wide tails compared with  $p$ .

For a more complicated example, suppose  $x \sim G(\alpha)$  has a gamma density

$$q(x | \alpha) = \frac{1}{\Gamma(\alpha)} x^{\alpha-1} \exp(-x)$$

(see Section A.4). Then there is a family of conjugate priors of the form

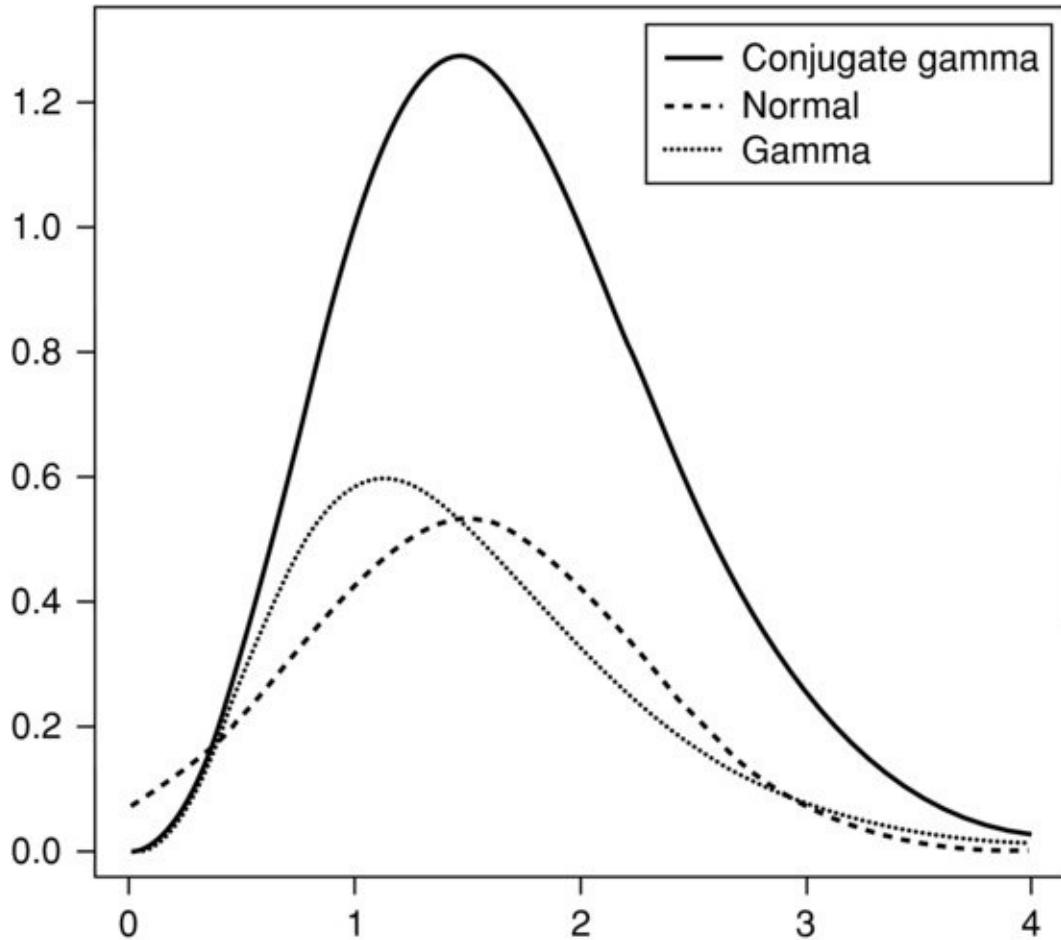
$$q(\alpha) \propto \left( \frac{1}{\Gamma(\alpha)} \right)^{\lambda} \xi^{\alpha-1},$$

where  $\lambda$  and  $\xi$  are hyperparameters, since the posterior is then equal to

$$q(\alpha | x) \propto \left( \frac{1}{\Gamma(\alpha)} \right)^{\lambda+1} (x\xi)^{\alpha-1}$$

(we write  $q$  rather than  $p$  in accordance with our notation for importance sampling). This does not come from any of the well-known distributions and the constant of proportionality is unknown, although it can be approximated by normal or gamma densities. In [Figure 10.1](#), we show the unnormalized density when  $\xi = 1$  and  $\lambda = 2$ . It appears symmetric about a value near 1.5 with most of the values between 0.75 and 2.25, so it seems reasonable to use as an importance function an approximation such as  $N(1.5, 0.75)$  or, since it is asymmetric, a gamma density with the same mean and variance, so a  $G(4, 0.375)$  distribution, and these two are also illustrated in the figure. There is a discrepancy in scale because the constant of proportionality of the density we are interested in is unknown.

[Figure 10.1](#) The conjugate gamma distribution.



While in the middle of the range either of these distributions fits reasonably well, the gamma approximation is considerably better in the tails. We can begin by finding the constant of proportionality  $k = \int_0^\infty q(\alpha) d\alpha$  using the importance sampling formula with  $f(x)=1$ , and  $p(x)$  taken as a gamma density with  $\alpha = 4$  and  $\beta = 0.375$ . Use of this method with a sample size of 10 000 resulted in a value for the constant of 2.264 where the true value is approximately 2.267. We can use the same sample to find a value of the mean, this time taking  $f(x)=x/k$  and finding a value of 1.694 where the true value is 1.699 (we would have taken a different importance function if we were concerned only with this integral, but when we want results for both  $f(x)=1$  and  $f(x)=x/k$  it saves time to use the same sample). In the same way we can obtain a value of 0.529 for the variance where the true value is 0.529.

### 10.1.1 Importance sampling to find HDRs

It has been pointed out (see Chen and Shao, 1998) that we can use importance

sampling to find HDRs. As we noted in Section 2.6, an HDR is an interval which is (for a given probability level) as short as possible. This being so, all we need to do is to take the sample generated by importance sampling, sort it and note that for any desired value  $\alpha$ , if the sorted values are  $y_1, y_2, \dots, y_n$  and  $k = [(1 - \alpha)n]$ , then any interval of the form  $(y_j, y_{j+k})$  will give an interval with an estimated posterior probability of  $(1 - \alpha)$ , so for an HDR we merely need to find the shortest such interval. It may help to consider the following segment of R code:

```
alpha <- 0.1
le <- (1-alpha)*n
lo <- 1:(n-le)
hi <- (le+1):n
y <- sort(samp)
r <- y[hi]-y[lo]
rm <- min(r)
lom <- min(lo[r==rm])
him <- max(hi[r==rm])
abline(v=y[lom])
abline(v=y[him])
```

### 10.1.2 Sampling importance re-sampling

Another technique sometimes employed is sampling importance re-sampling (*SIR*). This provides a method for generating samples from a probability distribution which is hard to handle and for which the constant of proportionality may be unknown. We suppose that the (unnormalized) density in question is  $q(x)$ . What we do is to generate a sample  $x = (x_1, \dots, x_n)$  of size  $n$  from a distribution with density  $p(x)$  which is easier to handle and is in some sense close to the distribution we are interested in. We then find the values of

$$w_i = q(x_i)/p(x_i),$$

and then normalize these to find

$$\pi_i = \frac{w_i}{\sum_i w_i}.$$

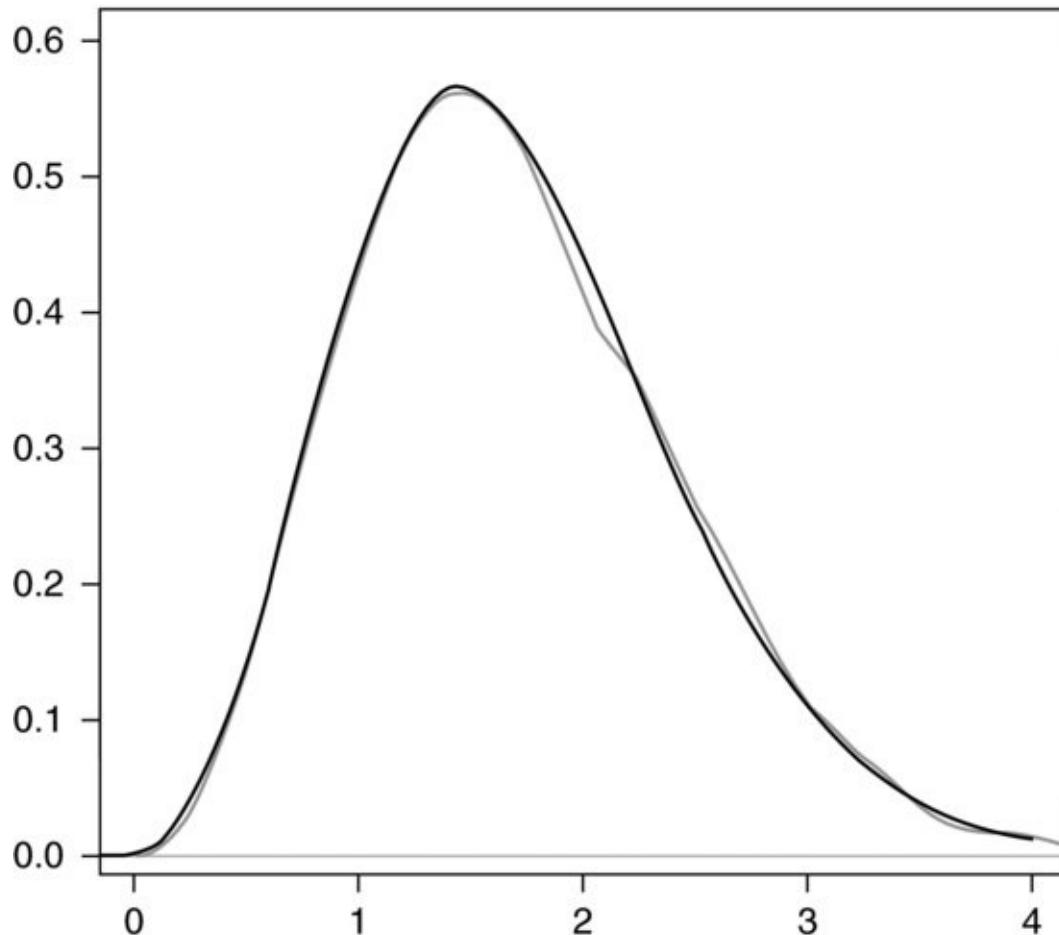
We then take a sample of the same size  $n$  without replacement from the values  $\{x_1, \dots, x_n\}$  with probabilities  $\{\pi_1, \dots, \pi_n\}$  to obtain a new sample  $x^* = (x_1^*, \dots, x_n^*)$ . This operation is easily carried out in R by

```
x <- random sample of size n from density p
w <- q(x)/p(x)
pi <- w/sum(w)
samp <- sample(x, size=n, prob=pi, replace=T)
```

The success of this method is illustrated in [Figure 10.2](#), where the normalized

conjugate density discussed earlier is compared with the smoothed density function found from such a random sample constructed in this way starting from the gamma approximation.

**Figure 10.2** Approximation to the conjugate gamma distribution.



We have used sampling *with* replacement as recommended by Albert (2009, Section 5.10), although Gelman *et al.* (2004, Section 10.5) prefer sampling *without* replacement, claiming that ‘in a bad case, with a few very large values and many small values [s]ampling with replacement will pick the same few values ... again and again; in contrast, sampling without replacement yields a more desirable intermediate approximation’. Nevertheless, at least in some cases such as the one discussed here, substituting `replace=F` results in a much poorer fit.

We can, of course, use the sample thus obtained to estimate the mean, this time as 1.689, and the variance, this time as 0.531. Further, we can use these values to estimate the median as 1.616 where the true value is 1.622.

### 10.1.3 Multidimensional applications

The *SIR* algorithm is most commonly applied in situations when we are trying to find the properties of the posterior distribution of a multidimensional unknown parameter  $\theta$ . Quite often, it turns out that a convenient choice of an approximating density  $p$  from which to generate a random sample is the multivariate t distribution. Since the multivariate t distribution has been avoided in this book, we shall not discuss this approach further; for more details about it see Albert (2009, Section 5.10).

## 10.2 Variational Bayesian methods: simple case

The integrals which arise in Bayesian inference are, as we have seen, frequently intractable unless we use priors which are jointly conjugate, and these are often difficult to deal with, as in the case of the normal/chi-squared distribution introduced in Section 2.13. The idea of variational Bayesian methods is to approximate the posterior by a density of a simpler form. They can be thought of as extensions of the *EM* technique discussed in Section 9.2.

When approximating one distribution by another, it is useful to have a measure of the closeness of the approximation, and for this purpose we shall use the Kullback–Leibler divergence or information  $\mathcal{I}(q : p)$ <sup>1</sup>

$$\mathcal{I}(q : p) = \int q(\theta) \log\{q(\theta)/p(\theta)\} d\theta$$

(cf. Section 3.11; the integral sign denotes a multiple integral when  $\theta$  is multidimensional and, of course, is replaced by a summation in discrete cases). This function satisfies

$$\mathcal{I}(q : p) \geq 0.$$

To show this, note that if we use natural logarithms it is easily shown that  $\log x \leq x - 1$  and hence for any densities  $p(\theta)$  and  $q(\theta)$ , we have

$$\begin{aligned} \int q(\theta) \log\{q(\theta)/p(\theta)\} d\theta &= - \int q(\theta) \log\{p(\theta)/q(\theta)\} d\theta \\ &\geq \int q(\theta)\{1 - p(\theta)/q(\theta)\} d\theta \\ &= \int q(\theta) d\theta - \int p(\theta) d\theta = 0. \end{aligned}$$

Note, however, that it is *not* symmetric [i.e. it is *not* generally true that  $\mathcal{I}(q : p) = \mathcal{I}(p : q)$ ], nor does it satisfy the triangle inequality, so that it is not a metric.

In variational inference, we try to find a distribution of a simpler form which approximates an intractable posterior distribution, so with a posterior  $p(\theta | x)$  we seek  $q(\theta)$  which minimizes

$$\mathcal{I}(q : p) = \int q(\theta) \log\{q(\theta)/p(\theta | x)\} d\theta$$

among a suitable class of densities  $q(\theta)$ .

We note that since  $p(\theta | x) = p(\theta, x)/p(x)$

$$\begin{aligned}\mathcal{I}(q : p) &= \int q(\theta) \log\{q(\theta)/p(\theta | x)\} d\theta \\ &= \int q(\theta) \log p(x) d\theta + \int q(\theta) \log\{q(\theta)/p(\theta, x)\} d\theta \\ &= \log p(x) - \int q(\theta) \log\{p(\theta, x)/q(\theta)\} d\theta.\end{aligned}$$

It follows that

$$\mathcal{I}(q : p) = \log p(x) - \mathcal{L}(q),$$

where

$$\mathcal{L}(q) = \int q(\theta) \log\{p(\theta, x)/q(\theta)\} d\theta = \int q(\theta) \log\{p(x | \theta)p(\theta)/q(\theta)\} d\theta.$$

Since  $\log p(x)$  is not a function of  $q$ , it follows that we minimize the divergence by maximizing  $(q)$ .<sup>2</sup>

### 10.2.1 Independent parameters

For the moment, let us suppose that there are just two unknown parameters  $\mu$  and  $\phi$ , so that  $\theta = (\mu, \phi)$ . It is often the case that the exact posterior  $p(\theta | x)$  is such that  $\mu$  and  $\phi$  are correlated and then it is useful to look for an approximation  $q(\theta)$  to the posterior which factorizes, so that

$$q(\theta) = q_1(\mu)q_2(\phi).$$

When this happens, the unknown parameters are, of course, independent. We look for a density of this form which minimizes the Kullback–Leibler divergence. We proceed sequentially, first taking the distribution of  $\phi$  as known and choosing that of  $\mu$  to minimize the divergence, and then taking that of  $\mu$  as known and choosing the distribution of  $\phi$  to minimize the divergence, and then iterating. Because these functions are densities, we have to apply the subsidiary conditions

$$\int q_1(\mu) d\mu = 1, \quad \int q_2(\phi) d\phi = 1.$$

For this purpose, we make use of a theorem in the calculus of variations given by Gelfand and Fomin (1963, Section 12.1, Theorem 1) which gives a way of

minimizing an integral  $J[y]$  subject to the condition that  $K[y]$  takes a fixed value  $l$ . The result we need is:

**Theorem 10.1** Given the functional

$$J[y] = \int_a^b F(x, y, y') dx,$$

let the admissible curves satisfy the conditions

$$y(a) = A, \quad y(b) = B, \quad K[y] = \int_a^b G(x, y, y') dx = l,$$

where  $K[y]$  is another functional, and let  $J[y]$  have an extremum for  $y=y(x)$ . Then, if  $y=y(x)$  is not an extremal of  $K[y]$ , there exists a constant  $\lambda$  such that  $y=y(x)$  is an extremal of the functional

$$\int_a^b (F + \lambda G) dx,$$

that is  $y=y(x)$  satisfies the differential equation

$$F_y - \frac{d}{dx} F_{y'} + \lambda \left( G_y - \frac{d}{dx} G_{y'} \right) = 0.$$

*Proof.* See Gelfand and Fomin ■

We can extend this result to integrals over the whole real line assuming convergence of the integrals.

Regarding the distribution of  $\phi$  as fixed for the moment, we now apply this theorem to  $q(\mu)$  in the case where we wish to maximize  $(q)$

$$\begin{aligned} \mathcal{L}(q) &= \int q(\theta) \log\{p(\theta, x)/q(\theta)\} d\theta \\ &= \int \int q_1(\mu) q_2(\phi) \log\{p(\mu, \phi, x)/(q_1(\mu)q_2(\phi))\} d\mu d\phi. \end{aligned}$$

We now take  $\mu$  for  $x$ ,  $q_1$  for  $y$ ,

$$\begin{aligned} q_1(\mu) \int q_2(\phi) \log\{p(\mu, \phi, x)/(q_1(\mu)q_2(\phi))\} d\phi \\ = q_1(\mu) \int q_2(\phi) \log p(\mu, \phi, x) d\phi - q_1(\mu) \log q_1(\mu) \\ - q_1(\mu) \int q_2(\phi) \log q_2(\phi) d\phi \end{aligned}$$

for  $F(x, y, y')$  and  $q_1(\mu)$  for  $G(x, y, y')$  (with  $l = 1$ ) in the theorem. As none of the integrals depend on  $q'_1(\mu)$  the terms

$$-\frac{d}{dx} F_{y'} + \lambda \left( -\frac{d}{dx} G_{y'} \right)$$

vanish automatically, while the term  $F_y + \lambda G_y$  becomes

$$\int q_2(\phi) \log p(\mu, \phi, x) d\phi - 1 \cdot \log q_1(\mu) - q_1(\mu) \frac{1}{q_1(\mu)} - \int q_2(\phi) \log q_2(\phi) d\phi + \lambda.$$

It follows that the stationarity condition given by the differential equation in the theorem takes the form

$$\log q_1(\mu) = \int q_2(\phi) \log p(\mu, \phi, x) d\phi + \text{constant},$$

and hence that

$$q_1(\mu) \propto \exp \left\{ \int q_2(\phi) \log p(\mu, \phi, x) d\phi \right\}.$$

The constant of proportionality is determined by the fact that  $\int q_1(\mu) d\mu = 1$ .

The theorem quoted earlier from Gelfand and Fomin requires us to check that  $y$  is not an extremal of  $K[y]$ . In our case  $G(x, y, y') = y$ , so that

$$G_y - \frac{d}{dx} G_{y'} = 1 \neq 0,$$

and hence  $y$  is *not* an extremal of  $K[y]$ .

It should be clear that, when we need to determine the distribution of  $\phi$  whilst regarding the distribution of  $\mu$  as fixed, we take

$$q_2(\phi) \propto \exp \left\{ \int q_1(\mu) \log p(\mu, \phi, x) d\mu \right\}.$$

The argument proceeds with no essential difference in cases where  $\mu$  and  $\phi$  are multidimensional.

### 10.2.2 Application to the normal distribution

As a simple example, we consider a case where we take observations  $x = (x_1, x_2, \dots, x_n)$  from a normal distribution  $N(\mu, \phi)$  with a log-likelihood  $L$  satisfying

$$-2L = -2 \log p(x | \mu, \phi) = n \log \phi + \sum (x_i - \mu)^2 / \phi.$$

We have already shown in Section 2.12 that the posteriors for  $\mu$  and  $\phi$  cannot be expected to be independent, but we wish to find an approximation to the posterior  $p(\theta | x)$  by a density  $q(\mu, \phi) = q_1(\mu)q_2(\phi)$ . We assume independent conjugate priors, so a normal prior  $N(\mu | \mu_0, \phi_0)$  for  $\mu$  and an inverse chi-squared prior  $S_0 \chi_{v_0}^{-2}$  for  $\phi$ . We thus have four ‘hyperparameters’  $\mu_0$ ,  $\phi_0$ ,  $S_0$  and  $v_0$  determining the prior distributions  $p_1(\mu)$  and  $p_2(\phi)$  which satisfy

$$-2 \log p_1(\mu) = (\mu - \mu_0)^2 / \phi_0 + \text{function of } \phi_0 \text{ alone}$$

$$-2 \log p_2(\phi) = (v_0 - 2) \log \phi + \frac{1}{2} S_0 / \phi + \text{function of } S_0 \text{ and } v_0 \text{ alone}$$

(cf. Sections A.1 and A.5). Because

$$p(\mu, \phi, x) = p(x | \mu, \phi)p(\mu, \phi) = p(x | \mu, \phi)p_1(\mu)p_2(\phi),$$

it follows that (up to a function of  $\phi_0$ ,  $S_0$  and  $v_0$  alone)

$$-2 \log p(\mu, \phi, x) = n \log \phi + \sum (x_i - \mu)^2 / \phi + (\mu - \mu_0)^2 / \phi_0 \\ + (v_0 - 2) \log \phi + S_0 / \phi.$$

We seek an approximation to  $p(\theta | x)$  of the form  $q_1(\mu)q_2(\phi)$ , and with this in mind we begin by setting the initial value of  $q_2(\phi)$  as  $p_2(\phi)$  and update  $q_1(\mu)$  using  $q_2(\phi)$  and then update  $q_2(\phi)$  using  $q_1(\mu)$ . This whole process has to be repeated until we obtain convergence.

### 10.2.3 Updating the mean

We update the mean using

$$-2 \log q_1(\mu) = \int \{-2 \log p(\mu, \phi, x)\} q_2(\phi) d\phi,$$

where  $q_2(\phi)$  has an inverse chi-squared distribution  $S\chi_{v-2}$ . Writing  $g(\phi)$  for the density of an  $S\chi_{v-2}$  random variable the right hand side becomes (ignoring some terms constant with respect to  $\mu$ )

$$\begin{aligned} & \int \left\{ n \log \phi + \sum (x_i - \mu)^2 / \phi + (\mu - \mu_0)^2 / \phi_0 + (v_0 - 2) \log \phi + S / \phi \right\} g(\phi) d\phi \\ &= (n + v_0 - 2) \int (\log \phi) g(\phi) d\phi + \sum (x_i - \mu)^2 \int \phi^{-1} g(\phi) d\phi \\ &+ (\mu - \mu_0)^2 \phi_0^{-1} \int g(\phi) d\phi + S_0 \int \phi^{-1} g(\phi) d\phi. \end{aligned}$$

The first and last terms are now seen to be constant with respect to  $\mu$ , while  $g(\phi)$  as a density integrates to unity and  $\int \phi^{-1} g(\phi) d\phi$  is easily shown to be  $v/S$ . Proceeding to complete the square much as we did in Section 2.3, it follows that (with  $c_1, c_2, c_3$  constant with respect to  $\mu$ )

$$\begin{aligned} -2 \log q_1(\mu) &= \sum (x_i - \mu)^2 v/S + (\mu - \mu_0)^2 / \phi_0 + c_1 \\ &= \mu^2 (nv/S + 1/\phi_0) - 2\mu(n\bar{x}v + \mu_0/\phi_0)/S + c_2 \\ &= (\mu - \mu_1)^2 / \phi_1 + c_3, \end{aligned}$$

where

$$\phi_1 = \left( \frac{1}{\phi_0} + \frac{v}{S/n} \right)^{-1}$$

$$\mu_1 = \phi_1 \left( \frac{\mu_0}{\phi_0} + \frac{v\bar{x}}{S/n} \right).$$

Note the similarity of the expression for  $\mu_1$  to the expression for  $\theta^{(t+1)}$  found using the *EM* algorithm in Subsection 9.2.3, headed ‘Semiconjugate prior with a normal likelihood’.

### 10.2.4 Updating the variance

Updating of the variance proceeds similarly, using

$$-2 \log q_2(\phi) = \int \{-2 \log p(\mu, \phi, x)\} q_1(\mu) d\mu.$$

The distribution of  $\mu$  is  $N(\mu_1, \phi_1)$  and consequently, writing  $h(\mu)$  for the density of a  $N(\mu_1, \phi_1)$  distribution, the right-hand side of the expression for  $-2 \log q_2(\phi)$  becomes (ignoring some terms constant with respect to  $\phi$ )

$$\begin{aligned} & \int \left\{ n \log \phi + \sum (x_i - \mu)^2 / \phi + (\mu - \mu_0)^2 / \phi_0 + (v_0 - 2) \log \phi + S_0 / \phi \right\} h(\mu) d\mu \\ &= (n + v_0 - 2) \log \phi + (1/\phi) \int \sum (x_i - \mu)^2 d\mu + (\mu - \mu_0)^2 / \phi_0 + S_0 / \phi \end{aligned}$$

(using  $\int h(\mu) d\mu = 1$ ). The third term is a constant with respect to  $\phi$ . We then use  $\int \mu h(\mu) d\mu = \mu_1$  and  $\int \mu^2 h(\mu) d\mu = \mu_1^2 + \phi_1$  to deduce that

$$\begin{aligned} \int \sum (x_i - \mu)^2 &= \int \left\{ \sum x_i^2 - 2\mu \sum x_i + n\mu^2 \right\} h(\mu) d\mu \\ &= \sum x_i^2 - 2\mu_1 \sum x_i + n(\mu_1^2 + \phi_1) \\ &= (SS + n\bar{x}^2) - 2n\bar{x}\mu_1 + n(\mu_1^2 + \phi_1) \end{aligned}$$

writing

$$SS = \sum (x_i - \bar{x})^2.$$

We thus find that (with  $c_4$  and  $c_5$  constant with respect to  $\phi$ )

$$\begin{aligned} -2 \log q_2(\phi) &= \{S_0 + (SS + n\bar{x}^2) - 2n\bar{x}\mu_1 + n(\mu_1^2 + \phi_1)\} / \phi \\ &\quad + (n + v_0 - 2) \log \phi + c_4 \\ &= (v_1 - 2) \log \phi + S_1 / \phi + c_5, \end{aligned}$$

where

$$S_1 = S_0 + (SS + n\bar{x}^2) - 2n\bar{x}\mu_1 + n(\mu_1^2 + \phi_1)$$

$$v_1 = v_0 + n.$$

## 10.2.5 Iteration

It will be seen that the expressions for  $\phi_1$  and  $\mu_1$  depend on  $S$  and  $v$  and that those for  $S$  and  $v$  depend on  $\phi_1$  and  $\mu_1$ . Consequently, we have to proceed iteratively, first using prior values of  $S$  and  $v$  to derive values of  $\phi_1$  and  $\mu_1$ , then using these values to find values  $S_1$  and  $v_1$ , then using these values to derive values  $\phi_2$  and  $\mu_2$ , and then to values  $S_2$  and  $v_2$ , and so on. The resultant sequences should reasonably quickly converge to values  $\phi^*$ ,  $\mu^*$ ,  $S^*$  and  $v^*$ .

## 10.2.6 Numerical example

In the subsection ‘Semiconjugate prior with a normal likelihood’ of Section 9.2, we investigated how the *EM* algorithm could be used with the data on wheat yield originally considered in the example towards the end of Section 2.13 in which  $n=12$ ,  $\bar{x} = 119$  and sum of squares about the mean  $SS=13\ 045$ . We took a prior for the variance which was  $S_0 \chi_{\nu_0}^{-2}$  with  $S_0=2700$  and  $\nu_0 = 11$  and an independent prior for the mean which was  $N(\mu_0, \phi_0)$  where  $\mu_0 = 110$  and  $\phi_0 = 20$  then<sup>3</sup> and we shall take the same priors now.

The following R program converges within half a dozen iterations to values  $\phi^* = 14.980 = 3.87^2$ ,  $\mu^* = 112.259$ ,  $S^* = 16470$  and  $\nu^* = 23$ . In particular, the distribution of  $\mu$  is  $N(112.259, 3.87^2)$ . Note that the value of 112.259 is close to the value 112.278 which we derived from the *EM* algorithm in Section 9.2.

```
r <- 10
phi <- rep(NA, r)
mu <- rep(NA, r)
S <- rep(NA, r)
n <- 12; xbar <- 119; SS <- 13045
phi0 <- 20; mu0 <- 110; S0 <- 2700; nu0 <- 11
S[1] <- S0
nustar <- nu0 + n
for (i in 2:r) {
  phi[i] <- (1/phi0 + n*nustar/S[i-1])^{-1}
  mu[i] <- phi[i]*(mu0/phi0 + n*xbar*nustar/S[i-1])
  S[i] <- S0 + (SS+n*xbar^2) - 2*n*xbar*mu[i+1] +
    n*(mu[i]^2 + phi[i])
  cat("i", i, "phi", phi[i], "mu", mu[i], "S", S[i], "\n")
}
mustar <- mu[r]; phistar <- phi[r]; Sstar <- S[r]
cat("mu has mean", mustar, "and s.d.", sqrt(phistar), "\n")
cat("phi has mean", Sstar/(nustar-2), "and s.d.",
  (Sstar/(nustar-2))*sqrt(2/(nustar-4)), "\n")
```

## 10.3 Variational Bayesian methods: general case

The variational Bayes approach can also be used in cases where we have more than two parameters or sets of parameters. In cases where  $\theta = (\theta_1, \theta_2, \dots, \theta_k)$ , we seek an approximation to the density  $p(\theta)$  of the form

$$q_1(\theta_1)q_2(\theta_2)\cdots q_n(\theta_n).$$

We need to write

$$\begin{aligned}\theta_{-i} &= (\theta_1, \dots, \theta_{i-1}, \theta_{i+1}, \dots, \theta_k), \\ q_{-i}(\theta_{-i}) &= q_1(\theta_1) \cdots q_{i-1}(\theta_{i-1}) q_{i+1}(\theta_{i+1}) \cdots q_k(\theta_k), \\ d\theta_{-i} &= d\theta_1 \cdots d\theta_{i-1} d\theta_{i+1} \cdots d\theta_k.\end{aligned}$$

With this notation, a straightforward generalization of the argument in the previous section leads to

$$\log q_i(\theta_i) = \int q_{-i}(\theta_{-i}) \log p(\theta, x) d\theta_{-i}$$

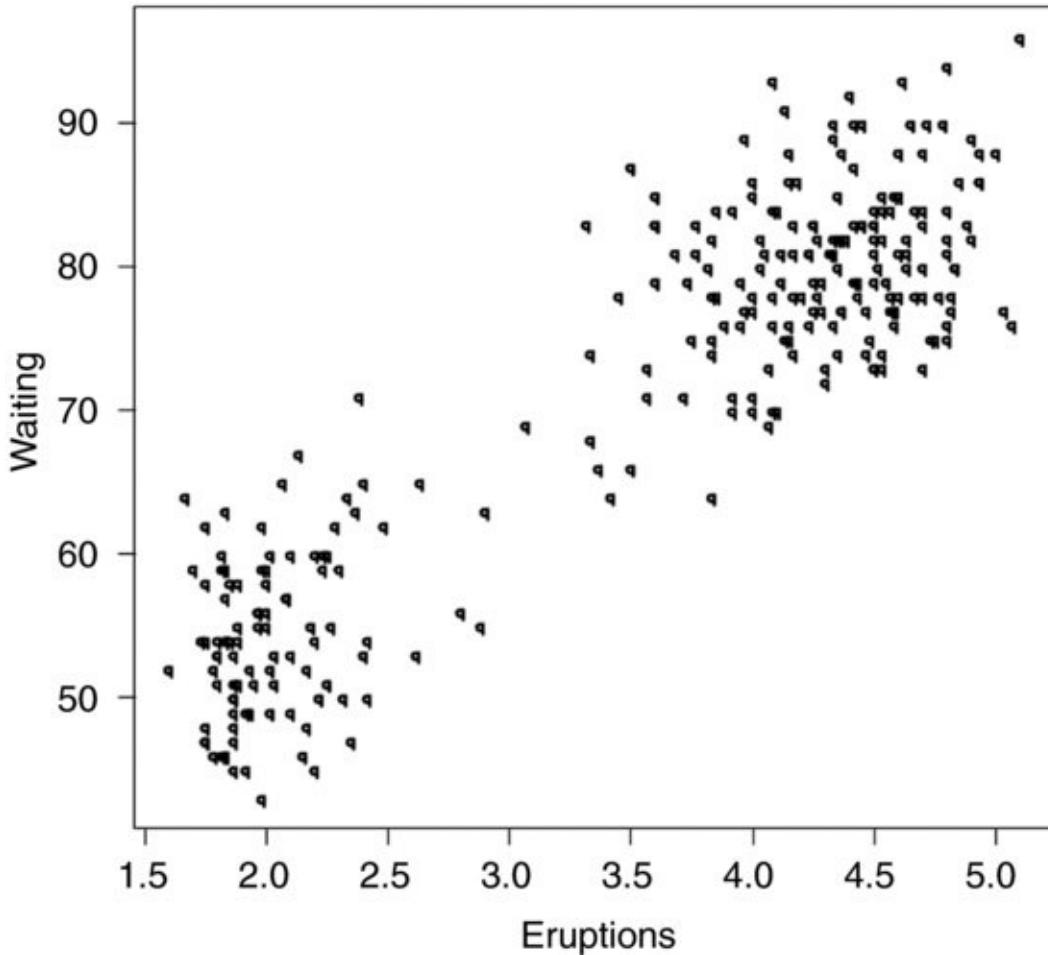
and so to

$$q_i(\theta_i) \propto \exp \left\{ \int q_{-i}(\theta_{-i}) \log p(\theta, x) d\theta_{-i} \right\}.$$

### 10.3.1 A mixture of multivariate normals

We sometimes encounter data in two or more dimensions which is clearly not adequately modelled by a multivariate normal distribution but which might well be modelled by a mixture of two or more such distributions. An example is the data on waiting time between eruptions and the duration of the eruption for the Old Faithful geyser in Yellowstone National Park, Wyoming, USA taken from Härdle (1991), which is illustrated in [Figure 10.3](#). A more complicated example of variational Bayes techniques can be illustrated by outlining a treatment of this problem.

[Figure 10.3](#) Härdle's old faithful data.



We shall write  $\pi = (\pi_1, \dots, \pi_M)$  for a vector of probabilities, so that  $\sum \pi_i = 1$ ,  $\mu = (\mu_1, \dots, \mu_M)$  for a set of  $M$  mean vectors of dimension  $k$ , and  $\Sigma = (\Sigma_1, \dots, \Sigma_M)$  for a set of  $M$  variance–covariance ( $k \times k$ ) matrices.

Suppose that we have a data set  $X$  comprising  $N$  observations  $x_i$  ( $i = 1, \dots, N$ ) which come from a mixture of sub-populations each of which has a  $k$ -dimensional multivariate normal distribution, so that with probability  $\pi_j$  (for  $j = 1, \dots, M$ ) their distribution is multivariate normal with mean  $\mu_j$  and variance–covariance matrix  $\Sigma_j$ , that is  $N(\mu_j, \Sigma_j)$ . Thus,

$$x_j | \pi, \mu, \Sigma \sim \sum_{j=1}^M \pi_j N(\mu_j, \Sigma_j)$$

(where by abuse of notation the expression on the right-hand side means a variable which with probability  $\pi_j$  has an  $N(\mu_j, \Sigma_j)$  distribution). We suppose that all the parameters are unknown, and we seek maximum likelihood estimates of them, so that we can find what the component distributions are and which observations come from which component.

In order to deal with this problem, it is helpful to augment the data with variables  $s_{ij}$  which are indicator variables denoting the sub-population to which each observation belongs, so that

$$s_{ij} \in \{0, 1\} \quad \text{and} \quad \sum_{j=1}^M s_{ij} = 1 \quad \text{with} \quad P(s_{ij} = 1) = \pi_j,$$

and the values for different values of  $i$  are independent of one another. We write  $s$  for the complete set of the  $s_{ij}$ . Estimating the values of the  $s_{ij}$  amounts to determining which sub-population each observation comes from.

In dealing with this problem, we begin by working as if the value of  $\pi$  were known. We then take independent multivariate normal priors of mean  $0$  and variance-covariance matrix  $\beta I$  for the  $\mu_i$ . For the variance-covariance matrices, we take independent inverse Wishart distributions; since we are only outlining a treatment of this problem it is not important exactly what this means and you can take it as given that this is the appropriate multidimensional analogue of the (multiples of) inverse chi-squared distributions we use as a prior for the variance in the one-dimensional case.

We can specify the prior distribution of  $s$  for a given value of  $\pi$  by saying that an individual  $i$  (for  $i = 1, 2, \dots, N$ ) belongs to sub-population  $j$  (for  $j = 1, 2, \dots, M$ ) with probability  $\pi_j$ .

It then follows that

$$p(x, \mu, \Sigma | \pi) = p(x | \mu, \Sigma, s)p(s | \pi)p(\mu)p(\Sigma).$$

In order to evaluate  $p(x | \pi)$ , it is necessary to marginalize this expression with respect to  $\theta = (\mu, \Sigma, s)$ , so that we seek

$$p(x | \pi) = \int p(x, \theta | \pi) d\theta.$$

Now as we found earlier

$$\log p(x | \pi) \geq \int q(\theta) \log\{p(x, \theta | \pi)/q(\theta)\} d\theta = \mathcal{L}(q).$$

We then seek a variational posterior distribution of the form

$$q(\theta) = q(\mu, \Sigma, s) = q_1(\mu)q_2(\Sigma)q_3(s).$$

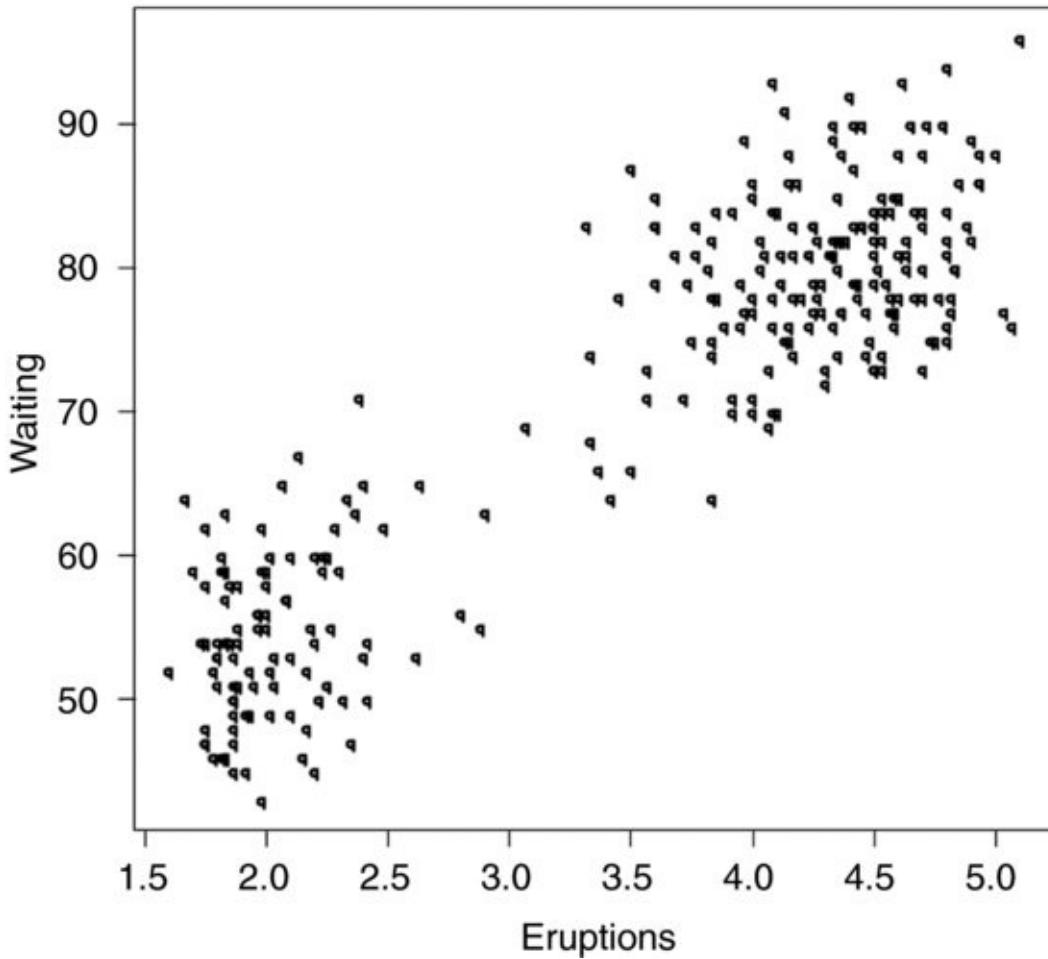
Since we used appropriate conjugate priors, we end up with a posterior for  $\mu$  which is multivariate normal and for  $\Sigma$  which is inverse Wishart. The posterior distribution of  $s$  for a given value of  $\pi$  is determined by a matrix  $(p_{ij})$  where  $p_{ij}$  is the probability that individual  $i$  belongs to sub-population  $j$ .

In this way we obtain a variational lower bound  $(q)$  which approximates the true marginal log-likelihood  $\log p(x | \pi)$ . This bound is, of course, dependent on  $\pi$  through the component  $p(x | \pi)$  of  $p(s, \theta | \pi)$ . By maximizing it with respect to  $\pi$ ,

we obtain our required estimates for the mixing coefficients. But, of course, the value of  $\theta$  and hence the value of the lower bound will depend on  $\pi$ . We, therefore, adopt an *EM* procedure in which we alternately maximize  $(q)$  with respect to  $\pi$  and then optimize  $q$  by updating of the variational solutions for  $q_1$ ,  $q_2$  and  $q_3$  (expectation step). We have so far assumed that  $M$  is fixed and known, but it is possible to try the procedure for various values of  $M$  and choose that value which maximizes the variational likelihood.

We omit the details, which can be found in Corduneanu and Bishop (2001). They find that the Old Faithful geyser data is well fitted by a three-component mixture of bivariate normals with  $\pi = (0.63, 0.33, 0.04)$ . This corresponds to the visual impression given by [Figure 10.3](#) in which it appears that there is one sub-population at the top right, one at the bottom left and possibly a small sub-population between the two.

[Figure 10.3](#) Härdle's old faithful data.



Some further applications of variational Bayesian methods can be found in

Ormerod and Wand (2010). In their words, ‘Variational methods are a much faster alternative to MCMC, especially for large models’.

## 10.4 ABC: Approximate Bayesian Computation

### 10.4.1 The *ABC* rejection algorithm

We sometimes find that we have a model with a likelihood which is mathematically difficult to deal with, and for such cases a technique called Approximate Bayesian Computation or *ABC* (sometimes referred to as *likelihood-free computation*) has been developed and has been quite widely employed, especially in genetics. There are several variants of this technique and we will illustrate each of them with the genetic linkage example we considered earlier in Sections 9.2 and 9.3.

Suppose we want to find the posterior distribution  $p(\theta | x)$  when we have observations  $x^* = (x_1^*, x_2^*, \dots, x_n^*)$  coming from a distribution  $p(x^* | \theta)$  and a (proper) prior distribution  $\pi(\theta)$  for  $\theta$ . Then if the observations come from a discrete distribution we can use the following algorithm to generate a sample of size  $k$  from the required posterior distribution.

#### 10.4.1.1 *ABC* algorithm

1. Generate  $\theta$  from the prior (discrete) density  $\pi(\theta)$ .
2. Generate  $\tilde{x}$  from the density  $p(x | \theta)$ .
3. If  $\tilde{x} = x^*$ , accept  $\theta$  as an observation from  $p(\theta | x^*)$ , and otherwise reject  $\tilde{x}$   
.
4. Repeat until a sample of size  $k$  has been obtained or a set maximum number of iterations has taken place.

This algorithm works because for any set  $A$  of possible values of  $\theta$  we find (using the ‘tilde’ notation introduced in Chapter 1)

$$\begin{aligned} P(\theta \in A \text{ and Accepted}) &= P(\tilde{x} = x^* \text{ and } \theta \in A) \\ &= \int_{\theta \in A} p(x^* | \theta) \pi(\theta) d\theta = \int_{\theta \in A} p(\theta | x^*) P(x^*) d\theta \\ &\propto \int_{\theta \in A} p(\theta | x^*) d\theta \end{aligned}$$

as required.

Although it works, this is an extremely inefficient algorithm. If, for example, we take six observations from a Poisson distribution such as the misprint data considered in Section 3.4

$$\begin{aligned} P(\text{Accepted}) &= \int_0^\infty \prod_{j=1}^6 \left( \frac{\lambda^{x_j^*}}{x_j^*!} \right) e^{-\lambda} \pi(\lambda) d\lambda \\ &= \frac{1}{\prod_{j=1}^6 x_j^*!} \int_0^\infty \lambda^{\sum_{j=1}^6 x_j^*} e^{-6\lambda} \pi(\lambda) d\lambda. \end{aligned}$$

We could calculate the value of this for any conjugate prior

$$\pi(\lambda) = \lambda^{\nu/2-1} \exp(-\frac{1}{2}S_0\lambda),$$

but for simplicity we consider the case where  $\nu = 2$  and  $S_0 = 2$ , so that our prior is  $\pi(\lambda) = e^{-\lambda}$ . With the data we considered in that example, namely,  $x^* = (3, 4, 2, 1, 2, 3)$ , the required probability becomes

$$\begin{aligned} \frac{1}{3!4!2!1!2!3!} \int_0^\infty \lambda^{15} e^{-7\lambda} d\lambda &= \frac{1}{6.24.2.1.2.6} 7^{-16} \int_0^\infty \mu^{15} e^{-\mu} d\mu \\ &= \frac{1}{6.24.2.1.2.6} 7^{-16} 15! \\ &= 0.0000114. \end{aligned}$$

Clearly this technique is not practical without modification, even in the discrete case, and since continuous random variables attain any particular value with probability zero, it cannot be used at all in the continuous case.

We can make a slight improvement by use of sufficient statistics. So in the case of a sample of size 6 from  $P(\lambda)$  we need consider only the sum  $T = \sum x_j$  of the observations, which has a  $P(6\lambda)$  distribution. With this in mind, we can replace step 3 of the algorithm by

3'. If  $T(\tilde{x}) = T(x^*)$ , accept  $\theta$  as an observation from  $p(\theta | x^*)$  and otherwise reject  $\tilde{x}$ .

By working with the sufficient statistic, the acceptance probability with the aforementioned data rises to

$$\begin{aligned} \int_0^\infty \frac{\lambda^T}{T!} e^{-6\lambda} e^{-\lambda} d\lambda &= \frac{1}{6!} \int_0^\infty \lambda^{-16} e^{-7\lambda} d\lambda = \frac{1}{6!} 7^{-16} 15! \\ &= 0.0000547. \end{aligned}$$

This is clearly not enough.

If there is a distance metric  $d$  available defining distances between values of the statistic  $T(x)$ , we can define an  $\varepsilon$ -approximate posterior distribution  $p_\varepsilon(\theta | x^*)$  as

$$p_\varepsilon(\theta | x^*) \propto p(\theta) \int_{R(\varepsilon)} p(T(x) | \theta) d\theta,$$

where

$$R(\varepsilon) = \{x; d(T(x), T(x^*)) \leq \varepsilon\}.$$

### 10.4.1.2 ABC-REJ algorithm

1. Generate  $\theta$  from the prior density  $\pi(\theta)$ .
2. Generate  $\tilde{x}$  from the density  $p(x | \theta)$ .
3. If  $d(T(\tilde{x}), T(x^*)) \leq \varepsilon$ , accept  $\theta$  as an observation from  $p(\theta | x^*)$  and otherwise reject  $\tilde{x}$ .
4. Repeat until a sample of size  $k$  has been obtained or a set maximum number of iterations has taken place.

With this modification, the method can be used with continuous random variables as well as with discrete ones. We observe that if  $\varepsilon = 0$  this algorithm reduces to the previous one with the modification concerning sufficient statistics.

Sometimes there is not a suitable sufficient statistic and we instead use a statistic  $S(x)$  which is in some sense close to being sufficient.

### 10.4.2 The genetic linkage example

We shall now apply this method to the genetic linkage. Recall that we considered observations  $x = (x_1, x_2, x_3, x_4)$  with cell probabilities

$$\left(\frac{1}{2} + \frac{1}{4}\eta, \frac{1}{4}(1 - \eta), \frac{1}{4}(1 - \eta), \frac{1}{4}\eta\right).$$

The values actually observed were  $x_1=125$ ,  $x_2=18$ ,  $x_3=20$ ,  $x_4=34$ , and we are interested in the posterior distribution of  $\eta$ . The likelihood is then

$$\left(\frac{1}{2} + \frac{1}{4}\eta\right)^{x_1} \left(\frac{1}{4}(1 - \eta)\right)^{x_2} \left(\frac{1}{4}(1 - \eta)\right)^{x_3} \left(\frac{1}{4}\eta\right)^{x_4} \propto \left(\frac{1}{2} + \frac{1}{4}\eta\right)^{x_1} \left(\frac{1}{2}(1 - \eta)\right)^{x_2+x_3} \left(\frac{1}{4}\eta\right)^{x_4},$$

and so  $T(x) = (x_1, x_2 + x_3, x_4)$  is sufficient. We take a uniform  $U(0, 1)$  prior for  $\eta$ , use the distance measure

$$d(x, y) = \sqrt{\sum (x_i - y_i)^2}$$

and fix a value for  $\varepsilon$ . The algorithm takes the form:

### 10.4.2.1 ABC-REJ algorithm for the genetic linkage example

1. Generate  $\eta$  from  $U(0, 1)$ .
2. Generate data  $\tilde{x}$  which has a trinomial distribution with index  $n$  and parameter

$$\rho = \left( \frac{1}{2} + \frac{1}{4}\eta, \frac{1}{2}(1-\eta), \frac{1}{4}\eta \right).$$

(We have not formally considered the trinomial distribution, although it was briefly mentioned in Exercise 15 on Chapter 2 and in Exercise 10 on Chapter 3, but it arises as a simple generalization of the binomial, giving the joint distribution of the total number of individuals taken from a population of size  $n$  falling in each of three classes when the individuals are independently assigned to the classes with the probabilities given by  $\rho$ .)

**3.** If  $d(T(\tilde{x}), T(x^*))\varepsilon$ , accept  $\theta$  as an observation from  $p(\theta|x^*)$  and otherwise reject  $\tilde{x}$ .

**4.** Repeat  $N$  times.

A program in R to do this is as follows:

```
N <- 1000000
etastar <- c(125, 18+20, 34)
n <- sum(etastar)
eps <- 3
d <- rep(NA, N)
trial <- rep(NA, N)
for (j in 1:N) {
  etatrial <- runif(1)
  inds <- sample(3, n, replace=T,
    p=c(0.5+etatrial/4, (1-etatrial)/2, etatrial/4))
  samp <- c(sum(inds==1), sum(inds==2), sum(inds==3))
  d <- sqrt(sum((etastar-samp)^2))
  if (d <= eps) trial[j] <- etatrial
}
eta <- trial[!is.na(trial)]
k <- length(eta)
m <- mean(eta)
s <- sd(eta)
cat("k", k, "m", m, "s", s, "\n")
```

A run of this program produced a sample of values of  $\eta$  of size  $k=1040$ , mean  $m=0.625$  and s.d.  $s=0.0511$ . For comparison, we recall that we found a value for  $\eta$  of 0.630 from the EM algorithm in Section 9.2, and a distribution with a posterior mode of 0.627 by data augmentation in Section 9.3.

An ad hoc rule sometimes quoted is that  $\varepsilon$  should be chosen such that approximately 1% of trials of  $\eta$  should be accepted, and the choice of 3 as a value for  $\varepsilon$  was made for this reason.

### 10.4.3 The ABC Markov Chain Monte Carlo algorithm

We can incorporate the idea of the Metropolis–Hastings algorithm introduced in

Section 9.6 into approximate Bayesian computation. In the aforementioned algorithm  $q(\theta | \theta_i)$  is a suitable proposal density (a transition probability density for a Markov chain) and

$$\alpha = \min \left[ \frac{\pi(\theta^*) q(\theta_i | \theta^*)}{\pi(\theta_i) q(\theta^* | \theta_i)} I_\varepsilon(x, x^*), 1 \right],$$

where

$$I_\varepsilon(x, x^*) = \begin{cases} 1 & \text{if } d(S(x), S(x^*)) \leq \varepsilon \\ 0 & \text{otherwise,} \end{cases}$$

the statistic  $S(x)$  being in some sense close to being sufficient.

#### 10.4.3.1 ABC-MCMC algorithm

1. Generate  $\theta_1$  from the prior density  $\pi(\theta)$ .
2. Generate  $\theta^*$  from the proposal density  $q(\theta | \theta_i)$ .
3. Generate  $X$  from the density  $p(x | \theta^*)$ .
4. Set  $\theta_{i+1} = \theta^*$  with probability  $\alpha$  and otherwise set  $\theta_{i+1} = \theta_i$ .
5. Repeat until a sample of size  $k$  has been obtained or a set maximum number of iterations has taken place.

In the genetic linkage example [using the sufficient statistic  $T(x) = (x_1, x_2 + x_3, x_4)$  as our  $S(x)$ ] a convenient choice is a uniform prior  $U(0, 1)$  and a normal proposal density

$$q(\theta | \theta_i) = (2\pi\phi)^{-1} \exp\{-\frac{1}{2}(\theta - \theta_0^2)/\phi\}$$

because, as is easily checked,  $[\pi(\theta^*) q(\theta_i | \theta^*)]/[\pi(\theta_i) q(\theta^* | \theta_i)]$  then reduces to unity, and so

$$\alpha = \begin{cases} 1 & \text{if } d(S(x), S(x^*)) \leq \varepsilon \\ 0 & \text{otherwise.} \end{cases}$$

The algorithm thus becomes:

#### 10.4.3.2 ABC-MCMC algorithm for the genetic linkage example

1. Generate  $\theta_1$  from the prior density  $U(0, 1)$ .
2. Generate  $\theta^*$  from the proposal density  $q(\theta | \theta_i)$ .
3. Generate  $X$  from the density  $p(x | \theta^*)$ .
4. Set  $\theta_{i+1} = \theta^*$  if  $d(S(x), S(x^*)) \leq \varepsilon$  and otherwise set  $\theta_{i+1} = \theta_i$ .
5. Repeat until a sample of size  $k$  has been obtained or a set maximum

number of iterations has taken place.

A program in R to do this is as follows:

```
N <- 100000
epsilon <- 3
sdev <- 0.05
k <- 0
etaset <- 0.5
xstar <- c(125, 18+20, 34)
n <- sum(xstar)
eta <- rep(NA, N)
for (j in 1:N) {
  etatrial = rnorm(1, etaset, sdev)
  if ((0<etatrial)&(etatrial<1)) {
    inds <- sample(3, n, replace=T,
      prob=c((0.5+etatrial/4), ((1-etatrial)/2),
      (etatrial/4)))
    x <- c(sum(inds==1), sum(inds==2), sum(inds==3))
    dist <- sqrt(sum((xstar-x)^2))
    if (dist <= epsilon) {
      etaset <- etatrial
      k <- k + 1
    }
  }
  eta[j] <- etaset
}
m <- mean(eta)
s <- sd(eta)
cat("k", k, "m", m, "s", s, "\n")}
```

A run of this program produced a sample of values of  $\eta$  of size  $k=4627$ , mean  $m=0.624$  and s.d.  $s=0.0499$ . Note the number of acceptances is some four and a half times as great as with the ABC-REJ algorithm.

A problem which can occur with this algorithm is pointed out by Sisson *et al.* (2007). In their words, ‘if the ABC-MCMC sampler enters an area of relatively low probability with a poor proposal mechanism, the efficiency of the algorithm is strongly reduced because it then becomes difficult to move anywhere with a reasonable chance of acceptance, and so the sampler “sticks” in that part of the state space for long periods of time.’

#### 10.4.4 The ABC Sequential Monte Carlo algorithm

In the ABC-SMC algorithm we try to improve on simple rejection sampling by adopting a sequential Monte Carlo based simulation approach. Here, a collection

of values  $\theta^{(1)}, \dots, \theta^{(N)}$  from the parameter space (termed particles) is propagated from an initial prior distribution through a sequence of intermediary distributions, until it ultimately represents the target distribution. The method can be thought of as a type of importance sampling (cf. Section 10.1).

We begin by taking  $\theta_1^{(1)}, \dots, \theta_1^{(N)}$  from an initial distribution  $\mu_1$  (which may or may not be the prior distribution  $p$ ). We then seek a target distribution  $f_T(\theta)$  which represents the posterior of  $\theta$  conditional on  $d(S(x^*), S(x)) \leq \varepsilon$  for observed data  $x^*$ . The standard importance sampling procedure would then indicate how well each particle  $\theta^{(i)}$  adheres to  $f_T(\theta)$  by specifying the importance weight  $W_T^{(i)} = f_T(\theta_1^{(i)})/\mu_1(\theta_1^{(i)})$  it should receive in the full population of  $N$  particles. The effectiveness of such a procedure is sensitive to the choice of  $\mu_1$  and it can be highly inefficient if  $\mu_1$  is diffuse relative to  $f_r$ .

The idea behind sequential sampling methods is to avoid the potential disparity between  $\mu_1$  and  $f_r$  by specifying a sequence of intermediary distributions  $f_1, \dots, f_{T-1}$  such that they evolve gradually from the initial distribution towards the target distribution. In the *ABC* setting, we may naturally define the sequence of distributions  $f_1, \dots, f_T$  as

$$f_t(\theta | d(S(x^*), S(x)) \leq \varepsilon_t),$$

where  $\varepsilon_1, \dots, \varepsilon_T = \varepsilon$  is a strictly decreasing sequence of tolerances.

The degree of sample degeneracy in a population is measured by the effective sample size *ESS* which represents the equivalent number of random samples needed to obtain an estimate such that its Monte Carlo variation is equal to that of the  $N$  weighted particles and it may be estimated as  $[\sum_{i=1}^N (W_t^{(i)})^2]^{-1}$ , which is between 1 and  $N$ . It is usual to set a re-sampling threshold  $E$ , which is often taken as  $N/2$ .

A difficulty with this approach is that, while some action occurs (e.g. a realization or move proposal is accepted) each time a non-zero likelihood is encountered, there is a large probability that the likelihood, and therefore the particle weight will be zero. Fortunately, the idea of partial rejection control leads to the modified notion of the *ABC-PRC* algorithm.

This process can be described in the following algorithm which assumes that data  $x_0$  is available and is taken from the 2009 corrections to Sisson *et al.* (2007):

#### 10.4.4.1 ABC-PRC algorithm

**1.** Initialize  $\varepsilon_1, \dots, \varepsilon_T$  and specify the initial sampling density  $\mu_1$  (which may or may not coincide with the prior distribution  $\pi(\theta)$ ). Set the population indicator  $t$  to 1.

**2.** Set the particle indicator to 1.

**a.** If  $t=1$  sample  $\theta^{**}$  independently from  $\mu_1$ . If  $t>1$  sample  $\theta^*$  from the previous population  $(\theta_{t-1}^{(i)})$  with weights  $(W_{t-1}^{(i)})$ , and move the particle to  $\theta^{**}$  according to a Markov transition density  $K_t$  which may or may not vary with  $t$ . Generate a data set  $x^{**}$  from the density  $f(x|\theta^{**})$ . If  $d(S(x^{**}), S(x_0)) > \varepsilon$  then go back to 2(a).

**b.** Set  $\theta_t^{(i)} = \theta^{**}$  and

$$W_t^{(i)} = \begin{cases} \pi(\theta_t^{(i)})/\mu_1(\theta_t^{(i)}) & (t = 1) \\ \pi(\theta_t^{(i)}) / \sum_{j=1}^N W_{t-1}(\theta_{t-1}^{(j)}) K_t(\theta_t^{(j)} | \theta_{t-1}^{(j)}) & (t > 1) \end{cases}$$

If  $i < N$  increment  $i$  to  $i+1$  and go back to 2(a).

**3.** Normalize the weights, so that  $\sum_{i=1}^N W_t^{(i)} = 1$ . If  $ESS = [\sum_{i=1}^N (W_t^{(i)})^2]^{-1} < E$  then re-sample with replacement the particles  $\theta_t^{(i)}$  with weights  $W_t^{(i)}$  and then set all weights  $W_t^{(i)}$  to  $1/N$ .

**4.** If  $t < T$  increment  $t$  to  $t+1$  and go back to step 2.

Note that as  $\pi$  is the prior density, when  $t=1$  step 2(b) gives the same weights as in importance sampling in Section 10.1, while the weights when  $t>1$  come from an obvious updating process.

An R program to carry this out for the genetic linkage example, using the normal random walk (restricted, so that  $\eta$  is always between 0 and 1) as the Markov transition density, is as follows:

```
N <- 1000
E <- N/2
x0 <- c(125, 18+20, 34)
n <- sum(x0)
sdev <- 0.01
pi <- function(x) dunif(x, 0, 1)
q <- function(eta, eta0) dnorm(eta, eta0, sdev)
epsilon <- c(9, 3, 1)
T <- length(epsilon)
eta <- matrix(NA, N, T)
etaold <- matrix(NA, N, T)
```

```

W <- matrix(NA,N,T)
X <- rep(NA,N)
W[1:N,1] <- 1/N
for (t in 1:T) {
  for (i in 1:N) {
    dist <- max(epsilon) + 1
    while (dist > epsilon[t]) {
      if (t==1) etastarstar <- runif(1,0,1)
      if (t>1) {
        etastar <-
          sample(eta[,t-1],1,replace=TRUE,prob=W[,t-1])
        etastarstar <- -1
        while (!((0<etastarstar)&(etastarstar<1))) {
          etastarstar <- rnorm(1,etastar,sdev)
        }
      }
      inds <- sample(3,n,replace=T,
                    p=c((0.5+etastarstar/4),((1-etastarstar)/2),
                          (etastarstar/4)))
      xstarstar <-
        c(sum(inds==1),sum(inds==2),sum(inds==3))
      dist <- sqrt(sum((xstarstar-x0)^2))
    }
    eta[i,t] <- etastarstar
  }
  for (i in 1:N) {
    if (t==1) W[1:N,1] <- 1/N
    if (t>1) {
      S <- sum(W[1:N,1]*q(eta[i,t],eta[1:N,t-1]))
      W[i,t] <- pi(eta[i,t])/S
    }
  }
  ESS <- 1/sum(W[1:N,t]^2)
  if (ESS < E) {
    etaold <- eta
    for (i in 1:N) eta[i,t] <-
      sample(eta[,t],1,replace=TRUE,prob=W[,t-1])
    W[1:N,t] <- 1/N
  }
}
plot(density(eta[,1]),lty=3,xlim=c(0.45,0.75),
      ylim=c(0,15),xlab="",ylab="",main="")
par(new=T)
plot(density(eta[,2]),lty=2,xlim=c(0.45,0.75),

```

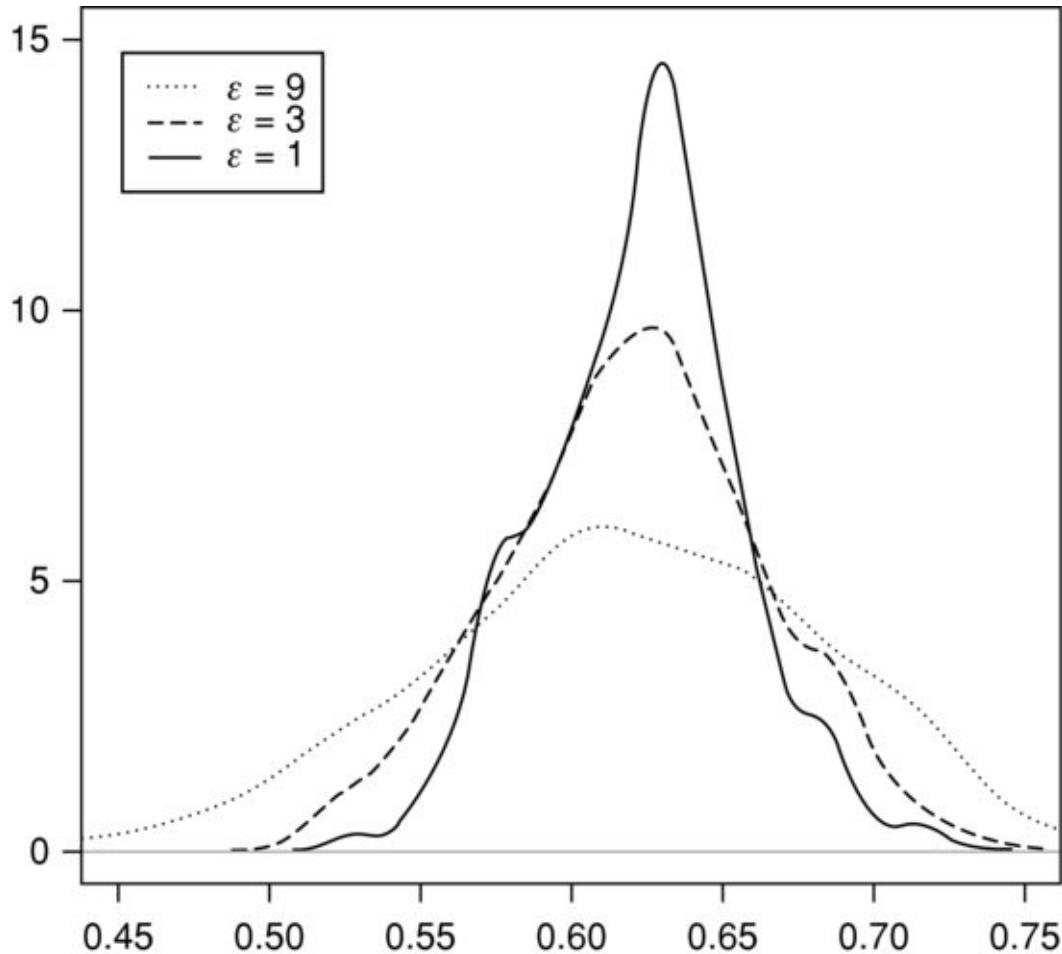
```

ylim=c(0,15),xlab="",ylab="",main="")
par(new=T)
plot(density(eta[,3]),lty=1,xlim=c(0.45,0.75),
      ylim=c(0,15),xlab="",ylab="",main="")
legend(0.5,12.5,c(expression(paste(epsilon," = 9")),
                     expression(paste(epsilon," = 3")),
                     expression(paste(epsilon," = 1))),lty=3:1)

```

Estimated density functions of  $\eta$  based on  $\varepsilon = 9$ ,  $\varepsilon = 3$  and  $\varepsilon = 1$  are shown in [Figure 10.4](#).

**Figure 10.4** The ABC-PRC algorithm for the genetic linkage example.



#### 10.4.5 The ABC local linear regression algorithm

Another variant of *ABC* which has been proposed uses local linear regression. We shall illustrate the technique in the case where there is a single unknown parameter  $\theta$ . We define the Epanechnikov kernel as the function

$$K_\varepsilon[t] = \begin{cases} \frac{3}{4}\varepsilon^{-1}[1 - (t/\varepsilon)^2] & (t \leq \varepsilon) \\ 0 & (t > \varepsilon), \end{cases}$$

so that  $K_\varepsilon[t] = K_\varepsilon[-t]$  and  $\int K_\varepsilon[t] dt = 1$ . We then use this function to give most weights to those simulations which result in the smallest values of  $d(S(x), S(x^*))$ . Thus, we could estimate the mean value of  $\theta$  by

$$\frac{\sum \theta_i K_\varepsilon[d(S(x_i), s(x^*))]}{\sum K_\varepsilon[d(S(x_i), s(x^*))]}.$$

But actually we can go further than this. We can seek values of  $\alpha$  and  $\beta$  which minimize

$$SS = \sum \{\theta_i - \alpha - \beta(S(x_i) - S(x^*))\}^2 K_\varepsilon[d(S(x_i), S(x^*))].$$

If instead of using the Epanechnikov kernel, we had taken  $K_\varepsilon(t) = 1$  for all  $t$  this would result in ordinary linear regression as we considered in Section 6.3, but instead we are giving more weight to observations for which  $d(S(x_i), S(x^*))$  is small. Straightforward differentiation shows that

$$-\frac{1}{2} \partial SS / \partial \alpha = \sum \{\theta_i - \alpha - \beta(S(x_i) - S(x^*))\} K_\varepsilon[d(S(x_i), S(x^*))]$$

$$-\frac{1}{2} \partial SS / \partial \beta = \sum (S(x_i) - S(x^*)) \{\theta_i - \alpha - \beta(S(x_i) - S(x^*))\} K_\varepsilon[d(S(x_i), S(x^*))].$$

By setting  $\partial SS / \partial \alpha = \partial SS / \partial \beta = 0$  we see that the minimum occurs when  $\alpha = \hat{\alpha}$  and  $\beta = \hat{\beta}$ , where

$$\sum \{\theta_i - \hat{\alpha} - \hat{\beta}(S(x_i) - S(x^*))\} K_\varepsilon[d(S(x_i), S(x^*))] = 0$$

$$\sum (S(x_i) - S(x^*)) \{\theta_i - \hat{\alpha} - \hat{\beta}(S(x_i) - S(x^*))\} K_\varepsilon[d(S(x_i), S(x^*))] = 0$$

and these equations are easily solved, albeit not quite as simply as in Section 6.3 since  $\hat{\beta}$  does not drop out of the first equation.

The natural way of solving such equations is by matrix algebra and with matrix techniques it is almost as easy to deal with cases where the statistic  $S(x)$  is multidimensional. For details see Beaumont *et al.* (2002).

#### 10.4.6 Other variants of ABC

Various other variants of the ABC technique have been considered, such as non-linear regression models, in particular fast-forward neural network regression models. For details, see Blum and François (2010).

It has also been applied to model selection (see Toni and Stumpf, 2010).

A forthcoming paper likely to be of interest in connection with ABC is Fearnhead and Prangle (2012) and a useful survey is Sisson and Fan (2011).

### 10.5 Reversible jump Markov chain Monte Carlo

The Metropolis–Hastings algorithm introduced in Section 9.6 can be generalized

to deal with a situation in which we have a number of available models, each with unknown parameters, and we wish to choose between them.

The basic idea is that as well as considering moves *within* a model as in the basic Metropolis–Hastings algorithm we also consider possible moves *between* models. We can regard the chain as moving between states which are specified by a model and a set of parameters for that model. For the time being, we shall suppose that we have two models  $M^{(1)}$  with parameters  $\theta^{(1)}$  and  $M^{(2)}$  with parameters  $\theta^{(2)}$ , so that at time  $t$  we have model  $M^{(i[t])}$  with parameters  $\theta^{(i[t])}$ .

### 10.5.1 RJMCMC algorithm

1. Initialize by setting  $t=0$  and choosing  $i[0]$  and hence model  $M^{(i[0])}$ , and then initial parameters  $\theta^{(i[0])}$ .
2. For
  - a. Update the parameters  $\theta^{(i[t-1])}$  of the current model  $M^{(i[t-1])}$  as in the usual Metropolis–Hastings algorithm.
  - b. Propose to move to a new model  $M'$  with parameters  $\phi$ . Calculate the acceptance probability for the move and decide whether to accept or reject the new model (and parameter values).

An example arises in connection with the coal-mining disaster data in which we could consider a model  $M^{(1)}$  in which all the data are supposed to be such that  $x_i \sim P(\kappa)$  for  $i = 1, 2, \dots, n$  and compare it with the model  $M^{(2)}$  discussed in Section 9.4 in which we suppose there is an unknown value  $k$  such that  $x_i \sim P(\lambda)$  for  $i = 1, 2, \dots, k$  and  $x_i \sim P(\mu)$  for  $i = k + 1, \dots, n$  (with  $\lambda$  and  $\mu$  unknown). In this case typical states of the chain would be  $(i = 1, \kappa = 1.5)$  and  $(i = 2, k = 40, \lambda = 3.2, \mu = 0.9)$ .

A problem which is immediately seen with models  $M^{(1)}$  and  $M^{(2)}$  described earlier is that while  $M^{(1)}$  has only one unknown parameter, namely  $\kappa$ ,  $M^{(2)}$  has three, namely  $(k, \lambda, \mu)$ . We accordingly add auxiliary variables  $u = (v, w)$  to the parameters of model  $M$ , so that the parameter spaces of the two models are of the same dimension. The auxiliary variables do not affect the modelling of the data but allow a one-to-one mapping between the two parameter spaces. We can easily map from the parameters of  $M^{(2)}$  to  $M^{(1)}$  by setting  $\kappa = \sqrt{\lambda\mu}$ , but in order to map from the parameters of  $M^{(1)}$  to uniquely determined parameters  $M^{(2)}$  we need these auxiliary variables. We can then, for example, take  $\lambda = \kappa e^v$ ,  $\mu = \kappa e^{-v}$  and  $k=w$ , where the random variables  $v$  and  $w$  are chosen to have suitable

distributions.

Because of the fact that we change the variables we use, we need to re-examine the detailed balance equation

$$\pi(\theta) q(\phi | \theta) \alpha(\phi | \theta) = \pi(\phi) q(\theta | \phi)$$

from which we deduced the Metropolis–Hastings algorithm in Section 9.6. We note that since we need to work in terms of probability elements in the continuous multivariate case this should be thought of as

$$\pi(\theta) q(\phi | \theta) \alpha(\phi | \theta) |d\theta| = \pi(\phi) q(\theta | \phi) |d\phi|.$$

In one dimension, the ratio of differentials would give rise to an ordinary derivative, but the analogue in many dimensions is, of course, the Jacobian  $|J| = |\partial\phi/\partial\theta|$ . For this reason, we find

$$\alpha(\phi | \theta) = \min \left[ \frac{\pi(\phi | x) q(\theta | \phi)}{\pi(\theta | x) q(\phi | \theta)} |J|, 1 \right].$$

Since the models have different parameters, it is clear that the priors cannot cancel out and so must be proper.

We can give a more explicit form to the probability of switching as follows. Let  $\theta$  be the parameters of a model  $M^{(i)}$  and let  $u$  be the associated auxiliary variables. Let  $\phi$  be the parameters of a model  $M^{(j)}$  and let  $u'$  be the associated auxiliary variables. Write  $\theta^* = (\theta, u)$  and  $\phi^* = (\phi, u')$ . Let  $r(u)$  be a proposal distribution for  $u$  and let  $r'(u')$  be the proposal distribution for  $u'$ . Then the probability of moving from model  $i$  with parameters  $\theta$  to model  $j$  with parameters  $\phi$  is

$$\min \left[ \frac{\pi(j, \phi | x) q(i | j) r(u)}{\pi(i, \theta | x) q(j | i) r'(u')} |J|, 1 \right],$$

where  $q(i | j)$  is the probability that a change of model from model  $i$  results in model  $j$  and  $|J| = |\partial(\phi^*)/\partial(\theta^*)|$  is the Jacobian.

In the case of the coal-mining disasters, we have only two possible models, so that  $q(2 | 1) = q(1 | 2) = 1$ . Further, as  $\lambda = \kappa e^v$ ,  $\mu = \kappa e^{-v}$  and  $k=w$ , we find

$$\begin{aligned} \frac{\partial(\theta^{(2)}, u^{(2)})}{\partial(\theta^{(1)}, u^{(1)})} &= \begin{vmatrix} \partial\lambda/\partial\kappa & \partial\lambda/\partial v & \partial\lambda/\partial w \\ \partial\mu/\partial\kappa & \partial\mu/\partial v & \partial\mu/\partial w \\ \partial k/\partial\kappa & \partial k/\partial v & \partial k/\partial w \end{vmatrix} = \begin{vmatrix} e^v & \kappa e^v & 0 \\ e^{-v} & -\kappa e^{-v} & 0 \\ 0 & 0 & 1 \end{vmatrix} \\ &= -2\kappa = -2\sqrt{\lambda\mu}, \end{aligned}$$

so that, depending on which way a jump is proposed, either  $|J| = 2\kappa = 2\sqrt{\lambda\mu}$  or  $|J| = 1/2\kappa = 1/2\sqrt{\lambda\mu}$ .

In this case, a suitable proposal density  $r^{(1)}(u^{(1)})$  is obtained by giving  $v$  a normal distribution with mean zero and a suitable variance, so  $v \sim N(0, \phi)$ , and  $w$  a discrete uniform distribution over  $(2, n-1)$ , so that  $w \sim UD(2, n-1)$ . Since  $u^{(2)}$

is null, we may take  $r^{(2)}(u^{(2)}) = 1$ . In fact, it turns out that there is strong evidence for the existence of a change point; if the chain is started in model  $M^{(2)}$  it always remains in that model (and the change point still being some time in 1889), although, if we allow the possibility of more than one change point, then Green (1995) shows that there is weak evidence for a second change point.

Another case where *RJMCMC* can be applied is in mixture models where we are unsure how many components are present. As a simple case, we might wish to tell whether a data set  $X$  consists of a single normal population of mean 0 or a 50:50 mixture of populations with mean 0 but different variances. More generally, we could allow for more components, for different means and for variable mixing coefficients. The method can also be applied in choosing between various regression models. In cases where there are  $m > 2$  models, values of  $q(j | i)$  will have to be supplied, but in some cases it will suffice simply to take  $q(j | i) = 1/(m-1)$  for all  $i$  and  $j$  (with  $j \neq i$ ), or at least to make all jumps to models which are in some sense ‘adjacent’ equally likely.

Some more details about this approach, including a detailed treatment of the coal-mining disaster data, can be found in Green (1995); see also Richardson and Green (1997).

More information can be found in Fan and Sisson (2011) or Gamerman and Lopes (2006). *RJMCMC* is in the process of being incorporated into WinBUGS; for information about this see the website associated with this book.

## 10.6 Exercises on Chapter 10

1. Show that in importance sampling the choice

$$p(x) = \frac{|f(x)|q(x)}{\int |f(\xi)|q(\xi) d\xi}$$

minimizes  $\mathcal{V}_w(x)$  even in cases where  $f(x)$  is not of constant sign.

2. Suppose that has a Cauchy distribution. It is easily shown that  $\eta = P(x > 2) = \tan^{-1}(\frac{1}{2})/\pi = 0.1475836$ , but we will consider Monte Carlo methods of evaluating this probability.

- a. Show that if  $k$  is the number of values taken from a random sample of size  $n$  with a Cauchy distribution, then  $k/n$  is an estimate with variance  $0.1258027/n$ .

- b. Let  $p(x)=2/x^2$ , so that  $\int_x^\infty p(\xi) d\xi = 2/x$ . Show that if  $x \sim U(0, 1)$  is uniformly distributed over the unit interval then  $y=2/x$  has the density  $p(x)$  and that all values of  $y$  satisfy  $y \geq 2$  and hence that

$$\sum_{i=1}^n \frac{1}{2\pi} \frac{y_i^2}{1+y_i^2}$$

gives an estimate of  $\eta$  by importance sampling.

- c. Deduce that if  $x_1, x_2, \dots, x_n$  are independent  $U(0, 1)$  variates then

$$\hat{\eta} = \frac{1}{n} \sum_{i=1}^n \frac{1}{2\pi} \frac{4}{4+x_i^2}$$

gives an estimate of  $\eta$ .

- d. Check that  $\hat{\eta}$  is an unbiased estimate of  $\eta$  and show that

$$E\hat{\eta}^2 = \frac{\tan^{-1}(\frac{1}{2}) + \frac{2}{3}}{4\pi^2}$$

and deduce that

$$\mathcal{V}\hat{\eta} = 0.0000955,$$

so that this estimator has a notably smaller variance than the estimate considered in (a).

3. Apply sampling importance re-sampling starting from random variables uniformly distributed over  $(0, 1)$  to estimate the mean and variance of a beta distribution  $Be(2, 3)$ .

4. Use the sample found in Section 10.5 to find a 90% HDR for  $Be(2, 3)$  and compare the resultant limits with the values found using the methodology of Section 3.1. Why do the values differ?

**5.** Apply the methodology used in the numerical example in Section 10.2 to the data set used in both Exercise 16 on Chapter 2 and Exercise 5 on Chapter 9.

**6.** Find the Kullback–Leibler divergence  $\mathcal{I}(q : p)$  when  $p$  is a binomial distribution  $B(n, \pi)$  and  $q$  is a binomial distribution  $B(n, \rho)$ . When does  $\mathcal{J}(q : p) = \mathcal{J}(p : q)$ ?

**7.** Find the Kullback–Leibler divergence  $\mathcal{I}(q : p)$  when  $p$  is a normal distribution  $N(\mu, \phi)$  and  $q$  is a normal distribution  $N(\nu, \psi)$ .

**8.** Let  $p$  be the density  $2(2\pi)^{-1/2} \exp(-\frac{1}{2}x^2)$  ( $x > 0$ ) of the modulus  $x=|z|$  of a standard normal variate  $z$  and let  $q$  be the density  $\beta^{-1} \exp(-x/\beta)$  ( $x > 0$ ) of an  $E(\beta)$  distribution. Find the value of  $\beta$  such that  $q$  is as close an approximation to  $p$  as possible in the sense that the Kullback–Leibler divergence  $\mathcal{I}(q : p)$  is a minimum.

**9.** The paper by Corduneanu and Bishop (2001) referred to in Section 10.3 can be found on the web at

[http://research.microsoft.com/pubs/67239/  
bishop-aistats01.pdf](http://research.microsoft.com/pubs/67239/bishop-aistats01.pdf).

Härdle's data set is available in R by going `data(faithful)`. Fill in the details of the analysis of a mixture of multivariate normals given in that section.

**10.** Carry out the calculations in Section 10.4 for the genetic linkage data quoted by Smith which was given in Exercise 3 on Chapter 9.

**11.** A group of  $n$  students sit two exams. Exam one is on history and exam two is on chemistry. Let  $x_i$  and  $y_i$  denote the  $i$ th student's score in the history and chemistry exams, respectively. The following linear regression model is proposed for the relationship between the two exam scores:

$$y_i = \alpha + \beta x_i + \varepsilon_i \quad (i = 1, 2, \dots, n),$$

where  $\varepsilon_i \sim N(0, 1/\tau)$ .

Assume that  $x = (x_1, x_2, \dots, x_n)$  and  $y = (y_1, y_2, \dots, y_n)$  and that  $\alpha$ ,  $\beta$  and  $\tau$  are unknown parameters to be estimated.

Describe a reversible jump MCMC algorithm including discussion of the acceptance probability, to move between the four competing models:

- 1.**  $y_i = \alpha + \varepsilon_i ;$
- 2.**  $y_i = \alpha + \beta x_i + \varepsilon_i ;$
- 3.**  $y_i = \alpha + \lambda t_i + \varepsilon_i ;$

**4.**  $y_i = \alpha + \beta x_i + \lambda t_i + \varepsilon_i$ .

Note that if  $z$  is a random variable with probability density function  $f$  given by

$$f(z) \propto \exp\left(-\frac{1}{2}A(z^2 - 2Bz)\right),$$

then  $z \sim N(B, 1/A)$  [due to P. Neal].

<sup>1</sup> Often denoted  $D_{\text{KL}}(q||p)$  or  $\text{KL}(q||p)$ .

<sup>2</sup> Those with a background in statistical physics sometimes refer to  $(q)$  as the (negative) variational free energy because it can be expressed as an ‘energy’

$$\mathbb{E} \log p(\theta, x) = \int q(\theta) \log p(\theta, x) d\theta = \int q(\theta) \log(p(x|\theta)p(\theta)) d\theta$$

plus the entropy

$$-\mathbb{E} \log q(\theta) = - \int q(\theta) \log q(\theta) d\theta$$

but it is not necessary to know about the reasons for this.

<sup>3</sup> In that subsection, we wrote  $S$  where we will now write  $SS$ , we wrote  $\nu$  where we will now write  $\nu_0$ , and we wrote  $\theta_0$  where we will now write  $\mu_0$ .

# Appendix A: Common statistical distributions

Some facts are given about various common statistical distributions. In the case of continuous distributions, the (probability) density (function)  $p(x)$  equals the derivative of the (cumulative) distribution function  $F(x) = P(X \leq x)$ . In the case of discrete distributions, the (probability) density (function)  $p(x)$  equals the probability that the random variable  $X$  takes the value  $x$ .

The mean or expectation is defined by

$$\mathbb{E}X = \int xp(x) dx \quad \text{or} \quad \sum xp(x)$$

depending on whether the random variable is discrete or continuous. The variance is defined as  $\mathcal{V}X = \int (x - \mathbb{E}X)^2 p(x) dx \quad \text{or} \quad \sum (x - \mathbb{E}X)^2 p(x)$

depending on whether the random variable is discrete or continuous. A mode is any value for which  $p(x)$  is a maximum; most common distributions have only one mode and so are called unimodal. A median is any value  $m$  such that both  $P(X \leq m) \geq \frac{1}{2}$  and  $P(X \geq m) \geq \frac{1}{2}$ .

In the case of most continuous distributions, there is a unique median  $m$  and  $F(m) = P(X \leq m) = \frac{1}{2}$ .

There is a well-known empirical relationship that

$$\text{mean} - \text{mode} \cong 3(\text{mean} - \text{median})$$

or equivalently

$$\text{median} \cong (2 \text{ mean} + \text{mode})/3.$$

Some theoretical grounds for this relationship based on Gram–Charlier or Edgeworth expansions can be found in Lee (1991) or Kendall, Stewart and Ord (1987, Section 2.11).

Further material can be found in Rothschild and Logothetis (1986) or Evans, Hastings and Peacock (1993), with a more detailed account in Johnson *et al.* (2005), Johnson *et al.* (1994–1995), Balakrishnan *et al.* (2012) and Fang, Kotz and Wang (1989).

## A.1 Normal distribution

$X$  is normal with mean  $\theta$  and variance  $\phi$ , denoted  $X \sim N(\theta, \phi)$

if it has density

$$p(X) = \frac{1}{\sqrt{2\pi\phi}} \exp(-\frac{1}{2}(X - \theta)^2/\phi) \quad (-\infty < X < \infty).$$

The mean and variance are

$$\begin{aligned}\mathbb{E}X &= \theta \\ \mathbb{V}X &= \phi.\end{aligned}$$

Because the distribution is symmetrical and unimodal, the median and mode

$$\text{median}(X) = \theta$$

both equal the mean, that is,  $\text{mode}(X) = \theta$ .

If  $\theta = 0$  and  $\phi = 1$ , that is,  $X \sim N(0, 1)$ ,  $X$  is said to have a standard normal distribution A.2 Chi-squared distribution

$X$  has a chi-squared distribution on  $v$  degrees of freedom, denoted  $X \sim \chi_v^2$

if it has the same distribution as

$$Z_1^2 + Z_2^2 + \cdots + Z_v^2,$$

where  $Z_1, Z_2, \dots, Z_v$  are independent standard normal variates, or equivalently if

it has density  $p(X) = \frac{1}{2^{v/2}\Gamma(v/2)} X^{v/2-1} \exp(-\frac{1}{2}X) \quad (0 < X < \infty)$ .

If  $Y = X/S$ , where  $S$  is a constant, then  $Y$  is a chi-squared variate on  $v$  degrees of freedom divided by  $S$ , denoted  $Y \sim S^{-1}\chi_v^2$

and it has density

$$p(Y) = \frac{S^{\nu/2}}{2^{\nu/2}\Gamma(\nu/2)} Y^{\nu/2-1} \exp(-\frac{1}{2}SY) \quad (0 < Y < \infty).$$

The mean and variance are

$$\mathbb{E}Y = \nu/S$$

$$\mathcal{V}Y = 2\nu/S^2.$$

The mode is

$$\text{mode}(Y) = (\nu - 2)/S \quad (\text{provided } \nu \geq 2)$$

and the approximate relationship between mean, mode and median implies that the median is approximately  $\text{median}(Y) = (\nu - (\frac{2}{3})) / S$  at least for reasonably large  $\nu$ , say  $\nu \geq 5$ .

## A.3 Normal approximation to chi-squared

If  $X \sim \chi^2_\nu$  then for large  $\nu$  we have that approximately  $\sqrt{(2X)} - \sqrt{(2\nu - 1)} \sim N(0, 1)$  has a standard normal distribution.

## A.4 Gamma distribution

$X$  has a (one-parameter) gamma distribution with parameter  $\alpha$ , denoted  $X \sim G(\alpha)$

if it has density

$$p(X) = \frac{1}{\Gamma(\alpha)} X^{\alpha-1} \exp(-X) \quad (0 < X < \infty).$$

This is simply another name for the distribution, we refer to as

$$\frac{1}{2} \chi_{2\alpha}^2.$$

If  $Y = \beta X$ , then  $Y$  has a two-parameter gamma distribution with parameters  $\alpha$  and  $\beta$  denoted  $Y \sim G(\alpha, \beta)$

and it has density

$$p(Y) = \frac{1}{\beta^\alpha \Gamma(\alpha)} Y^{\alpha-1} \exp(-Y/\beta) \quad (0 < Y < \infty),$$

so that its mean and variance are

$$\begin{aligned}\mathbb{E}X &= \alpha\beta \\ \mathbb{V}X &= \alpha\beta^2.\end{aligned}$$

This is simply another name for the distribution, we refer to as

$$\frac{1}{2}\beta \chi_{2\alpha}^2.$$

If  $\beta = 1$ , we recover the one-parameter gamma distribution; if  $\alpha = 1$ , so that the density is  $p(Y) = \beta^{-1} \exp(-Y/\beta)$

we obtain another special case sometimes called the (negative) exponential distribution and denoted  $Y \sim \text{E}(\beta)$ .

The distribution function of any variable with a gamma distribution is easily found in terms of the incomplete gamma function  $\gamma(\alpha, x) = \int_0^x \xi^{\alpha-1} \exp(-\xi) d\xi$  or in terms of Karl Pearson's incomplete gamma function

$$I(u, p) = \frac{1}{\Gamma(p+1)} \int_0^{u\sqrt{p+1}} t^p \exp(-t) dt.$$

Extensive tables can be found in Pearson (1924).

## A.5 Inverse chi-squared distribution

$X$  has an inverse chi-squared distribution on  $v$  degrees of freedom, denoted  $X \sim \chi_v^{-2}$

if  $1/X \sim \chi_v^2$ , or equivalently if it has density

$$p(X) = \frac{1}{2^{v/2} \Gamma(v/2)} X^{-v/2-1} \exp(-\frac{1}{2}X^{-1}) \quad (0 < X < \infty).$$

If  $Y = SX$ , so that  $1/Y \sim S^{-1} \chi_v^2$ , then  $Y$  is  $S$  times an inverse chi-squared distribution on  $v$  degrees of freedom, denoted  $Y \sim S\chi_v^{-2}$

and it has density

$$p(Y) = \frac{S^{\nu/2}}{2^{\nu/2}\Gamma(\nu/2)} Y^{-\nu/2-1} \exp\left(-\frac{1}{2}S/Y\right) \quad (0 < Y < \infty).$$

The mean and variance are

$$\mathbb{E}Y = S/(\nu - 2) \quad (\text{provided } \nu > 2)$$

$$\mathcal{V}Y = \frac{2S^2}{(\nu - 2)^2(\nu - 4)} \quad (\text{provided } \nu > 4).$$

The mode is

$$\text{mode}(Y) = S/(v + 2)$$

and the median is in the range

$$S/\left(v - \left(\frac{1}{2}\right)\right) < \text{median}(Y) < S/\left(v - \left(\frac{2}{3}\right)\right)$$

provided  $v \geq 1$ , with the upper limit approached closely when  $v > 5$  [see Novick and Jackson (1974, Section 7.5)].

## A.6 Inverse chi distribution

$X$  has an inverse chi distribution on  $v$  degrees of freedom, denoted  $X \sim \chi_v^{-1}$   
if  $1/X^2 \sim \chi_v^2$ , or equivalently if it has density

$$p(X) = \frac{1}{2^{v/2-1} \Gamma(v/2)} X^{-v-1} \exp\left(-\frac{1}{2}X^{-2}\right) \quad (0 < X < \infty).$$

If  $Y = S^{1/2}X$ , so that  $1/Y^2 \sim S^{-1}\chi_v^2$ , then  $Y$  is  $S^{1/2}$  times an inverse chi distribution on  $v$  degrees of freedom, denoted  $Y \sim S^{1/2}\chi_v^{-1}$

and it has density

$$p(Y) = \frac{S^{\nu/2}}{2^{\nu/2-1} \Gamma(\nu/2)} Y^{-\nu-1} \exp(-\frac{1}{2} S/Y^2) \quad (0 < Y < \infty).$$

The mean and variance do not greatly simplify. They are

$$\mathbb{E}Y = \frac{S^{\frac{1}{2}} \Gamma(\nu - 1)/2}{\sqrt{2} \Gamma(\nu/2)}$$

$$\mathcal{V}Y = S/(\nu - 2) - (\mathbb{E}Y)^2$$

$$\mathbb{E}Y = S^{\frac{1}{2}} / (\nu - (\frac{3}{2}))$$

$$\mathcal{V}Y = \frac{S}{2(\nu - 2)(\nu - (\frac{5}{3}))}$$

but very good approximations, at least if  $\nu \geq 5$ , are

[see Novick and Jackson (1974, Section 7.3)]. The mode is exactly

$$\text{mode}(Y) = S^{\frac{1}{2}} / \sqrt{(\nu + 1)}$$

and a good approximation to the median at least if  $\nu \geq 4$  is (*ibid.*)

$$\text{median}(Y) = S^{\frac{1}{2}} / \sqrt{(\nu - \frac{2}{3})}.$$

## A.7 Log chi-squared distribution

$X$  has a log chi-squared distribution on  $\nu$  degrees of freedom, denoted  $X \sim \log \chi_{\nu}^2$

if  $X = \log W$  where  $W \sim \chi_{\nu}^2$ , or equivalently if  $X$  has density

$$p(X) = \frac{1}{2^{\nu/2} \Gamma(\nu/2)} \exp\left\{\frac{1}{2}\nu X - \frac{1}{2} \exp(X)\right\} \quad (-\infty < X < \infty)$$

(note that unlike  $\chi_{\nu}^2$  itself this is a distribution over the whole line).

Because the logarithm of an  $S^{-1} \chi_{\nu}^2$  variable differs from a log chi-squared variable simply by an additive constant, it is not necessary to consider such variables in any detail.

By considering the  $t$ th moment of a  $\chi_{\nu}^2$  variable, it is easily shown that the moment generating function of a log chi-squared variable is  $2^t \Gamma(t + (\nu/2)) / \Gamma(\nu/2)$ .

Writing

$$\psi(z) = \frac{d}{dz} \log \Gamma(z) = \frac{\Gamma'(z)}{\Gamma(z)}$$

for the so-called digamma function, it follows that the mean and variance are

$$\mathbb{E}X = \log 2 + \psi(v/2)$$

$$\mathbb{V}X = \psi'(v/2)$$

or (using Stirling's approximation and its derivatives) approximately

$$\mathbb{E}X = \log v - v^{-1} \cong \log(v - 1)$$

$$\mathbb{V}X = 2/v.$$

The mode is

$$\text{mode}(X) = \log v.$$

## A.8 Student's t distribution

$X$  has a Student's t distribution on  $v$  degrees of freedom, denoted  $X \sim t_v$  if it has the same distribution as

$$\frac{Z}{\sqrt{(W/v)}}$$

where  $Z \sim N(0, 1)$  and  $W \sim \chi^2_v$  are independent, or equivalently if  $X$  has density

$$p(X) = \frac{\Gamma((v+1)/2)}{\sqrt{(\pi v)\Gamma(v/2)}} \left(1 + \frac{X^2}{v}\right)^{-(v+1)/2}$$

$$= B\left(\frac{v}{2}, \frac{1}{2}\right)^{-1} \left(1 + \frac{X^2}{v}\right)^{-(v+1)/2}.$$

$$\begin{aligned}\bar{X} &= \sum X_i/n \\ S &= \sum (X_i - \bar{X})^2\end{aligned}$$

It follows that if  $X_1, X_2, \dots, X_n$  are independently  $N(\mu, \sigma^2)$  and  $s^2 = S/(n-1)$ ,

then

$$\frac{(\bar{X} - \mu)}{s/\sqrt{n}} \sim t_{n-1}.$$

The mean and variance are

$$\begin{aligned} \mathbb{E}X &= 0 \\ \mathbb{V}X &= v/(v - 2). \end{aligned}$$

Because the distribution is symmetrical and unimodal, the median and mode both equal the mean, that is  $\text{median}(X) = 0$  and  $\text{mode}(X) = 0$ .

As  $v \rightarrow \infty$  the distribution approaches the standard normal form.

It may be noted that Student's t distribution on one degree of freedom is the standard Cauchy distribution  $C(0, 1)$ .

## A.9 Normal/chi-squared distribution

The ordered pair  $(X, Y)$  has a normal/chi-squared distribution if  $Y \sim S\chi_v^{-2}$  for some  $S$  and  $v$  and, conditional on  $Y$ ,  $X \sim N(\mu, Y/n)$  for some  $\mu$  and  $n$ . An equivalent condition is that the joint density function (for

$$\begin{aligned} p(X, Y) &= \frac{1}{\sqrt{(2\pi Y/n)}} \exp\left(-\frac{1}{2}(X - \mu)^2/(Y/n)\right) \\ &\quad \times \frac{S^{v/2}}{2^{v/2}\Gamma(v/2)} Y^{v/2-1} \exp(-S/Y) \\ &= \frac{\sqrt{n}S^{v/2}}{\sqrt{\pi}2^{(v+1)/2}\Gamma(v/2)} Y^{-(v+1)/2-1} \\ &\quad \times \exp\left(-\frac{1}{2}\{S + n(X - \mu)^2\}/Y\right) \\ &\propto Y^{-(v+1)/2-1} \exp\left(-\frac{1}{2}Q/Y\right), \end{aligned}$$

$-\infty < X < \infty$  and  $0 < Y < \infty$ ) is

where

$$\begin{aligned}Q &= S + n(X - \mu)^2 \\&= nX^2 - 2(n\mu)X + (n\mu^2 + S).\end{aligned}$$

If we define

$$s^2 = S/v$$

then the marginal distribution of  $X$  is given by the fact that

$$\frac{X - \mu}{s/\sqrt{n}} \sim t_v.$$

The marginal distribution of  $Y$  is of course  $Y \sim S \chi_v^{-2}$ .

Approximate methods of constructing two-dimensional highest density regions for this distribution are described in Box and Tiao (1992, Section 2.4).

## A.10 Beta distribution

$X$  has a beta distribution with parameters  $\alpha$  and  $\beta$ , denoted  $X \sim \text{Be}(\alpha, \beta)$

if it has density

$$p(X) = \frac{1}{B(\alpha, \beta)} X^{\alpha-1} (1-X)^{\beta-1} \quad (0 < X < 1)$$

where the beta function  $B(\alpha, \beta)$  is defined by  $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)}$ .

The mean and variance are

$$\mathbb{E}X = \alpha/(\alpha + \beta)$$

$$\mathbb{V}X = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}.$$

The mode is

$$\text{mode}(X) = (\alpha - 1)/(\alpha + \beta - 2)$$

and the approximate relationship between mean, mode and median can be used to find an approximate median.

The distribution function of any variable with a beta distribution is easily found in terms of the incomplete beta function  $I_x(\alpha, \beta) = \int_0^x \frac{1}{B(\alpha, \beta)} \xi^{\alpha-1} (1-\xi)^{\beta-1} d\xi$ . Extensive tables can be found in Pearson (1968) or in Pearson and Hartley (1966, Table 17).

## A.11 Binomial distribution

$X$  has a binomial distribution of index  $n$  and parameter  $\pi$ , denoted  $x \sim B(n, \pi)$  if it has a discrete distribution with density

$$p(X) = \binom{n}{X} \pi^X (1-\pi)^{n-X} \quad (X = 0, 1, 2, \dots, n).$$

The mean and variance are

$$\mathbb{E}X = n\pi$$
$$\mathcal{V}X = n\pi(1 - \pi).$$

Because

$$\frac{p(X+1)}{p(X)} = \frac{(n-X)\pi}{(X+1)(1-\pi)}$$

we see that  $p(X+1) > p(X)$  if and only if  $X < (n+1)\pi - 1$ ,

and hence that a mode occurs at

$$\text{mode}(X) = [(n+1)\pi]$$

the square brackets denoting ‘integer part of’, and this mode is unique unless  $(n+1)\pi$  is an integer.

Integration by parts shows that the distribution function is expressible in terms of the incomplete beta function, namely,

$$\sum_{\xi=1}^x \binom{n}{\xi} \pi^\xi (1-\pi)^{n-\xi} = I_{1-\pi}(n-x, x+1) = 1 - I_\pi(x+1, n-x)$$

see, for example, Kendall, Stewart and Ord (1987, Section 5.7).

## A.12 Poisson distribution

$X$  has a Poisson distribution of mean  $\lambda$ , denoted  $X \sim P(\lambda)$

if it has a discrete distribution with density

$$p(X) = \frac{\lambda^X}{X!} \exp(-\lambda) \quad (X = 0, 1, 2, \dots).$$

The mean and variance are

$$\mathbb{E}X = \lambda$$

$$\mathcal{V}X = \lambda.$$

Because

$$p(X+1)/p(X) = \lambda/(X+1)$$

we see that  $p(X+1) > p(X)$  if and only if  $X < \lambda - 1$ ,

and hence that a mode occurs at

$$\text{mode}(X) = [\lambda]$$

the square brackets denoting ‘integer part of’, and this mode is unique unless  $\lambda$  is an integer.

Integrating by parts shows that the distribution function is expressible in terms

of the incomplete gamma function, namely,  $\sum_{\xi=0}^x \frac{\lambda^\xi}{\xi!} \exp(-\lambda) = 1 - \frac{\gamma(x+1, \lambda)}{\Gamma(x+1)}$ ;  
see Kendall, Stewart and Ord (1987, Section 5.9).

The Poisson distribution often occurs as the limit of the binomial as

$$n \rightarrow \infty, \quad \pi \rightarrow 0, \quad n\pi \rightarrow \lambda.$$

## A.13 Negative binomial distribution

$X$  has a negative binomial distribution of index  $n$  and parameter  $\pi$ , denoted  $X \sim \text{NB}(n, \pi)$

if it has a discrete distribution with density

$$p(X) = \binom{n+X-1}{X} \pi^n (1-\pi)^X \quad (X = 0, 1, 2, \dots).$$

Because

$$(1+z)^{-n} = \sum_{x=0}^{\infty} \binom{n+x-1}{x} (-1)^x z^x$$

we sometimes use the notation

$$\binom{-n}{x} = \binom{n+x-1}{x} (-1)^x.$$

The mean and variance are

$$\mathbb{E}X = n(1 - \pi)/\pi$$

$$\mathcal{V}X = n(1 - \pi)/\pi^2.$$

Because

$$\frac{p(X+1)}{p(X)} = \frac{(n+X)}{(X+1)}(1-\pi)$$

we see that  $p(X+1) > p(X)$  if and only if  $X < \{n(1-\pi) - 1\}/\pi$ ,

and hence that a mode occurs at

$$\text{mode}(X) = [(n-1)(1-\pi)/\pi]$$

the square brackets denoting ‘integer part of’, and this mode is unique unless  $(n-1)(1-\pi)/\pi$  is an integer.

It can be shown that the distribution function can be found in terms of that of the binomial distribution, or equivalently in terms of the incomplete beta function; for details see Balakrishnan *et al.* (1992, Chapter 5, Section 6). Just as the Poisson distribution can arise as a limit of the binomial distribution, so it can as a limit of the negative binomial, but in this case as  $n \rightarrow \infty$ ,  $1-\pi \rightarrow 0$ ,  $n(1-\pi) \rightarrow \lambda$ .

The particular case where  $n=1$ , so that  $p(X) = \pi(1-\pi)^X$ , is sometimes referred to as the *geometric distribution*.

## A.14 Hypergeometric distribution

$X$  has a hypergeometric distribution of population size  $N$ , index  $n$  and parameter  $\pi$ , denoted  $X \sim H(N, n, \pi)$

if it has a discrete distribution with density

$$p(X) = \frac{\binom{N\pi}{X} \binom{N(1-\pi)}{n-X}}{\binom{N}{n}} \quad (X = 0, 1, 2, \dots, n).$$

The mean and variance are

$$\mathbb{E}X = n\pi$$

$$\mathcal{V}X = n\pi(1 - \pi)(N - n)/(N - 1).$$

Because

$$\frac{p(X+1)}{p(X)} = \frac{(n-X)(N\pi - X)}{(X+1)[N(1-\pi) - n + X + 1]}$$

we see that  $p(X+1) > p(X)$  if and only if  $X < (n+1)\pi - 1 + (n+1)(1-2\pi)/(N+2)$ ,  
and hence that if, as is usually the case,  $N$  is fairly large, if and only if  
 $X < (n+1)\pi - 1$ .

Hence, the mode occurs very close to the binomial value

$$\text{mode}(X) = [(n+1)\pi].$$

As  $N \rightarrow \infty$  this distribution approaches the binomial distribution  $B(n, \pi)$ .

Tables of it can be found in Lieberman and Owen (1961).

## A.15 Uniform distribution

$X$  has a uniform distribution on the interval  $(a, b)$  denoted  $X \sim U(a, b)$

if it has density

$$p(X) = (b - a)^{-1} I_{(a,b)}(X),$$

where

$$I_{(a,b)}(X) = \begin{cases} 1 & (X \in (a, b)) \\ 0 & (X \notin (a, b)) \end{cases}$$

is the *indicator function* of the set  $(a, b)$ . The mean and variance are

$$\mathbb{E}X = \frac{1}{2}(a + b)$$

$$\mathbb{V}X = (b - a)^2/12.$$

There is no unique mode, but the distribution is symmetrical, and hence  
 $\text{median}(X) = \frac{1}{2}(a + b)$ .

Sometimes we have occasion to refer to a discrete version;  $Y$  has a discrete uniform distribution on the interval  $[a, b]$  denoted  $Y \sim \text{UD}(a, b)$   
if it has a discrete distribution with density

$$p(Y) = (b - a + 1)^{-1} \quad (Y = a, a + 1, \dots, b).$$

The mean and variance are

$$\mathbb{E}X = \frac{1}{2}(a + b)$$

$$\mathbb{V}X = (b - a)(b - a + 2)/12.$$

using formulae for the sum and sum of squares of the first  $n$  natural numbers [the variance is best found by noting that the variance of  $\text{UD}(a, b)$  equals that of  $\text{UD}(1, n)$  where  $n=b-a+1$ ]. Again, there is no unique mode, but the distribution is symmetrical, and hence  $\text{median}(Y) = \frac{1}{2}(a + b)$ .

## A.16 Pareto distribution

$X$  has a Pareto distribution with parameters  $\xi$  and  $\gamma$ , denoted  $X \sim \text{Pa}(\xi, \gamma)$

if it has density

$$p(X) = \gamma \xi^\gamma X^{-\gamma-1} I_{(\xi, \infty)}(X),$$

where

$$I_{(\xi, \infty)}(X) = \begin{cases} 1 & (X > \xi) \\ 0 & (\text{otherwise}) \end{cases}$$

is the indicator function of the set  $(\xi, \infty)$ . The mean and variance are

$$\mathbb{E}X = \gamma \xi / (\gamma - 1) \quad (\text{provided } \gamma > 1)$$

$$\mathbb{V}X = \frac{\gamma \xi^2}{(\gamma - 1)^2(\gamma - 2)} \quad (\text{provided } \gamma > 2)$$

The distribution function is

$$F(x) = [1 - (\xi/x)^\gamma] I_{(\xi, \infty)}(x)$$

and in particular the median is

$$\text{median}(X) = 2^{1/\gamma} \xi.$$

The mode, of course, occurs at

$$\text{mode}(X) = \xi.$$

The ordered pair  $(Y, Z)$  has a bilateral bivariate Pareto distribution with parameters  $\xi$ ,  $\eta$  and  $\gamma$ , denoted  $(Y, Z) \sim \text{Pabb}(\xi, \eta, \gamma)$

if it has joint density function

$$p(Y, Z) = \gamma(\gamma + 1)(\xi - \eta)^\gamma (Y - Z)^{-\gamma - 2} I_{(\xi, \infty)}(Y) I_{(-\infty, \eta)}(Z).$$

The means and variances are

$$\begin{aligned}\mathbb{E}X &= (\gamma\xi - \eta)/(\gamma - 1) \\ \mathbb{E}Z &= (\gamma\eta - \xi)/(\gamma - 1) \\ \mathcal{V}Y = \mathcal{V}Z &= \frac{\gamma(\xi - \eta)^2}{(\gamma - 1)^2(\gamma - 2)}\end{aligned}$$

and the correlation coefficient between  $Y$  and  $Z$  is  $\rho(Y, Z) = -1/\gamma$ .

It is also sometimes useful that

$$\mathbb{E}(Y - Z) = (\gamma + 1)(\xi - \eta)/(\gamma - 1)$$

$$\mathcal{V}(Y - Z) = \frac{2(\gamma + 1)(\xi - \eta)^2}{(\gamma - 1)^2(\gamma - 2)}.$$

The marginal distribution function of  $Y$  is  $F(y) = \left[1 - \left(\frac{\xi - \eta}{y - \eta}\right)^\gamma\right] I_{(\xi, \infty)}(y)$   
and in particular the median is

$$\text{median}(Y) = \eta + 2^{1/\gamma}(\xi - \eta).$$

The distribution function of  $Z$  is similar, and in particular the median is  
 $\text{median}(Z) = \xi - 2^{1/\gamma}(\xi - \eta)$ .

The modes, of course, occur at

$$\text{mode}(Y) = \xi$$

$$\text{mode}(Z) = \eta.$$

The distribution is discussed in DeGroot (1970, Sections 4.11, 5.7 and 9.7).

## A.17 Circular normal distribution

$X$  has a circular normal or von Mises' distribution with mean  $\mu$  and concentration parameter  $\kappa$ , denoted  $X \sim \text{M}(\mu, \kappa)$

if it has density

$$p(X) = (2\pi I_0(\kappa))^{-1} \exp(\kappa \cos(X - \mu)),$$

where  $X$  is any angle, so  $0 < X < 2\pi$  and  $I_0(\kappa)$  is a constant called the modified Bessel function of the first kind and order zero (`besselI(kappa, 0)` in R). It turns out

$$\text{that } I_0(\kappa) = \sum_{r=0}^{\infty} \frac{1}{(r!)^2} \left(\frac{1}{2}\kappa\right)^{2r}$$

and that asymptotically for large  $\kappa$

$$I_0(\kappa) \sim \frac{1}{\sqrt{(2\pi\kappa)}} \exp(\kappa).$$

For large  $\kappa$ , we have approximately

$$X \sim N(\mu, 1/\kappa)$$

while for small  $\kappa$ , we have approximately  $p(X) = (2\pi)^{-1} \{1 + \frac{1}{2}\kappa \cos(X - \mu)\}$

which density is sometimes referred to as a cardioid distribution. The circular normal distribution is discussed by Mardia (1972), Mardia and Jupp (2001) and Batschelet (1981).

One point related to this distribution arises in a Bayesian context in connection with the reference prior  $p(\mu, \kappa) \propto 1$

$$\begin{aligned} c &= n^{-1} \sum \cos X_i \\ s &= n^{-1} \sum \sin X_i \\ \rho &= \sqrt{(c^2 + s^2)} \end{aligned}$$

when we have observations  $X_1, X_2, \dots, X_n$  such that  $\hat{\mu} = \tan^{-1}(s/c)$ .

The only sensible estimator of  $\mu$  on the basis of the posterior distribution is  $\hat{\mu}$ .

The mode of the posterior distribution of  $\kappa$  is approximately  $\hat{\kappa}$ , where

$$\hat{\kappa} = \begin{cases} \rho \left( \frac{2-\rho^2}{1-\rho^2} \right) & (\rho < 2/3) \\ \frac{\rho+1}{4\rho(1-\rho)} & (\rho > 2/3) \end{cases}$$

(both of which are approximately 1.87 when  $\rho = 2/3$ ) according to Schmitt (1969, Section 10.2). Because of the skewness of the distribution of  $\kappa$ , its posterior mean is greater than its posterior mode.

## A.18 Behrens' distribution

$X$  is said to have Behrens' (or Behrens–Fisher or Fisher–Behrens) distribution with degrees of freedom  $v_1$  and  $v_2$  and angle  $\phi$ , denoted  $X \sim BF(v_1, v_2, \phi)$  if  $X$  has the same distribution as

$$T_2 \cos \phi - T_1 \sin \phi,$$

where  $T_1$  and  $T_2$  are independent and  $T \sim t_{\nu_1}$  and  $T \sim t_{\nu_2}$ .

Equivalently,  $X$  has density

$$p(X) = k \int_{-\infty}^{\infty} g(z) dz,$$

where

$$g(z) = \left[ 1 + \frac{(z \cos \phi - X \sin \phi)^2}{\nu_1} \right]^{-(\nu_1+1)/2} \left[ 1 + \frac{(z \sin \phi + X \cos \phi)^2}{\nu_2} \right]^{-(\nu_2+1)/2}$$

over the whole real line, where

$$k^{-1} = B\left(\frac{1}{2}\nu_1, \frac{1}{2}\right) B\left(\frac{1}{2}\nu_2, \frac{1}{2}\right) \sqrt{(\nu_1\nu_2)}.$$

This distribution naturally arises as the posterior distribution of

$$\theta_2 - \theta_1$$

when we have samples of size  $n_1 = \nu_1 + 1$  from  $N(\theta_1, \sigma_1^2)$  and of size  $n_2 = \nu_2 + 1$  from  $N(\theta_2, \sigma_2^2)$  and neither  $\sigma_1^2$  nor  $\sigma_2^2$  is known, and conventional priors are adopted. In this case, in a fairly obvious notation

$$T_1 = \frac{\theta_1 - \bar{X}_1}{s_1/\sqrt{n_1}} \quad \text{and} \quad T_2 = \frac{\theta_2 - \bar{X}_2}{s_2/\sqrt{n_2}}$$

$$\tan \phi = \frac{s_1/\sqrt{n_1}}{s_2/\sqrt{n_2}}$$

An approximation to this distribution due to Patil (1965) is as follows.

Define

$$\begin{aligned}f_1 &= \left(\frac{\nu_2}{\nu_2 - 2}\right) \cos^2 \phi + \left(\frac{\nu_1}{\nu_1 - 2}\right) \sin^2 \phi \\f_2 &= \frac{\nu_2^2}{(\nu_2 - 2)^2(\nu_2 - 4)} \cos^4 \phi + \frac{\nu_1^2}{(\nu_1 - 2)^2(\nu_1 - 4)} \sin^4 \phi \\a &= \sqrt{\{f_1(b-2)/b\}} \\b &= 4 + (f_1^2/f_2) \\s^2 &= (s_1^2/n_1) + (s_2^2/n_2) \\T &= \frac{(\theta_2 - \theta_1) - (\bar{X}_2 - \bar{X}_1)}{s} = T_1 \sin \phi - T_2 \cos \phi.\end{aligned}$$

Then, approximately,

$$T/a \sim t_b.$$

Obviously  $b$  is usually not an integer, and consequently this approximation requires interpolation in the  $t$  tables.

Clearly Behrens' distribution has mean and variance

$$\mathbb{E}X = 0$$

$$\mathbb{V}X = \frac{\nu_2 \cos^2 \phi}{\nu_2 - 2} + \frac{\nu_1 \sin^2 \phi}{\nu_1 - 2}$$

using the mean and variance of  $t$  distributions and the independence of  $T_1$  and  $T_2$ . The distribution is symmetrical and unimodal and hence the mean, mode and median are all equal, so  $\text{median}(X) = 0$ .

## A.19 Snedecor's F distribution

$X$  has an F distribution on  $\nu_1$  and  $\nu_2$  degrees of freedom, denoted  $X \sim F_{\nu_1, \nu_2}$  if  $X$  has the same distribution as

$$\frac{W_1/\nu_1}{W_2/\nu_2},$$

where  $W_1$  and  $W_2$  are independent and  $W_1 \sim \chi_{\nu_1}^2$  and  $W_2 \sim \chi_{\nu_2}^2$ .

Equivalently,  $X$  has density

$$p(X) = \frac{\nu_1^{(\nu_1/2)} \nu_2^{(\nu_2/2)}}{B\left(\frac{1}{2}\nu_1, \frac{1}{2}\nu_2\right)} \frac{X^{\nu_1/2-1}}{(\nu_2 + \nu_1 X)^{(\nu_1+\nu_2)/2}}.$$

The mean and variance are

$$\mathbb{E}X = v_2/(v_2 - 2)$$

$$\mathcal{V}X = \frac{2v_2^2(v_1 + v_2 - 2)}{v_1(v_2 - 2)^2(v_2 - 4)}.$$

The mode is

$$\text{mode}(X) = \frac{\nu_2}{\nu_2 + 1} \frac{\nu_1 - 2}{\nu_1}.$$

If  $X \sim F_{\nu_1, \nu_2}$ , then

$$\frac{\nu_1 X}{\nu_2 + \nu_1 X} \sim \text{Be}\left(\frac{1}{2}\nu_1, \frac{1}{2}\nu_2\right).$$

Conversely, if  $Y \sim \text{Be}(\delta_1, \delta_2)$  then

$$X = \frac{\delta_2 Y}{\delta_1(1 - Y)} \sim F_{2\delta_1, 2\delta_2}.$$

## A.20 Fisher's z distribution

$X$  has a z distribution on  $\nu_1$  and  $\nu_2$  degrees of freedom, denoted  $X \sim z_{\nu_1, \nu_2}$

if  $Y = \exp(2X) \sim F_{\nu_1, \nu_2}$ , or equivalently if it has the density

$$p(X) = 2 \frac{\nu_1^{\nu_1/2} \nu_2^{\nu_2/2}}{B(\nu_1/2, \nu_2/2)} \frac{e^{\nu_1 X}}{(\nu_2 + \nu_1 e^{2X})^{(\nu_1+\nu_2)/2}}.$$

Another definition is that if  $Y \sim F_{\nu_1, \nu_2}$ , then  $X = \frac{1}{2} \log Y \sim z_{\nu_1, \nu_2}$ .

The mean and variance are easily deduced from those of the log chi-squared

$$\mathbb{E}X = \frac{1}{2} \log(\nu_2/\nu_1) + \frac{1}{2}\psi(\nu_1/2) - \frac{1}{2}\psi(\nu_2/2)$$

distribution; they are  $\mathbb{V}X = \frac{1}{4}\psi'(\nu_1/2) + \frac{1}{4}\psi'(\nu_2/2)$ ,

where  $\psi$  is (as above) the digamma function, or approximately  $\mathbb{E}X = \frac{1}{2}(\nu_2^{-1} - \nu_1^{-1}) \cong \frac{1}{2} \log \left[ (1 - \nu_1^{-1}) / (1 - \nu_2^{-1}) \right]$

$$\mathbb{V}X = \frac{1}{2}(\nu_1^{-1} + \nu_2^{-1}).$$

The mode is zero.

Unless  $\nu_1$  and  $\nu_2$  are very small, the distribution of z is approximately normal. The z distribution was introduced by Fisher (1924).

## A.21 Cauchy distribution

$X$  has a Cauchy distribution with location parameter  $\mu$  and scale parameter  $\sigma^2$ , denoted  $X \sim C(\mu, \sigma^2)$

if

$$p(X) = \frac{1}{\pi} \frac{\sigma}{\sigma^2 + (X - \mu)^2}.$$

Because the relevant integral is not absolutely convergent, this distribution does not have a finite mean, nor *a fortiori*, a finite variance. However, it is symmetrical about  $\mu$ , and hence  $\text{median}(X) = \mu$   $\text{mode}(X) = \mu$ .

The distribution function is

$$F(x) = \frac{1}{2} + \pi^{-1} \tan^{-1} \{(x - \mu)/\sigma\},$$

so that when  $x = \mu - \sigma$  then  $F(x)=1/4$  and when  $x = \mu + \sigma$  then  $F(x)=3/4$ . Thus,  $\mu - \sigma$  and  $\mu + \sigma$  are, respectively, the lower and upper quartiles, and hence  $\sigma$  may be thought of as the *semi-interquartile range*. Note that for a normal  $N(\mu, \sigma^2)$  distribution the semi-interquartile range is  $0.67449\sigma$  rather than  $\sigma$ .

It may be noted that the  $C(0, 1)$  distribution is also the Student's t distribution on 1 degree of freedom A.22 The probability that one beta variable is greater than another Suppose  $\pi$  and  $\rho$  have independent beta distributions  $\pi \sim Be(\alpha, \beta)$ ,  $\rho \sim Be(\gamma, \delta)$ .

Then

$$P(\pi < \rho) = \sum_{\kappa=\max(\gamma-\beta, 0)}^{\gamma-1} \frac{\binom{\gamma+\delta-1}{\kappa} \binom{\alpha+\beta-1}{\alpha+\gamma-1-\kappa}}{\binom{\alpha+\beta+\gamma+\delta-2}{\alpha+\gamma-1}}$$

[see Altham (1969)]. For an expression (albeit a complicated one) for  $P(\pi/\rho \leq c)$ ; see Weisberg (1972).

When  $\alpha, \beta, \gamma$  and  $\delta$  are large we can approximate the beta variables by normal variates of the same means and variances and hence approximate the distribution of  $\pi - \rho$  by a normal distribution.

## A.23 Bivariate normal distribution

The ordered pair  $(X, Y)^T$  of observations has a bivariate normal distribution, denoted  $\begin{pmatrix} X \\ Y \end{pmatrix} \sim N \left( \begin{pmatrix} \mu \\ \nu \end{pmatrix}, \begin{pmatrix} \phi & \rho\sqrt{\phi\psi} \\ \rho\sqrt{\phi\psi} & \psi \end{pmatrix} \right)$

if it has joint density

$$p(X, Y | \lambda, \mu, \phi, \psi, \rho) = \frac{1}{2\pi \sqrt{\{\phi\psi(1 - \rho^2)\}}} \exp\left(-\frac{1}{2(1 - \rho^2)} Q\right),$$

where

$$Q = \frac{(X - \lambda)^2}{\phi} - \frac{2\rho(X - \lambda)(Y - \mu)}{\sqrt{\phi\psi}} + \frac{(Y - \mu)^2}{\psi}.$$

The means and variances are

$$\mathbb{E}X = \lambda, \quad \mathbb{E}Y = \mu, \quad \mathbb{V}X = \phi, \quad \mathbb{V}Y = \psi$$

and  $X$  and  $Y$  have correlation coefficient and covariance  $\rho(X, Y) = \rho$ ,  $\mathcal{C}(X, Y) = \rho\sqrt{\phi\psi}$ .

Most properties follow from those of the ordinary, univariate and normal distribution. One point worth noting (which is clear from the form of the joint density function) is that if  $X$  and  $Y$  have a bivariate normal distribution, then they are independent if they are uncorrelated (a result which is *not* true in general).

## A.24 Multivariate normal distribution

An  $n$ -dimensional random vector  $X$  has a multivariate normal distribution with mean vector  $\mu$  and variance–covariance matrix  $\Sigma$ , denoted  $X \sim N(\mu, \Sigma)$  if it has joint density function

$$\frac{1}{(2\pi)^{n/2}\sqrt{\det \Sigma}} \exp\left\{-\frac{1}{2}(X - \mu)^T \Sigma^{-1} (X - \mu)\right\}.$$

It can be checked by finding the determinant of the  $2 \times 2$  variance–covariance

$$\text{matrix } \Sigma = \begin{pmatrix} \phi & \rho\sqrt{\phi\psi} \\ \rho\sqrt{\phi\psi} & \psi \end{pmatrix}$$

and inverting it that the bivariate normal distribution is a special case.

## A.25 Distribution of the correlation coefficient

If the prior density of the correlation coefficient is  $\rho$ , then its posterior density, given  $n$  pairs of observations  $(X_i, Y_i)$  with sample correlation coefficient  $r$ , is

$$\begin{aligned} p(\rho | X, Y) &\propto p(\rho)(1 - \rho^2)^{(n-1)/2} \int_0^\infty \omega^{-1} (\omega + \omega^{-1} - 2\rho r)^{-(n-1)} d\omega \\ &\propto p(\rho)(1 - \rho^2)^{(n-1)/2} \int_0^\infty (\cosh t - \rho r)^{-(n-1)} dt \\ &\propto p(\rho)(1 - \rho^2)^{(n-1)/2} \int_0^\infty (\cosh t + \cos \theta)^{-(n-1)} dt \end{aligned}$$

given by

on writing  $-\rho r = \cos \theta$ . It can also be shown that

$$p(\rho | X, Y) \propto p(\rho)(1 - \rho^2)^{(n-1)/2} \left(\frac{\partial}{\sin \theta \partial \theta}\right)^{n-2} \left(\frac{\theta}{\sin \theta}\right).$$

When  $r=0$  the density simplifies to

$$p(\rho | X, Y) \propto p(\rho)(1 - \rho^2)^{(n-1)/2}$$

and so if the prior is of the form

$$p(\rho) \propto (1 - \rho^2)^{(v_0-3)/2},$$

it can be shown that

$$(\nu_0 + n - 2)^{\frac{1}{2}} \frac{\rho}{(1 - \rho^2)^{\frac{1}{2}}} \sim t_{\nu_0+n-2}$$

has a Student's t distribution on  $\nu_0 + n - 2$  degrees of freedom.

Going back to the general case, it can be shown that

$$\begin{aligned} p(\rho | X, Y) &\propto p(\rho) \frac{(1 - \rho^2)^{(n-1)/2}}{(1 - \rho r)^{n-(3/2)}} \\ &\quad \times \int_0^1 \frac{(1-u)^{n-2}}{(2u)^{\frac{1}{2}}} [1 - \frac{1}{2}(1 + \rho r)u]^{-\frac{1}{2}} du. \end{aligned}$$

Expanding the term in square brackets as a power series in  $u$ , we can express the last integral as a sum of beta functions. Taking only the first term, we have as an

$$\text{approximation } p(\rho | X, Y) \propto p(\rho) \frac{(1 - \rho^2)^{(n-1)/2}}{(1 - \rho r)^{n-(3/2)}}.$$

On writing

$$\rho = \tanh \zeta, \quad r = \tanh z,$$

it can be shown that

$$p(\zeta | X, Y) \propto \frac{p(\zeta)}{\cosh^{5/2}(\zeta) \cosh^{n-(3/2)}(\zeta - z)},$$

and hence that for large  $n$

$$\zeta \sim N(z, 1/n)$$

approximately [whatever  $p(\rho)$  is]. A better approximation is

$$\zeta \sim N\left(z - 5r/2n, \left\{n - \left(\frac{3}{2}\right) + \left(\frac{5}{2}\right)(1 - r^2)\right\}^{-1}\right).$$

# Appendix B: Tables

**Table B.1** Percentage points of the Behrens–Fisher distribution.

	$\psi = 0^\circ$					$\psi = 15^\circ$					
	$v_2$	75%	90%	95%	97.5%	$v_2$	75%	90%	95%	97.5%	
$v_1 = 6$	6	0.72	1.44	1.94	2.45	$v_1 = 6$	6	0.72	1.45	1.95	2.45
	8	0.71	1.40	1.86	2.31		8	0.72	1.41	1.87	2.32
	12	0.70	1.36	1.78	2.18		12	0.71	1.37	1.80	2.19
	24	0.68	1.32	1.71	2.06		24	0.69	1.34	1.73	2.09
	$\infty$	0.67	1.28	1.65	1.96		$\infty$	0.68	1.30	1.67	2.00
	$\psi = 0^\circ$					$\psi = 15^\circ$					
	$v_2$	75%	90%	95%	97.5%	$v_2$	75%	90%	95%	97.5%	
$v_1 = 8$	6	0.72	1.44	1.94	2.45	$v_1 = 8$	6	0.72	1.44	1.94	2.44
	8	0.71	1.40	1.86	2.31		8	0.71	1.40	1.86	2.31
	12	0.70	1.36	1.78	2.18		12	0.70	1.37	1.79	2.18
	24	0.68	1.32	1.71	2.06		24	0.69	1.33	1.72	2.08
	$\infty$	0.67	1.28	1.65	1.96		$\infty$	0.68	1.30	1.66	1.98
	$\psi = 0^\circ$					$\psi = 15^\circ$					
	$v_2$	75%	90%	95%	97.5%	$v_2$	75%	90%	95%	97.5%	
$v_1 = 12$	6	0.72	1.44	1.94	2.45	$v_1 = 12$	6	0.72	1.44	1.94	2.43
	8	0.71	1.40	1.86	2.31		8	0.71	1.40	1.86	2.30
	12	0.70	1.36	1.78	2.18		12	0.70	1.36	1.78	2.18
	24	0.68	1.32	1.71	2.06		24	0.69	1.32	1.72	2.07
	$\infty$	0.67	1.28	1.65	1.96		$\infty$	0.68	1.29	1.66	1.98

		$\psi = 0^\circ$				$\psi = 15^\circ$					
		$v_2$	75%	90%	95%	97.5%	$v_2$	75%	90%	95%	97.5%
$v_1 = 24$	6	0.72	1.44	1.94	2.45		6	0.72	1.43	1.93	2.43
	8	0.71	1.40	1.86	2.31	$v_1 = 24$	8	0.71	1.39	1.85	2.29
	12	0.70	1.36	1.78	2.18		12	0.70	1.36	1.78	2.17
	24	0.68	1.32	1.71	2.06		24	0.69	1.32	1.71	2.06
	$\infty$	0.67	1.28	1.65	1.96		$\infty$	0.68	1.29	1.65	1.97
		$\psi = 0^\circ$				$\psi = 15^\circ$					
		$v_2$	75%	90%	95%	97.5%	$v_2$	75%	90%	95%	97.5%
$v_1 = \infty$	6	0.72	1.44	1.94	2.45		6	0.71	1.43	1.93	2.42
	8	0.71	1.40	1.86	2.31	$v_1 = \infty$	8	0.70	1.39	1.85	2.29
	12	0.70	1.36	1.78	2.18		12	0.70	1.35	1.77	2.16
	24	0.68	1.32	1.71	2.06		24	0.69	1.32	1.71	2.06
	$\infty$	0.67	1.28	1.65	1.96		$\infty$	0.67	1.28	1.65	1.96
		$\psi = 30^\circ$				$\psi = 45^\circ$					
		$v_2$	75%	90%	95%	97.5%	$v_2$	75%	90%	95%	97.5%
$v_1 = 6$	6	0.74	1.47	1.96	2.45		6	0.75	1.48	1.97	2.45
	8	0.73	1.44	1.90	2.34	$v_1 = 6$	8	0.74	1.45	1.93	2.37
	12	0.72	1.40	1.84	2.24		12	0.72	1.42	1.88	2.32
	24	0.71	1.37	1.79	2.16		24	0.71	1.39	1.84	2.27
	$\infty$	0.69	1.34	1.74	2.10		$\infty$	0.70	1.37	1.81	2.23
		$v = 30^\circ$				$v = 45^\circ$					
		$v_2$	75%	90%	95%	97.5%	$v_2$	75%	90%	95%	97.5%
$v_1 = 8$	6	0.73	1.45	1.94	2.41		6	0.74	1.45	1.93	2.37
	8	0.73	1.42	1.87	2.30	$v_1 = 8$	8	0.73	1.43	1.88	2.30
	12	0.72	1.39	1.81	2.20		12	0.72	1.40	1.84	2.23
	24	0.70	1.36	1.76	2.12		24	0.71	1.38	1.80	2.18
	$\infty$	0.69	1.32	1.71	2.05		$\infty$	0.70	1.35	1.77	2.14
		$v = 30^\circ$				$v = 45^\circ$					
		$v_2$	75%	90%	95%	97.5%	$v_2$	75%	90%	95%	97.5%
$v_1 = 12$	6	0.72	1.43	1.91	2.39		6	0.72	1.42	1.88	2.32
	8	0.72	1.40	1.85	2.27	$v_1 = 12$	8	0.72	1.40	1.84	2.23
	12	0.71	1.37	1.79	2.17		12	0.71	1.38	1.79	2.17
	24	0.70	1.34	1.73	2.09		24	0.70	1.35	1.75	2.11
	$\infty$	0.69	1.31	1.69	2.01		$\infty$	0.69	1.33	1.72	2.07

		$\nu = 30^\circ$				$\nu = 45^\circ$					
		$v_2$	75%	90%	95%	97.5%	$v_2$	75%	90%	95%	97.5%
$v_1 = 24$	6	0.71	1.42	1.89	2.36		6	0.71	1.39	1.84	2.27
	8	0.71	1.39	1.83	2.25	$v_1 = 24$	8	0.71	1.38	1.80	2.18
	12	0.70	1.36	1.77	2.15		12	0.70	1.35	1.75	2.11
	24	0.69	1.33	1.71	2.06		24	0.69	1.33	1.71	2.06
	$\infty$	0.68	1.30	1.66	1.99		$\infty$	0.68	1.30	1.68	2.01
		$\nu = 30^\circ$				$\nu = 45^\circ$					
		$v_2$	75%	90%	95%	97.5%	$v_2$	75%	90%	95%	97.5%
$v_1 = \infty$	6	0.70	1.40	1.88	2.34		6	0.70	1.37	1.81	2.23
	8	0.70	1.37	1.81	2.23	$v_1 = \infty$	8	0.70	1.35	1.77	2.14
	12	0.69	1.34	1.75	2.13		12	0.69	1.33	1.72	2.07
	24	0.69	1.31	1.70	2.04		24	0.68	1.30	1.68	2.01
	$\infty$	0.67	1.28	1.65	1.96		$\infty$	0.67	1.28	1.65	1.96

**Table B.2** Highest density regions for the chi-squared distribution.

$v$	50%		60%		67%		70%		75%	
3	0.259	2.543	0.170	3.061	0.120	3.486	0.099	3.731	0.070	4.155
4	0.871	3.836	0.684	4.411	0.565	4.876	0.506	5.143	0.420	5.603
5	1.576	5.097	1.315	5.730	1.141	6.238	1.052	6.527	0.918	7.023
6	2.327	6.330	2.004	7.016	1.783	7.563	1.669	7.874	1.493	8.404
7	3.107	7.540	2.729	8.276	2.467	8.860	2.331	9.190	2.118	9.753
8	3.907	8.732	3.480	9.515	3.181	10.133	3.025	10.482	2.779	11.075
9	4.724	9.911	4.251	10.737	3.917	11.387	3.742	11.754	3.465	12.376
10	5.552	11.079	5.037	11.946	4.671	12.626	4.479	13.009	4.174	13.658
11	6.391	12.238	5.836	13.143	5.440	13.851	5.231	14.250	4.899	14.925
12	7.238	13.388	6.645	14.330	6.221	15.066	5.997	15.480	5.639	16.179
13	8.093	14.532	7.464	15.508	7.013	16.271	6.774	16.699	6.392	17.422
14	8.954	15.669	8.290	16.679	7.814	17.467	7.560	17.909	7.155	18.654
15	9.821	16.801	9.124	17.844	8.622	18.656	8.355	19.111	7.928	19.878
16	10.692	17.929	9.963	19.003	9.438	19.838	9.158	20.306	8.709	21.094
17	11.568	19.052	10.809	20.156	10.260	21.014	9.968	21.494	9.497	22.303
18	12.448	20.171	11.659	21.304	11.088	22.184	10.783	22.677	10.293	23.505
19	13.331	21.287	12.514	22.448	11.921	23.349	11.605	23.854	11.095	24.701
20	14.218	22.399	13.373	23.589	12.759	24.510	12.431	25.026	11.902	25.892
21	15.108	23.509	14.235	24.725	13.602	25.667	13.262	26.194	12.715	27.078
22	16.001	24.615	15.102	25.858	14.448	26.820	14.098	27.357	13.532	28.259
23	16.897	25.719	15.971	26.987	15.298	27.969	14.937	28.517	14.354	29.436
24	17.794	26.821	16.844	28.114	16.152	29.114	15.780	29.672	15.180	30.609
25	18.695	27.921	17.719	29.238	17.008	30.257	16.627	30.825	16.010	31.778
26	19.597	29.018	18.597	30.360	17.868	31.396	17.477	31.974	16.843	32.944
27	20.501	30.113	19.478	31.479	18.731	32.533	18.329	33.121	17.680	34.106
28	21.407	31.207	20.361	32.595	19.596	33.667	19.185	34.264	18.520	35.265
29	22.315	32.299	21.246	33.710	20.464	34.798	20.044	35.405	19.363	36.421
30	23.225	33.389	22.133	34.822	21.335	35.927	20.905	36.543	20.209	37.575
35	27.795	38.818	26.597	40.357	25.718	41.541	25.245	42.201	24.477	43.304
40	32.396	44.216	31.099	45.853	30.146	47.112	29.632	47.812	28.797	48.981
45	37.023	49.588	35.633	51.318	34.611	52.646	34.059	53.383	33.161	54.616
50	41.670	54.940	40.194	56.757	39.106	58.150	38.518	58.924	37.560	60.216
55	46.336	60.275	44.776	62.174	43.626	63.629	43.004	64.437	41.990	65.784
60	51.017	65.593	49.378	67.572	48.169	69.086	47.514	69.926	46.446	71.328

$v$	80%		90%		95%		99%		99.5%	
3	0.046	4.672	0.012	6.260	0.003	7.817	0.000	11.346	0.000	12.840
4	0.335	6.161	0.168	7.864	0.085	9.530	0.017	13.287	0.009	14.860
5	0.779	7.622	0.476	9.434	0.296	11.191	0.101	15.128	0.064	16.771
6	1.308	9.042	0.883	10.958	0.607	12.802	0.264	16.903	0.186	18.612
7	1.891	10.427	1.355	12.442	0.989	14.369	0.496	18.619	0.372	20.390
8	2.513	11.784	1.875	13.892	1.425	15.897	0.786	20.295	0.614	22.116
9	3.165	13.117	2.431	15.314	1.903	17.393	1.122	21.931	0.904	23.802
10	3.841	14.430	3.017	16.711	2.414	18.860	1.498	23.532	1.233	25.450
11	4.535	15.727	3.628	18.087	2.953	20.305	1.906	25.108	1.596	27.073
12	5.246	17.009	4.258	19.447	3.516	21.729	2.344	26.654	1.991	28.659
13	5.970	18.279	4.906	20.789	4.099	23.135	2.807	28.176	2.410	30.231
14	6.707	19.537	5.570	22.119	4.700	24.525	3.291	29.685	2.853	31.777
15	7.454	20.786	6.246	23.437	5.317	25.901	3.795	31.171	3.317	33.305
16	8.210	22.026	6.935	24.743	5.948	27.263	4.315	32.644	3.797	34.821
17	8.975	23.258	7.634	26.039	6.591	28.614	4.853	34.099	4.296	36.315
18	9.747	24.483	8.343	27.325	7.245	29.955	5.404	35.539	4.811	37.788
19	10.527	25.701	9.060	28.604	7.910	31.285	5.968	36.972	5.339	39.253
20	11.312	26.913	9.786	29.876	8.584	32.608	6.545	38.388	5.879	40.711
21	12.104	28.120	10.519	31.140	9.267	33.921	7.132	39.796	6.430	42.160
22	12.900	29.322	11.259	32.398	9.958	35.227	7.730	41.194	6.995	43.585
23	13.702	30.519	12.005	33.649	10.656	36.526	8.337	42.583	7.566	45.016
24	14.508	31.711	12.756	34.896	11.362	37.817	8.951	43.969	8.152	46.421
25	15.319	32.899	13.514	36.136	12.073	39.103	9.574	45.344	8.742	47.832
26	16.134	34.083	14.277	37.372	12.791	40.384	10.206	46.708	9.341	49.232
27	16.952	35.264	15.044	38.603	13.515	41.657	10.847	48.062	9.949	50.621
28	17.774	36.441	15.815	39.830	14.243	42.927	11.491	49.419	10.566	52.000
29	18.599	37.615	16.591	41.052	14.977	44.191	12.143	50.764	11.186	53.381
30	19.427	38.786	17.372	42.271	15.715	45.452	12.804	52.099	11.815	54.752
35	23.611	44.598	21.327	48.311	19.473	51.687	16.179	58.716	15.051	61.504
40	27.855	50.352	25.357	54.276	23.319	57.836	19.668	65.223	18.408	68.143
45	32.146	56.059	29.449	60.182	27.238	63.913	23.257	71.624	21.871	74.672
50	36.478	61.726	33.591	66.037	31.217	69.931	26.919	77.962	25.416	81.127
55	40.842	67.360	37.777	71.849	35.249	75.896	30.648	84.230	29.036	87.501
60	45.236	72.965	41.999	77.625	39.323	81.821	34.436	90.440	32.717	93.818

**Table B.3** HDRs for the inverse chi-squared distribution.

$v$	50%		60%		67%		70%		75%	
3	0.106	0.446	0.093	0.553	0.085	0.653	0.082	0.716	0.076	0.837
4	0.098	0.320	0.087	0.380	0.080	0.435	0.077	0.469	0.072	0.532
5	0.089	0.249	0.081	0.289	0.075	0.324	0.072	0.346	0.068	0.385
6	0.082	0.204	0.075	0.233	0.070	0.258	0.067	0.273	0.064	0.299
7	0.075	0.173	0.069	0.195	0.065	0.213	0.063	0.224	0.060	0.244
8	0.070	0.150	0.064	0.167	0.061	0.182	0.059	0.190	0.056	0.206
9	0.065	0.133	0.060	0.147	0.057	0.158	0.055	0.165	0.053	0.177
10	0.061	0.119	0.056	0.130	0.054	0.140	0.052	0.146	0.050	0.156
11	0.057	0.107	0.053	0.117	0.051	0.125	0.049	0.130	0.047	0.138
12	0.054	0.098	0.050	0.106	0.048	0.113	0.047	0.118	0.045	0.125
13	0.051	0.090	0.048	0.097	0.045	0.104	0.044	0.107	0.042	0.113
14	0.048	0.083	0.045	0.090	0.043	0.095	0.042	0.098	0.041	0.104
15	0.046	0.078	0.043	0.083	0.041	0.088	0.040	0.091	0.039	0.096
16	0.044	0.072	0.041	0.078	0.039	0.082	0.038	0.084	0.037	0.089
17	0.042	0.068	0.039	0.073	0.038	0.077	0.037	0.079	0.036	0.083
18	0.040	0.064	0.038	0.068	0.036	0.072	0.035	0.074	0.034	0.077
19	0.038	0.061	0.036	0.065	0.035	0.068	0.034	0.069	0.033	0.073
20	0.037	0.057	0.035	0.061	0.033	0.064	0.033	0.066	0.032	0.068
21	0.035	0.055	0.033	0.058	0.032	0.061	0.032	0.062	0.031	0.065
22	0.034	0.052	0.032	0.055	0.031	0.058	0.031	0.059	0.030	0.061
23	0.033	0.050	0.031	0.053	0.030	0.055	0.030	0.056	0.029	0.058
24	0.032	0.047	0.030	0.050	0.029	0.052	0.029	0.054	0.028	0.056
25	0.031	0.046	0.029	0.048	0.028	0.050	0.028	0.051	0.027	0.053
26	0.030	0.044	0.028	0.046	0.027	0.048	0.027	0.049	0.026	0.051
27	0.029	0.042	0.027	0.044	0.027	0.046	0.026	0.047	0.025	0.049
28	0.028	0.040	0.027	0.043	0.026	0.044	0.025	0.045	0.025	0.047
29	0.027	0.039	0.026	0.041	0.025	0.043	0.025	0.043	0.024	0.045
30	0.026	0.038	0.025	0.039	0.024	0.041	0.024	0.042	0.023	0.043
35	0.023	0.032	0.022	0.034	0.021	0.035	0.021	0.035	0.021	0.036
40	0.020	0.028	0.020	0.029	0.019	0.030	0.019	0.031	0.018	0.031
45	0.018	0.025	0.018	0.026	0.017	0.026	0.017	0.027	0.017	0.028
50	0.017	0.022	0.016	0.023	0.016	0.024	0.016	0.024	0.015	0.025
55	0.015	0.020	0.015	0.021	0.015	0.021	0.014	0.022	0.014	0.022
60	0.014	0.018	0.014	0.019	0.014	0.019	0.013	0.020	0.013	0.020

$v$	80%		90%		95%		99%		99.5%	
3	0.070	1.005	0.057	1.718	0.048	2.847	0.036	8.711	0.033	13.946
4	0.067	0.616	0.055	0.947	0.047	1.412	0.036	3.370	0.033	4.829
5	0.063	0.436	0.053	0.627	0.046	0.878	0.036	1.807	0.033	2.433
6	0.060	0.334	0.050	0.460	0.044	0.616	0.035	1.150	0.032	1.482
7	0.056	0.270	0.048	0.359	0.042	0.466	0.033	0.810	0.031	1.013
8	0.053	0.225	0.045	0.292	0.040	0.370	0.032	0.610	0.030	0.746
9	0.050	0.193	0.043	0.245	0.038	0.305	0.031	0.481	0.029	0.578
10	0.047	0.168	0.041	0.210	0.037	0.257	0.030	0.393	0.028	0.466
11	0.045	0.149	0.039	0.184	0.035	0.222	0.029	0.330	0.027	0.386
12	0.043	0.134	0.037	0.163	0.034	0.195	0.028	0.282	0.026	0.327
13	0.041	0.121	0.036	0.146	0.032	0.173	0.027	0.245	0.025	0.282
14	0.039	0.110	0.034	0.132	0.031	0.155	0.026	0.216	0.024	0.247
15	0.037	0.102	0.033	0.120	0.030	0.140	0.025	0.193	0.024	0.219
16	0.036	0.094	0.032	0.111	0.029	0.128	0.024	0.174	0.023	0.196
17	0.034	0.087	0.031	0.102	0.028	0.118	0.024	0.158	0.022	0.177
18	0.033	0.082	0.029	0.095	0.027	0.109	0.023	0.144	0.022	0.161
19	0.032	0.076	0.028	0.089	0.026	0.101	0.022	0.132	0.021	0.147
20	0.031	0.072	0.028	0.083	0.025	0.094	0.022	0.122	0.020	0.136
21	0.029	0.068	0.027	0.078	0.025	0.088	0.021	0.114	0.020	0.126
22	0.029	0.064	0.026	0.074	0.024	0.083	0.020	0.106	0.019	0.117
23	0.028	0.061	0.025	0.070	0.023	0.078	0.020	0.099	0.019	0.109
24	0.027	0.058	0.024	0.066	0.022	0.074	0.019	0.093	0.018	0.102
25	0.026	0.055	0.024	0.063	0.022	0.070	0.019	0.088	0.018	0.096
26	0.025	0.053	0.023	0.060	0.021	0.067	0.018	0.083	0.018	0.091
27	0.024	0.051	0.022	0.057	0.021	0.063	0.018	0.079	0.017	0.086
28	0.024	0.049	0.022	0.055	0.020	0.061	0.018	0.075	0.017	0.081
29	0.023	0.047	0.021	0.052	0.020	0.058	0.017	0.071	0.016	0.077
30	0.023	0.045	0.021	0.050	0.019	0.055	0.017	0.068	0.016	0.073
35	0.020	0.038	0.018	0.042	0.017	0.046	0.015	0.055	0.015	0.059
40	0.018	0.032	0.017	0.036	0.016	0.039	0.014	0.046	0.013	0.049
45	0.016	0.028	0.015	0.031	0.014	0.034	0.013	0.039	0.012	0.042
50	0.015	0.025	0.014	0.027	0.013	0.030	0.012	0.034	0.011	0.036
55	0.014	0.023	0.013	0.025	0.012	0.026	0.011	0.030	0.011	0.032
60	0.013	0.021	0.012	0.022	0.011	0.024	0.010	0.027	0.010	0.029

**Table B.4** Chi-squared corresponding to HDRs for log chi-squared.

$v$	50%		60%		67%		70%		75%	
3	1.576	5.097	1.315	5.730	1.141	6.238	1.052	6.527	0.918	7.023
4	2.327	6.330	2.004	7.016	1.783	7.563	1.669	7.874	1.493	8.404
5	3.107	7.540	2.729	8.276	2.467	8.860	2.331	9.190	2.118	9.753
6	3.907	8.732	3.480	9.515	3.181	10.133	3.025	10.482	2.779	11.075
7	4.724	9.911	4.251	10.737	3.917	11.387	3.742	11.754	3.465	12.376
8	5.552	11.079	5.037	11.946	4.671	12.626	4.479	13.009	4.174	13.658
9	6.391	12.238	5.836	13.143	5.440	13.851	5.231	14.250	4.899	14.925
10	7.238	13.388	6.645	14.330	6.221	15.066	5.997	15.480	5.639	16.179
11	8.093	14.532	7.464	15.508	7.013	16.271	6.774	16.699	6.392	17.422
12	8.954	15.669	8.290	16.679	7.814	17.467	7.560	17.909	7.155	18.654
13	9.821	16.801	9.124	17.844	8.622	18.656	8.355	19.111	7.928	19.878
14	10.692	17.929	9.963	19.003	9.438	19.838	9.158	20.306	8.709	21.094
15	11.568	19.052	10.809	20.156	10.260	21.014	9.968	21.494	9.497	22.303
16	12.448	20.171	11.659	21.304	11.088	22.184	10.783	22.677	10.293	23.505
17	13.331	21.287	12.514	22.448	11.921	23.349	11.605	23.854	11.095	24.701
18	14.218	22.399	13.373	23.589	12.759	24.510	12.431	25.026	11.902	25.892
19	15.108	23.509	14.235	24.725	13.602	25.667	13.262	26.194	12.715	27.078
20	16.001	24.615	15.102	25.858	14.448	26.820	14.098	27.357	13.532	28.259
21	16.897	25.719	15.971	26.987	15.298	27.969	14.937	28.517	14.354	29.436
22	17.794	26.821	16.844	28.114	16.152	29.114	15.780	29.672	15.180	30.609
23	18.695	27.921	17.719	29.238	17.008	30.257	16.627	30.825	16.010	31.778
24	19.597	29.018	18.597	30.360	17.868	31.396	17.477	31.974	16.843	32.944
25	20.501	30.113	19.478	31.479	18.731	32.533	18.329	33.121	17.680	34.106
26	21.407	31.207	20.361	32.595	19.596	33.667	19.185	34.264	18.520	35.265
27	22.315	32.299	21.246	33.710	20.464	34.798	20.044	35.405	19.363	36.421
28	23.225	33.389	22.133	34.822	21.335	35.927	20.905	36.543	20.209	37.575
29	24.136	34.478	23.022	35.933	22.207	37.054	21.769	37.679	21.058	38.725
30	25.048	35.565	23.913	37.042	23.082	38.179	22.635	38.812	21.909	39.873
35	29.632	40.980	28.393	42.560	27.485	43.774	26.995	44.450	26.199	45.581
40	34.244	46.368	32.909	48.042	31.929	49.329	31.399	50.044	30.538	51.240
45	38.879	51.731	37.455	53.496	36.406	54.851	35.839	55.603	34.917	56.860
50	43.534	57.076	42.024	58.926	40.911	60.345	40.309	61.132	39.329	62.447
55	48.206	62.404	46.615	64.335	45.440	65.815	44.805	66.635	43.769	68.005
60	52.893	67.717	51.223	69.726	49.991	71.263	49.323	72.117	48.234	73.539

<i>v</i>	80%		90%		95%		99%		99.5%	
3	0.779	7.622	0.476	9.434	0.296	11.191	0.101	15.128	0.064	16.771
4	1.308	9.042	0.883	10.958	0.607	12.802	0.264	16.903	0.186	18.612
5	1.891	10.427	1.355	12.442	0.989	14.369	0.496	18.619	0.372	20.390
6	2.513	11.784	1.875	13.892	1.425	15.897	0.786	20.295	0.614	22.116
7	3.165	13.117	2.431	15.314	1.903	17.393	1.122	21.931	0.904	23.802
8	3.841	14.430	3.017	16.711	2.414	18.860	1.498	23.532	1.233	25.450
9	4.535	15.727	3.628	18.087	2.953	20.305	1.906	25.108	1.596	27.073
10	5.246	17.009	4.258	19.447	3.516	21.729	2.344	26.654	1.991	28.659
11	5.970	18.279	4.906	20.789	4.099	23.135	2.807	28.176	2.410	30.231
12	6.707	19.537	5.570	22.119	4.700	24.525	3.291	29.685	2.853	31.777
13	7.454	20.786	6.246	23.437	5.317	25.901	3.795	31.171	3.317	33.305
14	8.210	22.026	6.935	24.743	5.948	27.263	4.315	32.644	3.797	34.821
15	8.975	23.258	7.634	26.039	6.591	28.614	4.853	34.099	4.296	36.315
16	9.747	24.483	8.343	27.325	7.245	29.955	5.404	35.539	4.811	37.788
17	10.527	25.701	9.060	28.604	7.910	31.285	5.968	36.972	5.339	39.253
18	11.312	26.913	9.786	29.876	8.584	32.608	6.545	38.388	5.879	40.711
19	12.104	28.120	10.519	31.140	9.267	33.921	7.132	39.796	6.430	42.160
20	12.900	29.322	11.259	32.398	9.958	35.227	7.730	41.194	6.995	43.585
21	13.702	30.519	12.005	33.649	10.656	36.526	8.337	42.583	7.566	45.016
22	14.508	31.711	12.756	34.896	11.362	37.817	8.951	43.969	8.152	46.421
23	15.319	32.899	13.514	36.136	12.073	39.103	9.574	45.344	8.742	47.832
24	16.134	34.083	14.277	37.372	12.791	40.384	10.206	46.708	9.341	49.232
25	16.952	35.264	15.044	38.603	13.515	41.657	10.847	48.062	9.949	50.621
26	17.774	36.441	15.815	39.830	14.243	42.927	11.491	49.419	10.566	52.000
27	18.599	37.615	16.591	41.052	14.977	44.191	12.143	50.764	11.186	53.381
28	19.427	38.786	17.372	42.271	15.715	45.452	12.804	52.099	11.815	54.752
29	20.259	39.953	18.156	43.486	16.459	46.706	13.467	53.436	12.451	56.111
30	21.093	41.119	18.944	44.696	17.206	47.958	14.138	54.761	13.091	57.473
35	25.303	46.906	22.931	50.705	21.002	54.156	17.563	61.330	16.384	64.165
40	29.566	52.640	26.987	56.645	24.879	60.275	21.094	67.792	19.782	70.766
45	33.874	58.330	31.100	62.530	28.823	66.326	24.711	74.172	23.277	77.269
50	38.220	63.983	35.260	68.366	32.824	72.324	28.401	80.480	26.857	83.681
55	42.597	69.605	39.461	74.164	36.873	78.272	32.158	86.717	30.499	90.039
60	47.001	75.200	43.698	79.926	40.965	84.178	35.966	92.908	34.207	96.324

**Table B.5** Values of F corresponding to HDRs for log F.

$(v_1 = 3)$													
$v_2$	50%		67%		75%		90%		95%		99%		
3	0.42	2.36	0.29	3.48	0.22	4.47	0.11	9.28	0.06	15.44	0.02	47.45	
$(v_1 = 4)$													
$v_2$	50%		67%		75%		90%		95%		99%		
3	0.46	2.24	0.33	3.26	0.26	4.16	0.14	8.48	0.09	14.00	0.04	42.61	
4	0.48	2.06	0.35	2.86	0.28	3.52	0.16	6.39	0.10	9.60	0.04	23.15	
$(v_1 = 5)$													
$v_2$	50%		67%		75%		90%		95%		99%		
3	0.48	2.17	0.35	3.13	0.29	3.98	0.17	8.02	0.12	13.17	0.05	39.74	
4	0.51	2.00	0.38	2.74	0.31	3.35	0.19	6.00	0.13	8.97	0.06	21.45	
5	0.53	1.89	0.40	2.52	0.33	3.02	0.20	5.05	0.14	7.15	0.07	14.94	
$(v_1 = 6)$													
$v_2$	50%		67%		75%		90%		95%		99%		
3	0.50	2.13	0.37	3.05	0.31	3.86	0.19	7.72	0.13	12.63	0.07	37.93	
4	0.53	1.95	0.40	2.65	0.34	3.24	0.21	5.75	0.15	8.56	0.08	20.34	
5	0.55	1.85	0.42	2.44	0.35	2.91	0.22	4.82	0.16	6.79	0.08	14.10	
6	0.56	1.78	0.43	2.30	0.37	2.71	0.23	4.28	0.17	5.82	0.09	11.07	
$(v_1 = 7)$													
$v_2$	50%		67%		75%		90%		95%		99%		
3	0.51	2.10	0.38	2.99	0.32	3.77	0.20	7.50	0.15	12.25	0.08	36.66	
4	0.54	1.92	0.42	2.60	0.35	3.16	0.22	5.57	0.17	8.27	0.09	19.55	
5	0.56	1.82	0.44	2.38	0.37	2.84	0.24	4.66	0.18	6.54	0.10	13.50	
6	0.58	1.75	0.45	2.25	0.39	2.64	0.25	4.13	0.19	5.59	0.11	10.57	
7	0.59	1.70	0.46	2.15	0.40	2.50	0.26	3.79	0.20	4.99	0.11	8.89	

$(v_1 = 8)$													
$v_2$	50%		67%		75%		90%		95%		99%		
3	0.52	2.07	0.39	2.94	0.33	3.71	0.21	7.35	0.16	11.97	0.09	35.69	
4	0.55	1.89	0.43	2.55	0.37	3.10	0.24	5.44	0.18	8.05	0.10	18.97	
5	0.57	1.79	0.45	2.34	0.39	2.78	0.26	4.54	0.20	6.35	0.11	13.05	
6	0.59	1.72	0.47	2.20	0.40	2.58	0.27	4.01	0.21	5.42	0.12	10.20	
7	0.60	1.68	0.48	2.11	0.42	2.44	0.28	3.68	0.22	4.83	0.13	8.55	
8	0.61	1.64	0.49	2.04	0.43	2.34	0.29	3.44	0.23	4.43	0.13	7.50	
$(v_1 = 9)$													
$v_2$	50%		67%		75%		90%		95%		99%		
3	0.53	2.05	0.40	2.91	0.34	3.66	0.22	7.22	0.17	11.75	0.10	34.99	
4	0.56	1.88	0.44	2.52	0.38	3.05	0.25	5.34	0.19	7.88	0.11	18.52	
5	0.58	1.77	0.46	2.30	0.40	2.73	0.27	4.44	0.21	6.20	0.13	12.71	
6	0.60	1.70	0.48	2.17	0.42	2.53	0.28	3.92	0.22	5.28	0.13	9.91	
7	0.61	1.66	0.49	2.07	0.43	2.40	0.30	3.59	0.23	4.70	0.14	8.29	
8	0.62	1.62	0.50	2.01	0.44	2.30	0.31	3.35	0.24	4.31	0.15	7.26	
9	0.63	1.59	0.51	1.95	0.45	2.22	0.31	3.18	0.25	4.03	0.15	6.54	
$(v_1 = 10)$													
$v_2$	50%		67%		75%		90%		95%		99%		
3	0.53	2.04	0.41	2.88	0.35	3.62	0.23	7.13	0.18	11.58	0.11	34.41	
4	0.57	1.86	0.45	2.49	0.38	3.01	0.26	5.25	0.20	7.75	0.12	18.17	
5	0.59	1.76	0.47	2.28	0.41	2.69	0.28	4.37	0.22	6.08	0.14	12.44	
6	0.61	1.69	0.49	2.14	0.43	2.50	0.30	3.85	0.23	5.17	0.15	9.67	
7	0.62	1.64	0.50	2.05	0.44	2.36	0.31	3.52	0.25	4.60	0.15	8.09	
8	0.63	1.60	0.51	1.98	0.45	2.26	0.32	3.28	0.25	4.21	0.16	7.07	
9	0.64	1.57	0.52	1.92	0.46	2.19	0.33	3.11	0.26	3.93	0.17	6.36	
10	0.64	1.55	0.53	1.88	0.47	2.13	0.34	2.98	0.27	3.72	0.17	5.85	
$(v_1 = 11)$													
$v_2$	50%		67%		75%		90%		95%		99%		
3	0.54	2.03	0.42	2.85	0.36	3.58	0.24	7.05	0.18	11.44	0.11	33.97	
4	0.57	1.85	0.45	2.47	0.39	2.98	0.27	5.19	0.21	7.64	0.13	17.89	
5	0.59	1.74	0.48	2.25	0.42	2.66	0.29	4.30	0.23	5.99	0.14	12.22	
6	0.61	1.67	0.50	2.12	0.44	2.47	0.31	3.79	0.24	5.09	0.16	9.48	
7	0.63	1.63	0.51	2.02	0.45	2.33	0.32	3.46	0.26	4.52	0.16	7.92	
8	0.64	1.59	0.52	1.95	0.46	2.23	0.33	3.23	0.27	4.13	0.17	6.91	
9	0.65	1.56	0.53	1.90	0.47	2.16	0.34	3.06	0.27	3.85	0.18	6.21	
10	0.65	1.54	0.54	1.86	0.48	2.10	0.35	2.92	0.28	3.64	0.18	5.70	
11	0.66	1.52	0.55	1.82	0.49	2.05	0.35	2.82	0.29	3.47	0.19	5.32	

(v <sub>1</sub> = 12)													
v <sub>2</sub>	50%		67%		75%		90%		95%		99%		
3	0.54	2.02	0.42	2.84	0.36	3.56	0.24	6.99	0.19	11.33	0.12	33.58	
4	0.58	1.84	0.46	2.45	0.40	2.96	0.27	5.13	0.22	7.55	0.14	17.64	
5	0.60	1.73	0.48	2.24	0.42	2.64	0.30	4.25	0.24	5.91	0.15	12.03	
6	0.62	1.66	0.50	2.10	0.44	2.44	0.32	3.74	0.25	5.01	0.16	9.33	
7	0.63	1.61	0.52	2.01	0.46	2.30	0.33	3.41	0.27	4.45	0.17	7.77	
8	0.64	1.58	0.53	1.94	0.47	2.21	0.34	3.18	0.28	4.07	0.18	6.78	
9	0.65	1.55	0.54	1.88	0.48	2.13	0.35	3.01	0.29	3.79	0.19	6.09	
10	0.66	1.53	0.55	1.84	0.49	2.07	0.36	2.88	0.29	3.58	0.19	5.59	
11	0.67	1.51	0.56	1.80	0.50	2.02	0.37	2.77	0.30	3.41	0.20	5.20	
12	0.67	1.49	0.56	1.78	0.50	1.98	0.37	2.69	0.31	3.28	0.20	4.91	
(v <sub>1</sub> = 13)													
v <sub>2</sub>	50%		67%		75%		90%		95%		99%		
3	0.54	2.01	0.43	2.82	0.37	3.54	0.25	6.93	0.20	11.23	0.12	33.26	
4	0.58	1.83	0.46	2.43	0.40	2.94	0.28	5.08	0.22	7.48	0.14	17.45	
5	0.60	1.72	0.49	2.22	0.43	2.62	0.30	4.21	0.24	5.84	0.16	11.88	
6	0.62	1.65	0.51	2.08	0.45	2.42	0.32	3.70	0.26	4.95	0.17	9.20	
7	0.64	1.60	0.53	1.99	0.47	2.28	0.34	3.37	0.27	4.39	0.18	7.66	
8	0.65	1.57	0.54	1.92	0.48	2.18	0.35	3.14	0.29	4.01	0.19	6.67	
9	0.66	1.54	0.55	1.87	0.49	2.11	0.36	2.97	0.29	3.73	0.20	5.99	
10	0.67	1.51	0.56	1.82	0.50	2.05	0.37	2.84	0.30	3.52	0.20	5.49	
11	0.67	1.50	0.56	1.79	0.51	2.00	0.38	2.73	0.31	3.36	0.21	5.11	
12	0.68	1.48	0.57	1.76	0.51	1.96	0.38	2.65	0.32	3.22	0.21	4.81	
13	0.68	1.47	0.58	1.73	0.52	1.93	0.39	2.58	0.32	3.12	0.22	4.57	
(v <sub>1</sub> = 14)													
v <sub>2</sub>	50%		67%		75%		90%		95%		99%		
3	0.55	2.00	0.43	2.81	0.37	3.52	0.25	6.88	0.20	11.15	0.13	32.99	
4	0.58	1.82	0.47	2.42	0.41	2.92	0.29	5.04	0.23	7.41	0.15	17.28	
5	0.61	1.71	0.49	2.21	0.44	2.60	0.31	4.17	0.25	5.79	0.16	11.75	
6	0.63	1.65	0.51	2.07	0.46	2.40	0.33	3.67	0.27	4.90	0.18	9.09	
7	0.64	1.60	0.53	1.98	0.47	2.27	0.34	3.34	0.28	4.34	0.19	7.55	
8	0.65	1.56	0.54	1.91	0.49	2.17	0.36	3.11	0.29	3.96	0.20	6.57	
9	0.66	1.53	0.55	1.85	0.50	2.09	0.37	2.94	0.30	3.68	0.21	5.90	
10	0.67	1.51	0.56	1.81	0.51	2.03	0.38	2.80	0.31	3.47	0.21	5.40	
11	0.68	1.49	0.57	1.77	0.51	1.98	0.38	2.70	0.32	3.31	0.22	5.03	
12	0.68	1.47	0.58	1.74	0.52	1.94	0.39	2.61	0.33	3.18	0.22	4.73	
13	0.69	1.46	0.58	1.72	0.53	1.91	0.40	2.54	0.33	3.07	0.23	4.49	
14	0.69	1.44	0.59	1.70	0.53	1.88	0.40	2.48	0.34	2.98	0.23	4.30	

$(\nu_1 = 15)$													
$\nu_2$	50%		67%		75%		90%		95%		99%		
3	0.55	1.99	0.43	2.79	0.37	3.50	0.26	6.84	0.20	11.08	0.13	32.77	
4	0.59	1.81	0.47	2.41	0.41	2.90	0.29	5.01	0.23	7.35	0.15	17.12	
5	0.61	1.71	0.50	2.19	0.44	2.58	0.32	4.14	0.26	5.74	0.17	11.63	
6	0.63	1.64	0.52	2.06	0.46	2.39	0.34	3.64	0.27	4.85	0.18	8.98	
7	0.65	1.59	0.54	1.96	0.48	2.25	0.35	3.31	0.29	4.30	0.20	7.47	
8	0.66	1.55	0.55	1.89	0.49	2.15	0.36	3.08	0.30	3.92	0.21	6.49	
9	0.67	1.52	0.56	1.84	0.50	2.07	0.37	2.91	0.31	3.64	0.21	5.82	
10	0.67	1.50	0.57	1.80	0.51	2.01	0.38	2.78	0.32	3.43	0.22	5.33	
11	0.68	1.48	0.58	1.76	0.52	1.97	0.39	2.67	0.33	3.27	0.23	4.96	
12	0.69	1.46	0.58	1.73	0.53	1.93	0.40	2.59	0.33	3.14	0.23	4.66	
13	0.69	1.45	0.59	1.71	0.53	1.89	0.41	2.51	0.34	3.03	0.24	4.43	
14	0.70	1.44	0.60	1.68	0.54	1.86	0.41	2.45	0.34	2.94	0.24	4.23	
15	0.70	1.43	0.60	1.67	0.54	1.84	0.42	2.40	0.35	2.86	0.25	4.07	
$(\nu_1 = 16)$													
$\nu_2$	50%		67%		75%		90%		95%		99%		
3	0.55	1.99	0.44	2.78	0.38	3.48	0.26	6.81	0.21	11.02	0.13	32.54	
4	0.59	1.81	0.47	2.40	0.42	2.89	0.29	4.98	0.24	7.31	0.16	17.01	
5	0.61	1.70	0.50	2.18	0.44	2.57	0.32	4.11	0.26	5.69	0.18	11.53	
6	0.63	1.63	0.52	2.05	0.47	2.37	0.34	3.61	0.28	4.81	0.19	8.90	
7	0.65	1.58	0.54	1.95	0.48	2.24	0.36	3.28	0.29	4.26	0.20	7.39	
8	0.66	1.54	0.55	1.88	0.50	2.14	0.37	3.05	0.31	3.88	0.21	6.42	
9	0.67	1.52	0.56	1.83	0.51	2.06	0.38	2.88	0.32	3.61	0.22	5.75	
10	0.68	1.49	0.57	1.78	0.52	2.00	0.39	2.75	0.33	3.40	0.23	5.26	
11	0.69	1.47	0.58	1.75	0.53	1.95	0.40	2.65	0.33	3.23	0.23	4.89	
12	0.69	1.46	0.59	1.72	0.53	1.91	0.41	2.56	0.34	3.10	0.24	4.60	
13	0.70	1.44	0.60	1.69	0.54	1.88	0.41	2.49	0.35	3.00	0.25	4.37	
14	0.70	1.43	0.60	1.67	0.55	1.85	0.42	2.43	0.35	2.90	0.25	4.17	
15	0.71	1.42	0.61	1.65	0.55	1.82	0.42	2.38	0.36	2.83	0.25	4.01	
16	0.71	1.41	0.61	1.64	0.56	1.80	0.43	2.33	0.36	2.76	0.26	3.87	

$(\nu_1 = 17)$													
$\nu_2$	50%		67%		75%		90%		95%		99%		
3	0.55	1.98	0.44	2.78	0.38	3.47	0.26	6.78	0.21	10.96	0.14	32.38	
4	0.59	1.80	0.48	2.39	0.42	2.88	0.30	4.95	0.24	7.26	0.16	16.89	
5	0.62	1.70	0.51	2.18	0.45	2.56	0.32	4.09	0.27	5.66	0.18	11.44	
6	0.64	1.63	0.53	2.04	0.47	2.36	0.34	3.59	0.28	4.78	0.19	8.83	
7	0.65	1.58	0.54	1.94	0.49	2.22	0.36	3.26	0.30	4.22	0.21	7.32	
8	0.66	1.54	0.56	1.87	0.50	2.12	0.38	3.03	0.31	3.85	0.22	6.36	
9	0.67	1.51	0.57	1.82	0.51	2.05	0.39	2.86	0.32	3.57	0.23	5.69	
10	0.68	1.49	0.58	1.77	0.52	1.99	0.40	2.73	0.33	3.37	0.23	5.20	
11	0.69	1.47	0.59	1.74	0.53	1.94	0.41	2.62	0.34	3.20	0.24	4.83	
12	0.70	1.45	0.59	1.71	0.54	1.90	0.41	2.54	0.35	3.07	0.25	4.55	
13	0.70	1.44	0.60	1.68	0.55	1.86	0.42	2.47	0.35	2.96	0.25	4.31	
14	0.71	1.42	0.61	1.66	0.55	1.83	0.43	2.41	0.36	2.87	0.26	4.12	
15	0.71	1.41	0.61	1.64	0.56	1.81	0.43	2.36	0.37	2.80	0.26	3.96	
16	0.71	1.40	0.62	1.63	0.56	1.79	0.44	2.31	0.37	2.73	0.27	3.82	
17	0.72	1.39	0.62	1.61	0.57	1.77	0.44	2.27	0.37	2.67	0.27	3.71	
$(\nu_1 = 18)$													
$\nu_2$	50%		67%		75%		90%		95%		99%		
3	0.56	1.98	0.44	2.77	0.38	3.46	0.27	6.75	0.21	10.92	0.14	32.23	
4	0.59	1.80	0.48	2.38	0.42	2.87	0.30	4.93	0.24	7.22	0.16	16.78	
5	0.62	1.69	0.51	2.17	0.45	2.55	0.33	4.07	0.27	5.62	0.18	11.37	
6	0.64	1.62	0.53	2.03	0.47	2.35	0.35	3.56	0.29	4.75	0.20	8.76	
7	0.65	1.57	0.55	1.93	0.49	2.21	0.37	3.24	0.30	4.19	0.21	7.26	
8	0.67	1.53	0.56	1.86	0.51	2.11	0.38	3.01	0.32	3.82	0.22	6.30	
9	0.68	1.50	0.57	1.81	0.52	2.04	0.39	2.84	0.33	3.55	0.23	5.64	
10	0.69	1.48	0.58	1.77	0.53	1.98	0.40	2.71	0.34	3.34	0.24	5.15	
11	0.69	1.46	0.59	1.73	0.54	1.93	0.41	2.60	0.35	3.18	0.25	4.79	
12	0.70	1.44	0.60	1.70	0.54	1.89	0.42	2.52	0.35	3.04	0.25	4.50	
13	0.70	1.43	0.61	1.68	0.55	1.85	0.43	2.45	0.36	2.94	0.26	4.26	
14	0.71	1.42	0.61	1.65	0.56	1.82	0.43	2.39	0.37	2.85	0.26	4.07	
15	0.71	1.41	0.62	1.63	0.56	1.80	0.44	2.33	0.37	2.77	0.27	3.91	
16	0.72	1.40	0.62	1.62	0.57	1.78	0.44	2.29	0.38	2.70	0.27	3.78	
17	0.72	1.39	0.62	1.60	0.57	1.76	0.45	2.25	0.38	2.65	0.28	3.66	
18	0.72	1.38	0.63	1.59	0.58	1.74	0.45	2.22	0.39	2.60	0.28	3.56	

( $v_1 = 19$ )

$v_2$	50%		67%		75%		90%		95%		99%	
3	0.56	1.97	0.44	2.76	0.38	3.45	0.27	6.73	0.22	10.87	0.14	32.09
4	0.60	1.79	0.48	2.37	0.42	2.86	0.30	4.91	0.25	7.19	0.17	16.70
5	0.62	1.69	0.51	2.16	0.45	2.54	0.33	4.05	0.27	5.59	0.19	11.30
6	0.64	1.62	0.53	2.02	0.48	2.34	0.35	3.55	0.29	4.72	0.20	8.70
7	0.66	1.57	0.55	1.93	0.49	2.20	0.37	3.22	0.31	4.17	0.22	7.21
8	0.67	1.53	0.57	1.86	0.51	2.10	0.38	2.99	0.32	3.79	0.23	6.25
9	0.68	1.50	0.58	1.80	0.52	2.03	0.40	2.82	0.33	3.52	0.24	5.59
10	0.69	1.48	0.59	1.76	0.53	1.97	0.41	2.69	0.34	3.31	0.25	5.11
11	0.70	1.46	0.60	1.72	0.54	1.92	0.42	2.58	0.35	3.15	0.25	4.74
12	0.70	1.44	0.60	1.69	0.55	1.88	0.42	2.50	0.36	3.02	0.26	4.45
13	0.71	1.42	0.61	1.67	0.56	1.84	0.43	2.43	0.37	2.91	0.27	4.22
14	0.71	1.41	0.62	1.65	0.56	1.81	0.44	2.37	0.37	2.82	0.27	4.03
15	0.72	1.40	0.62	1.63	0.57	1.79	0.44	2.32	0.38	2.74	0.28	3.87
16	0.72	1.39	0.63	1.61	0.57	1.77	0.45	2.27	0.38	2.68	0.28	3.74
17	0.72	1.38	0.63	1.60	0.58	1.75	0.45	2.23	0.39	2.62	0.28	3.62
18	0.73	1.38	0.63	1.58	0.58	1.73	0.46	2.20	0.39	2.57	0.29	3.52
19	0.73	1.37	0.64	1.57	0.58	1.71	0.46	2.17	0.40	2.53	0.29	3.43

( $v_1 = 20$ )

$v_2$	50%		67%		75%		90%		95%		99%	
3	0.56	1.97	0.44	2.75	0.39	3.44	0.27	6.70	0.22	10.83	0.15	31.95
4	0.60	1.79	0.48	2.37	0.43	2.85	0.31	4.89	0.25	7.16	0.17	16.63
5	0.62	1.68	0.51	2.15	0.46	2.53	0.33	4.03	0.28	5.57	0.19	11.24
6	0.64	1.61	0.54	2.02	0.48	2.33	0.36	3.53	0.30	4.69	0.21	8.64
7	0.66	1.56	0.55	1.92	0.50	2.19	0.37	3.20	0.31	4.14	0.22	7.16
8	0.67	1.52	0.57	1.85	0.51	2.09	0.39	2.97	0.33	3.77	0.23	6.20
9	0.68	1.49	0.58	1.79	0.52	2.02	0.40	2.80	0.34	3.50	0.24	5.55
10	0.69	1.47	0.59	1.75	0.54	1.96	0.41	2.67	0.35	3.29	0.25	5.07
11	0.70	1.45	0.60	1.71	0.54	1.91	0.42	2.57	0.36	3.13	0.26	4.70
12	0.70	1.43	0.61	1.68	0.55	1.87	0.43	2.48	0.37	3.00	0.27	4.41
13	0.71	1.42	0.61	1.66	0.56	1.83	0.44	2.41	0.37	2.89	0.27	4.18
14	0.72	1.41	0.62	1.64	0.57	1.80	0.44	2.35	0.38	2.80	0.28	3.99
15	0.72	1.40	0.62	1.62	0.57	1.78	0.45	2.30	0.38	2.72	0.28	3.83
16	0.72	1.39	0.63	1.60	0.58	1.76	0.45	2.26	0.39	2.66	0.29	3.70
17	0.73	1.38	0.63	1.59	0.58	1.74	0.46	2.22	0.39	2.60	0.29	3.58
18	0.73	1.37	0.64	1.57	0.58	1.72	0.46	2.18	0.40	2.55	0.29	3.48
19	0.73	1.36	0.64	1.56	0.59	1.70	0.47	2.15	0.40	2.50	0.30	3.40
20	0.74	1.36	0.64	1.55	0.59	1.69	0.47	2.12	0.41	2.46	0.30	3.32

$(v_1 = 21)$ 

$v_2$	50%		67%		75%		90%		95%		99%	
3	0.56	1.97	0.45	2.75	0.39	3.43	0.27	6.68	0.22	10.80	0.15	31.85
4	0.60	1.79	0.49	2.36	0.43	2.84	0.31	4.87	0.25	7.13	0.17	16.55
5	0.63	1.68	0.52	2.15	0.46	2.52	0.34	4.01	0.28	5.54	0.19	11.18
6	0.65	1.61	0.54	2.01	0.48	2.32	0.36	3.51	0.30	4.67	0.21	8.59
7	0.66	1.56	0.56	1.91	0.50	2.19	0.38	3.19	0.32	4.12	0.22	7.11
8	0.67	1.52	0.57	1.84	0.52	2.09	0.39	2.96	0.33	3.75	0.24	6.16
9	0.68	1.49	0.58	1.79	0.53	2.01	0.41	2.79	0.34	3.48	0.25	5.51
10	0.69	1.47	0.59	1.74	0.54	1.95	0.42	2.66	0.35	3.27	0.26	5.03
11	0.70	1.45	0.60	1.71	0.55	1.90	0.43	2.55	0.36	3.11	0.26	4.66
12	0.71	1.43	0.61	1.68	0.56	1.86	0.43	2.47	0.37	2.98	0.27	4.38
13	0.71	1.42	0.62	1.65	0.56	1.82	0.44	2.40	0.38	2.87	0.28	4.15
14	0.72	1.40	0.62	1.63	0.57	1.79	0.45	2.34	0.38	2.78	0.28	3.96
15	0.72	1.39	0.63	1.61	0.58	1.77	0.45	2.28	0.39	2.70	0.29	3.80
16	0.73	1.38	0.63	1.60	0.58	1.75	0.46	2.24	0.39	2.64	0.29	3.67
17	0.73	1.37	0.64	1.58	0.59	1.73	0.46	2.20	0.40	2.58	0.30	3.55
18	0.73	1.37	0.64	1.57	0.59	1.71	0.47	2.17	0.40	2.53	0.30	3.45
19	0.74	1.36	0.64	1.56	0.59	1.69	0.47	2.14	0.41	2.48	0.30	3.36
20	0.74	1.35	0.65	1.55	0.60	1.68	0.48	2.11	0.41	2.44	0.31	3.29
21	0.74	1.35	0.65	1.54	0.60	1.67	0.48	2.08	0.42	2.41	0.31	3.22

 $(v_1 = 22)$ 

$v_2$	50%		67%		75%		90%		95%		99%	
3	0.56	1.96	0.45	2.74	0.39	3.43	0.27	6.67	0.22	10.77	0.15	31.75
4	0.60	1.78	0.49	2.36	0.43	2.83	0.31	4.86	0.26	7.11	0.18	16.48
5	0.63	1.68	0.52	2.14	0.46	2.51	0.34	4.00	0.28	5.52	0.20	11.12
6	0.65	1.61	0.54	2.00	0.48	2.31	0.36	3.50	0.30	4.65	0.21	8.55
7	0.66	1.56	0.56	1.91	0.50	2.18	0.38	3.17	0.32	4.10	0.23	7.07
8	0.68	1.52	0.57	1.84	0.52	2.08	0.40	2.94	0.33	3.73	0.24	6.13
9	0.69	1.49	0.59	1.78	0.53	2.00	0.41	2.77	0.35	3.46	0.25	5.47
10	0.70	1.46	0.60	1.74	0.54	1.94	0.42	2.64	0.36	3.25	0.26	4.99
11	0.70	1.44	0.61	1.70	0.55	1.89	0.43	2.54	0.37	3.09	0.27	4.63
12	0.71	1.43	0.61	1.67	0.56	1.85	0.44	2.45	0.37	2.96	0.28	4.35
13	0.72	1.41	0.62	1.65	0.57	1.82	0.45	2.38	0.38	2.85	0.28	4.12
14	0.72	1.40	0.63	1.62	0.57	1.79	0.45	2.32	0.39	2.76	0.29	3.93
15	0.73	1.39	0.63	1.61	0.58	1.76	0.46	2.27	0.39	2.68	0.29	3.77
16	0.73	1.38	0.64	1.59	0.58	1.74	0.46	2.23	0.40	2.62	0.30	3.64
17	0.73	1.37	0.64	1.57	0.59	1.72	0.47	2.19	0.40	2.56	0.30	3.52
18	0.74	1.36	0.64	1.56	0.59	1.70	0.47	2.15	0.41	2.51	0.31	3.42
19	0.74	1.36	0.65	1.55	0.60	1.69	0.48	2.12	0.41	2.47	0.31	3.33
20	0.74	1.35	0.65	1.54	0.60	1.67	0.48	2.09	0.42	2.43	0.31	3.26
21	0.74	1.34	0.65	1.53	0.60	1.66	0.49	2.07	0.42	2.39	0.32	3.19
22	0.75	1.34	0.66	1.52	0.61	1.65	0.49	2.05	0.42	2.36	0.32	3.12

$(\nu_1 = 23)$													
$\nu_2$	50%		67%		75%		90%		95%		99%		
3	0.56	1.96	0.45	2.74	0.39	3.42	0.28	6.65	0.22	10.74	0.15	31.66	
4	0.60	1.78	0.49	2.35	0.43	2.82	0.31	4.84	0.26	7.08	0.18	16.42	
5	0.63	1.67	0.52	2.14	0.46	2.51	0.34	3.98	0.28	5.50	0.20	11.09	
6	0.65	1.60	0.54	2.00	0.49	2.31	0.37	3.49	0.31	4.63	0.22	8.51	
7	0.66	1.55	0.56	1.90	0.51	2.17	0.38	3.16	0.32	4.08	0.23	7.04	
8	0.68	1.51	0.58	1.83	0.52	2.07	0.40	2.93	0.34	3.71	0.24	6.09	
9	0.69	1.48	0.59	1.78	0.53	1.99	0.41	2.76	0.35	3.44	0.25	5.44	
10	0.70	1.46	0.60	1.73	0.54	1.93	0.42	2.63	0.36	3.23	0.26	4.96	
11	0.71	1.44	0.61	1.70	0.55	1.88	0.43	2.53	0.37	3.07	0.27	4.60	
12	0.71	1.42	0.62	1.67	0.56	1.84	0.44	2.44	0.38	2.94	0.28	4.32	
13	0.72	1.41	0.62	1.64	0.57	1.81	0.45	2.37	0.39	2.83	0.29	4.09	
14	0.72	1.40	0.63	1.62	0.58	1.78	0.46	2.31	0.39	2.74	0.29	3.90	
15	0.73	1.38	0.63	1.60	0.58	1.75	0.46	2.26	0.40	2.67	0.30	3.74	
16	0.73	1.37	0.64	1.58	0.59	1.73	0.47	2.21	0.40	2.60	0.30	3.61	
17	0.74	1.37	0.64	1.57	0.59	1.71	0.47	2.17	0.41	2.54	0.31	3.49	
18	0.74	1.36	0.65	1.56	0.60	1.69	0.48	2.14	0.41	2.49	0.31	3.39	
19	0.74	1.35	0.65	1.54	0.60	1.68	0.48	2.11	0.42	2.45	0.32	3.30	
20	0.74	1.35	0.66	1.53	0.60	1.66	0.49	2.08	0.42	2.41	0.32	3.23	
21	0.75	1.34	0.66	1.52	0.61	1.65	0.49	2.06	0.43	2.37	0.32	3.16	
22	0.75	1.33	0.66	1.51	0.61	1.64	0.49	2.03	0.43	2.34	0.33	3.10	
23	0.75	1.33	0.66	1.51	0.61	1.63	0.50	2.01	0.43	2.31	0.33	3.04	

$(\nu_1 = 24)$														
$\nu_2$	50%		67%		75%		90%		95%		99%			
3	0.56	1.96	0.45	2.73	0.39	3.41	0.28	6.64	0.23	10.71	0.15	31.55		
4	0.60	1.78	0.49	2.35	0.43	2.82	0.32	4.83	0.26	7.06	0.18	16.36		
5	0.63	1.67	0.52	2.13	0.47	2.50	0.35	3.97	0.29	5.48	0.20	11.04		
6	0.65	1.60	0.54	1.99	0.49	2.30	0.37	3.47	0.31	4.61	0.22	8.48		
7	0.67	1.55	0.56	1.90	0.51	2.17	0.39	3.15	0.33	4.07	0.23	7.01		
8	0.68	1.51	0.58	1.83	0.52	2.06	0.40	2.92	0.34	3.69	0.25	6.06		
9	0.69	1.48	0.59	1.77	0.54	1.99	0.42	2.75	0.35	3.42	0.26	5.41		
10	0.70	1.46	0.60	1.73	0.55	1.93	0.43	2.62	0.36	3.22	0.27	4.94		
11	0.71	1.44	0.61	1.69	0.56	1.88	0.44	2.52	0.37	3.06	0.28	4.57		
12	0.71	1.42	0.62	1.66	0.57	1.84	0.45	2.43	0.38	2.93	0.28	4.29		
13	0.72	1.40	0.63	1.64	0.57	1.80	0.45	2.36	0.39	2.82	0.29	4.06		
14	0.73	1.39	0.63	1.61	0.58	1.77	0.46	2.30	0.40	2.73	0.30	3.87		
15	0.73	1.38	0.64	1.59	0.59	1.75	0.47	2.25	0.40	2.65	0.30	3.71		
16	0.73	1.37	0.64	1.58	0.59	1.72	0.47	2.20	0.41	2.59	0.31	3.58		
17	0.74	1.36	0.65	1.56	0.60	1.70	0.48	2.16	0.41	2.53	0.31	3.47		
18	0.74	1.36	0.65	1.55	0.60	1.69	0.48	2.13	0.42	2.48	0.32	3.37		
19	0.74	1.35	0.65	1.54	0.60	1.67	0.49	2.10	0.42	2.43	0.32	3.28		
20	0.75	1.34	0.66	1.53	0.61	1.66	0.49	2.07	0.43	2.39	0.32	3.20		
21	0.75	1.34	0.66	1.52	0.61	1.64	0.49	2.05	0.43	2.36	0.33	3.13		
22	0.75	1.33	0.66	1.51	0.62	1.63	0.50	2.02	0.43	2.33	0.33	3.07		
23	0.75	1.33	0.67	1.50	0.62	1.62	0.50	2.00	0.44	2.30	0.33	3.02		
24	0.76	1.32	0.67	1.49	0.62	1.61	0.50	1.98	0.44	2.27	0.34	2.97		

$(\nu_1 = 25)$													
$\nu_2$	50%		67%		75%		90%		95%		99%		
3	0.56	1.96	0.45	2.73	0.39	3.41	0.28	6.62	0.23	10.69	0.16	31.48	
4	0.60	1.78	0.49	2.34	0.44	2.81	0.32	4.82	0.26	7.04	0.18	16.32	
5	0.63	1.67	0.52	2.13	0.47	2.50	0.35	3.96	0.29	5.46	0.20	11.00	
6	0.65	1.60	0.55	1.99	0.49	2.30	0.37	3.46	0.31	4.60	0.22	8.44	
7	0.67	1.55	0.57	1.89	0.51	2.16	0.39	3.14	0.33	4.05	0.24	6.97	
8	0.68	1.51	0.58	1.82	0.53	2.06	0.40	2.91	0.34	3.68	0.25	6.03	
9	0.69	1.48	0.59	1.77	0.54	1.98	0.42	2.74	0.36	3.41	0.26	5.38	
10	0.70	1.45	0.60	1.72	0.55	1.92	0.43	2.61	0.37	3.20	0.27	4.91	
11	0.71	1.43	0.61	1.69	0.56	1.87	0.44	2.50	0.38	3.04	0.28	4.55	
12	0.72	1.42	0.62	1.66	0.57	1.83	0.45	2.42	0.39	2.91	0.29	4.26	
13	0.72	1.40	0.63	1.63	0.58	1.80	0.46	2.35	0.39	2.80	0.30	4.03	
14	0.73	1.39	0.63	1.61	0.58	1.77	0.46	2.29	0.40	2.71	0.30	3.85	
15	0.73	1.38	0.64	1.59	0.59	1.74	0.47	2.24	0.41	2.64	0.31	3.69	
16	0.74	1.37	0.65	1.57	0.59	1.72	0.48	2.19	0.41	2.57	0.31	3.56	
17	0.74	1.36	0.65	1.56	0.60	1.70	0.48	2.15	0.42	2.51	0.32	3.44	
18	0.74	1.35	0.65	1.54	0.60	1.68	0.49	2.12	0.42	2.46	0.32	3.34	
19	0.75	1.34	0.66	1.53	0.61	1.66	0.49	2.09	0.43	2.42	0.33	3.26	
20	0.75	1.34	0.66	1.52	0.61	1.65	0.49	2.06	0.43	2.38	0.33	3.18	
21	0.75	1.33	0.66	1.51	0.62	1.64	0.50	2.03	0.44	2.34	0.33	3.11	
22	0.75	1.33	0.67	1.50	0.62	1.62	0.50	2.01	0.44	2.31	0.34	3.05	
23	0.76	1.32	0.67	1.49	0.62	1.61	0.51	1.99	0.44	2.28	0.34	2.99	
24	0.76	1.32	0.67	1.49	0.62	1.60	0.51	1.97	0.45	2.25	0.34	2.94	
25	0.76	1.31	0.68	1.48	0.63	1.59	0.51	1.96	0.45	2.23	0.35	2.90	

$(\nu_1 = 26)$													
$\nu_2$	50%		67%		75%		90%		95%		99%		
3	0.57	1.96	0.45	2.72	0.40	3.40	0.28	6.61	0.23	10.67	0.16	31.39	
4	0.60	1.77	0.49	2.34	0.44	2.81	0.32	4.81	0.26	7.03	0.18	16.27	
5	0.63	1.67	0.52	2.12	0.47	2.49	0.35	3.95	0.29	5.45	0.21	10.96	
6	0.65	1.60	0.55	1.99	0.49	2.29	0.37	3.45	0.31	4.58	0.22	8.41	
7	0.67	1.54	0.57	1.89	0.51	2.15	0.39	3.13	0.33	4.04	0.24	6.94	
8	0.68	1.51	0.58	1.82	0.53	2.05	0.41	2.90	0.35	3.67	0.25	6.01	
9	0.69	1.48	0.59	1.76	0.54	1.98	0.42	2.73	0.36	3.39	0.26	5.36	
10	0.70	1.45	0.61	1.72	0.55	1.92	0.43	2.60	0.37	3.19	0.27	4.88	
11	0.71	1.43	0.61	1.68	0.56	1.87	0.44	2.49	0.38	3.03	0.28	4.52	
12	0.72	1.41	0.62	1.65	0.57	1.82	0.45	2.41	0.39	2.90	0.29	4.24	
13	0.72	1.40	0.63	1.63	0.58	1.79	0.46	2.34	0.40	2.79	0.30	4.01	
14	0.73	1.39	0.64	1.60	0.59	1.76	0.47	2.28	0.40	2.70	0.31	3.82	
15	0.73	1.38	0.64	1.59	0.59	1.73	0.47	2.23	0.41	2.62	0.31	3.67	
16	0.74	1.37	0.65	1.57	0.60	1.71	0.48	2.18	0.42	2.56	0.32	3.53	
17	0.74	1.36	0.65	1.55	0.60	1.69	0.48	2.14	0.42	2.50	0.32	3.42	
18	0.75	1.35	0.66	1.54	0.61	1.67	0.49	2.11	0.43	2.45	0.33	3.32	
19	0.75	1.34	0.66	1.53	0.61	1.66	0.49	2.08	0.43	2.41	0.33	3.23	
20	0.75	1.34	0.66	1.52	0.61	1.64	0.50	2.05	0.44	2.37	0.33	3.16	
21	0.75	1.33	0.67	1.51	0.62	1.63	0.50	2.02	0.44	2.33	0.34	3.09	
22	0.76	1.32	0.67	1.50	0.62	1.62	0.51	2.00	0.44	2.30	0.34	3.03	
23	0.76	1.32	0.67	1.49	0.62	1.61	0.51	1.98	0.45	2.27	0.34	2.97	
24	0.76	1.31	0.68	1.48	0.63	1.60	0.51	1.96	0.45	2.24	0.35	2.92	
25	0.76	1.31	0.68	1.48	0.63	1.59	0.52	1.95	0.45	2.22	0.35	2.88	
26	0.77	1.31	0.68	1.47	0.63	1.58	0.52	1.93	0.46	2.19	0.35	2.84	

$(\nu_1 = 27)$													
$\nu_2$	50%		67%		75%		90%		95%		99%		
3	0.57	1.95	0.45	2.72	0.40	3.40	0.28	6.60	0.23	10.65	0.16	31.35	
4	0.61	1.77	0.50	2.34	0.44	2.80	0.32	4.80	0.27	7.01	0.19	16.23	
5	0.63	1.66	0.53	2.12	0.47	2.49	0.35	3.94	0.29	5.43	0.21	10.93	
6	0.65	1.59	0.55	1.98	0.49	2.29	0.37	3.44	0.32	4.57	0.23	8.38	
7	0.67	1.54	0.57	1.89	0.51	2.15	0.39	3.12	0.33	4.02	0.24	6.92	
8	0.68	1.50	0.58	1.81	0.53	2.05	0.41	2.89	0.35	3.65	0.26	5.98	
9	0.69	1.47	0.60	1.76	0.54	1.97	0.42	2.72	0.36	3.38	0.27	5.33	
10	0.70	1.45	0.61	1.71	0.55	1.91	0.44	2.59	0.37	3.18	0.28	4.86	
11	0.71	1.43	0.62	1.68	0.56	1.86	0.45	2.49	0.38	3.02	0.29	4.50	
12	0.72	1.41	0.63	1.65	0.57	1.82	0.45	2.40	0.39	2.89	0.30	4.22	
13	0.73	1.40	0.63	1.62	0.58	1.78	0.46	2.33	0.40	2.78	0.30	3.99	
14	0.73	1.38	0.64	1.60	0.59	1.75	0.47	2.27	0.41	2.69	0.31	3.80	
15	0.74	1.37	0.64	1.58	0.59	1.73	0.48	2.22	0.41	2.61	0.32	3.65	
16	0.74	1.36	0.65	1.56	0.60	1.71	0.48	2.17	0.42	2.55	0.32	3.52	
17	0.74	1.35	0.65	1.55	0.60	1.69	0.49	2.13	0.43	2.49	0.33	3.40	
18	0.75	1.35	0.66	1.54	0.61	1.67	0.49	2.10	0.43	2.44	0.33	3.30	
19	0.75	1.34	0.66	1.52	0.61	1.65	0.50	2.07	0.44	2.39	0.33	3.21	
20	0.75	1.33	0.67	1.51	0.62	1.64	0.50	2.04	0.44	2.35	0.34	3.14	
21	0.76	1.33	0.67	1.50	0.62	1.62	0.51	2.01	0.44	2.32	0.34	3.07	
22	0.76	1.32	0.67	1.49	0.62	1.61	0.51	1.99	0.45	2.29	0.35	3.01	
23	0.76	1.32	0.68	1.49	0.63	1.60	0.51	1.97	0.45	2.26	0.35	2.95	
24	0.76	1.31	0.68	1.48	0.63	1.59	0.52	1.95	0.45	2.23	0.35	2.90	
25	0.77	1.31	0.68	1.47	0.63	1.58	0.52	1.94	0.46	2.20	0.35	2.86	
26	0.77	1.30	0.68	1.46	0.64	1.57	0.52	1.92	0.46	2.18	0.36	2.82	
27	0.77	1.30	0.69	1.46	0.64	1.57	0.52	1.90	0.46	2.16	0.36	2.78	

$(\nu_1 = 28)$													
$\nu_2$	50%		67%		75%		90%		95%		99%		
3	0.57	1.95	0.45	2.72	0.40	3.39	0.28	6.59	0.23	10.63	0.16	31.30	
4	0.61	1.77	0.50	2.33	0.44	2.80	0.32	4.79	0.27	6.99	0.19	16.20	
5	0.63	1.66	0.53	2.12	0.47	2.48	0.35	3.93	0.29	5.42	0.21	10.90	
6	0.66	1.59	0.55	1.98	0.50	2.28	0.38	3.44	0.32	4.55	0.23	8.35	
7	0.67	1.54	0.57	1.88	0.52	2.14	0.40	3.11	0.34	4.01	0.25	6.89	
8	0.69	1.50	0.59	1.81	0.53	2.04	0.41	2.88	0.35	3.64	0.26	5.96	
9	0.70	1.47	0.60	1.76	0.55	1.97	0.43	2.71	0.36	3.37	0.27	5.31	
10	0.71	1.45	0.61	1.71	0.56	1.91	0.44	2.58	0.38	3.17	0.28	4.84	
11	0.71	1.43	0.62	1.67	0.57	1.86	0.45	2.48	0.39	3.00	0.29	4.48	
12	0.72	1.41	0.63	1.64	0.58	1.81	0.46	2.39	0.40	2.87	0.30	4.20	
13	0.73	1.39	0.63	1.62	0.58	1.78	0.47	2.32	0.40	2.77	0.31	3.97	
14	0.73	1.38	0.64	1.60	0.59	1.75	0.47	2.26	0.41	2.68	0.31	3.78	
15	0.74	1.37	0.65	1.58	0.60	1.72	0.48	2.21	0.42	2.60	0.32	3.63	
16	0.74	1.36	0.65	1.56	0.60	1.70	0.49	2.16	0.42	2.53	0.32	3.50	
17	0.75	1.35	0.66	1.54	0.61	1.68	0.49	2.12	0.43	2.48	0.33	3.38	
18	0.75	1.34	0.66	1.53	0.61	1.66	0.50	2.09	0.43	2.43	0.33	3.28	
19	0.75	1.34	0.67	1.52	0.62	1.65	0.50	2.06	0.44	2.38	0.34	3.20	
20	0.76	1.33	0.67	1.51	0.62	1.63	0.51	2.03	0.44	2.34	0.34	3.12	
21	0.76	1.32	0.67	1.50	0.62	1.62	0.51	2.01	0.45	2.31	0.35	3.05	
22	0.76	1.32	0.68	1.49	0.63	1.61	0.51	1.98	0.45	2.27	0.35	2.99	
23	0.76	1.31	0.68	1.48	0.63	1.60	0.52	1.96	0.45	2.24	0.35	2.93	
24	0.77	1.31	0.68	1.47	0.63	1.59	0.52	1.94	0.46	2.22	0.36	2.88	
25	0.77	1.30	0.68	1.47	0.64	1.58	0.52	1.93	0.46	2.19	0.36	2.84	
26	0.77	1.30	0.69	1.46	0.64	1.57	0.53	1.91	0.46	2.17	0.36	2.80	
27	0.77	1.30	0.69	1.45	0.64	1.56	0.53	1.90	0.47	2.15	0.36	2.76	
28	0.77	1.29	0.69	1.45	0.64	1.55	0.53	1.88	0.47	2.13	0.37	2.72	

$(\nu_1 = 29)$													
$\nu_2$	50%		67%		75%		90%		95%		99%		
3	0.57	1.95	0.46	2.71	0.40	3.39	0.28	6.58	0.23	10.61	0.16	31.23	
4	0.61	1.77	0.50	2.33	0.44	2.79	0.32	4.78	0.27	6.98	0.19	16.15	
5	0.64	1.66	0.53	2.11	0.47	2.48	0.35	3.92	0.30	5.41	0.21	10.87	
6	0.66	1.59	0.55	1.98	0.50	2.28	0.38	3.43	0.32	4.54	0.23	8.33	
7	0.67	1.54	0.57	1.88	0.52	2.14	0.40	3.10	0.34	4.00	0.25	6.87	
8	0.69	1.50	0.59	1.81	0.53	2.04	0.41	2.88	0.35	3.63	0.26	5.94	
9	0.70	1.47	0.60	1.75	0.55	1.96	0.43	2.71	0.37	3.36	0.27	5.29	
10	0.71	1.44	0.61	1.71	0.56	1.90	0.44	2.57	0.38	3.15	0.28	4.82	
11	0.71	1.42	0.62	1.67	0.57	1.85	0.45	2.47	0.39	2.99	0.29	4.46	
12	0.72	1.41	0.63	1.64	0.58	1.81	0.46	2.38	0.40	2.86	0.30	4.18	
13	0.73	1.39	0.64	1.61	0.59	1.78	0.47	2.31	0.41	2.76	0.31	3.95	
14	0.73	1.38	0.64	1.59	0.59	1.75	0.48	2.25	0.41	2.67	0.32	3.77	
15	0.74	1.37	0.65	1.57	0.60	1.72	0.48	2.20	0.42	2.59	0.32	3.61	
16	0.74	1.36	0.65	1.56	0.60	1.70	0.49	2.16	0.43	2.52	0.33	3.48	
17	0.75	1.35	0.66	1.54	0.61	1.68	0.49	2.12	0.43	2.47	0.33	3.36	
18	0.75	1.34	0.66	1.53	0.61	1.66	0.50	2.08	0.44	2.42	0.34	3.26	
19	0.75	1.33	0.67	1.52	0.62	1.64	0.50	2.05	0.44	2.37	0.34	3.18	
20	0.76	1.33	0.67	1.50	0.62	1.63	0.51	2.02	0.45	2.33	0.35	3.10	
21	0.76	1.32	0.67	1.49	0.63	1.61	0.51	2.00	0.45	2.29	0.35	3.03	
22	0.76	1.32	0.68	1.49	0.63	1.60	0.52	1.98	0.46	2.26	0.35	2.97	
23	0.76	1.31	0.68	1.48	0.63	1.59	0.52	1.95	0.46	2.23	0.36	2.92	
24	0.77	1.31	0.68	1.47	0.64	1.58	0.52	1.94	0.46	2.21	0.36	2.87	
25	0.77	1.30	0.0	1.46	0.64	1.57	0.53	1.92	0.47	2.18	0.36	2.82	
26	0.77	1.30	0.69	1.46	0.64	1.56	0.53	1.90	0.47	2.16	0.37	2.78	
27	0.77	1.29	0.69	1.45	0.64	1.56	0.53	1.89	0.47	2.14	0.37	2.74	
28	0.77	1.29	0.69	1.44	0.65	1.55	0.53	1.87	0.47	2.12	0.37	2.71	
29	0.78	1.29	0.70	1.44	0.65	1.54	0.54	1.86	0.48	2.10	0.37	2.67	

$(\nu_1 = 30)$													
$\nu_2$	50%		67%		75%		90%		95%		99%		
3	0.57	1.95	0.46	2.71	0.40	3.38	0.29	6.57	0.23	10.60	0.16	31.19	
4	0.61	1.77	0.50	2.33	0.44	2.79	0.33	4.77	0.27	6.97	0.19	16.12	
5	0.64	1.66	0.53	2.11	0.47	2.47	0.36	3.92	0.30	5.39	0.21	10.84	
6	0.66	1.59	0.55	1.97	0.50	2.27	0.38	3.42	0.32	4.53	0.23	8.31	
7	0.67	1.54	0.57	1.88	0.52	2.14	0.40	3.10	0.34	3.99	0.25	6.85	
8	0.69	1.50	0.59	1.80	0.53	2.04	0.42	2.87	0.36	3.62	0.26	5.92	
9	0.70	1.47	0.60	1.75	0.55	1.96	0.43	2.70	0.37	3.35	0.28	5.27	
10	0.71	1.44	0.61	1.70	0.56	1.90	0.44	2.57	0.38	3.14	0.29	4.80	
11	0.72	1.42	0.62	1.67	0.57	1.85	0.45	2.46	0.39	2.98	0.30	4.44	
12	0.72	1.40	0.63	1.64	0.58	1.81	0.46	2.38	0.40	2.85	0.30	4.16	
13	0.73	1.39	0.64	1.61	0.59	1.77	0.47	2.31	0.41	2.75	0.31	3.94	
14	0.74	1.38	0.64	1.59	0.59	1.74	0.48	2.24	0.42	2.66	0.32	3.75	
15	0.74	1.37	0.65	1.57	0.60	1.71	0.49	2.19	0.42	2.58	0.33	3.59	
16	0.74	1.36	0.66	1.55	0.61	1.69	0.49	2.15	0.43	2.51	0.33	3.46	
17	0.75	1.35	0.66	1.54	0.61	1.67	0.50	2.11	0.44	2.46	0.34	3.35	
18	0.75	1.34	0.67	1.52	0.62	1.65	0.50	2.07	0.44	2.41	0.34	3.25	
19	0.76	1.33	0.67	1.51	0.62	1.64	0.51	2.04	0.45	2.36	0.35	3.16	
20	0.76	1.32	0.67	1.50	0.63	1.62	0.51	2.02	0.45	2.32	0.35	3.09	
21	0.76	1.32	0.68	1.49	0.63	1.61	0.52	1.99	0.45	2.28	0.35	3.02	
22	0.76	1.31	0.68	1.48	0.63	1.60	0.52	1.97	0.46	2.25	0.36	2.96	
23	0.77	1.31	0.68	1.47	0.64	1.59	0.52	1.95	0.46	2.22	0.36	2.90	
24	0.77	1.30	0.69	1.47	0.64	1.58	0.53	1.93	0.47	2.20	0.36	2.85	
25	0.77	1.30	0.69	1.46	0.64	1.57	0.53	1.91	0.47	2.17	0.37	2.81	
26	0.77	1.30	0.69	1.45	0.64	1.56	0.53	1.89	0.47	2.15	0.37	2.76	
27	0.77	1.29	0.69	1.45	0.65	1.55	0.54	1.88	0.47	2.13	0.37	2.73	
28	0.78	1.29	0.70	1.44	0.65	1.54	0.54	1.87	0.48	2.11	0.38	2.69	
29	0.78	1.29	0.70	1.43	0.65	1.54	0.54	1.85	0.48	2.09	0.38	2.66	
30	0.78	1.28	0.70	1.43	0.65	1.53	0.54	1.84	0.48	2.07	0.38	2.63	

$(\nu_1 = 35)$													
$\nu_2$	50%		67%		75%		90%		95%		99%		
3	0.57	1.94	0.46	2.70	0.40	3.37	0.29	6.53	0.24	10.53	0.17	30.96	
4	0.61	1.76	0.50	2.31	0.45	2.77	0.33	4.74	0.28	6.91	0.20	15.98	
5	0.64	1.65	0.53	2.10	0.48	2.46	0.36	3.88	0.30	5.35	0.22	10.73	
6	0.66	1.58	0.56	1.96	0.50	2.26	0.39	3.39	0.33	4.49	0.24	8.21	
7	0.68	1.53	0.58	1.86	0.52	2.12	0.41	3.07	0.35	3.95	0.26	6.76	
8	0.69	1.49	0.59	1.79	0.54	2.02	0.42	2.84	0.36	3.58	0.27	5.84	
9	0.70	1.46	0.61	1.74	0.56	1.94	0.44	2.67	0.38	3.31	0.29	5.20	
10	0.71	1.43	0.62	1.69	0.57	1.88	0.45	2.54	0.39	3.10	0.30	4.73	
11	0.72	1.41	0.63	1.65	0.58	1.83	0.46	2.43	0.40	2.94	0.31	4.37	
12	0.73	1.40	0.64	1.62	0.59	1.79	0.47	2.35	0.41	2.81	0.32	4.09	
13	0.74	1.38	0.65	1.60	0.60	1.75	0.48	2.27	0.42	2.70	0.32	3.86	
14	0.74	1.37	0.65	1.57	0.60	1.72	0.49	2.21	0.43	2.61	0.33	3.68	
15	0.75	1.36	0.66	1.56	0.61	1.70	0.50	2.16	0.44	2.54	0.34	3.52	
16	0.75	1.35	0.66	1.54	0.62	1.67	0.50	2.12	0.44	2.47	0.35	3.39	
17	0.76	1.34	0.67	1.52	0.62	1.65	0.51	2.08	0.45	2.41	0.35	3.28	
18	0.76	1.33	0.67	1.51	0.63	1.63	0.51	2.04	0.45	2.36	0.36	3.18	
19	0.76	1.32	0.68	1.50	0.63	1.62	0.52	2.01	0.46	2.32	0.36	3.09	
20	0.77	1.31	0.68	1.49	0.64	1.60	0.52	1.98	0.46	2.28	0.37	3.02	
21	0.77	1.31	0.69	1.47	0.64	1.59	0.53	1.96	0.47	2.24	0.37	2.95	
22	0.77	1.30	0.69	1.47	0.64	1.58	0.53	1.93	0.47	2.21	0.37	2.89	
23	0.77	1.30	0.69	1.46	0.65	1.57	0.54	1.91	0.48	2.18	0.38	2.83	
24	0.78	1.29	0.70	1.45	0.65	1.56	0.54	1.89	0.48	2.15	0.38	2.78	
25	0.78	1.29	0.70	1.44	0.65	1.55	0.54	1.88	0.48	2.13	0.38	2.74	
26	0.78	1.29	0.70	1.44	0.66	1.54	0.55	1.86	0.49	2.11	0.39	2.70	
27	0.78	1.28	0.70	1.43	0.66	1.53	0.55	1.85	0.49	2.08	0.39	2.66	
28	0.78	1.28	0.71	1.42	0.66	1.52	0.55	1.83	0.49	2.07	0.39	2.62	
29	0.79	1.27	0.71	1.42	0.66	1.52	0.56	1.82	0.50	2.05	0.40	2.59	
30	0.79	1.27	0.71	1.41	0.67	1.51	0.56	1.81	0.50	2.03	0.40	2.56	
35	0.79	1.26	0.72	1.39	0.68	1.48	0.57	1.76	0.51	1.96	0.41	2.44	

$(\nu_1 = 40)$													
$\nu_2$	50%		67%		75%		90%		95%		99%		
3	0.57	1.94	0.46	2.69	0.41	3.35	0.29	6.50	0.24	10.47	0.17	30.78	
4	0.61	1.75	0.51	2.31	0.45	2.76	0.33	4.71	0.28	6.87	0.20	15.87	
<i>S</i>	0.64	1.65	0.54	2.09	0.48	2.44	0.37	3.86	0.31	5.31	0.23	10.65	
6	0.66	1.58	0.56	1.95	0.51	2.25	0.39	3.37	0.33	4.45	0.25	8.14	
7	0.68	1.52	0.58	1.85	0.53	2.11	0.41	3.04	0.35	3.91	0.27	6.70	
8	0.70	1.48	0.60	1.78	0.55	2.01	0.43	2.81	0.37	3.54	0.28	5.77	
9	0.71	1.45	0.61	1.73	0.56	1.93	0.45	2.64	0.39	3.27	0.29	5.13	
10	0.72	1.43	0.62	1.68	0.57	1.87	0.46	2.51	0.40	3.07	0.31	4.67	
11	0.73	1.41	0.63	1.64	0.58	1.82	0.47	2.41	0.41	2.91	0.32	4.31	
12	0.73	1.39	0.64	1.61	0.59	1.77	0.48	2.32	0.42	2.78	0.33	4.04	
13	0.74	1.37	0.65	1.59	0.60	1.74	0.49	2.25	0.43	2.67	0.33	3.81	
14	0.75	1.36	0.66	1.56	0.61	1.71	0.50	2.19	0.44	2.58	0.34	3.63	
15	0.75	1.35	0.67	1.54	0.62	1.68	0.51	2.14	0.45	2.51	0.35	3.47	
16	0.76	1.34	0.67	1.53	0.62	1.66	0.51	2.09	0.45	2.44	0.36	3.34	
17	0.76	1.33	0.68	1.51	0.63	1.64	0.52	2.05	0.46	2.38	0.36	3.23	
18	0.76	1.32	0.68	1.50	0.63	1.62	0.52	2.02	0.47	2.33	0.37	3.13	
19	0.77	1.31	0.69	1.48	0.64	1.60	0.53	1.99	0.47	2.29	0.37	3.04	
20	0.77	1.31	0.69	1.47	0.64	1.59	0.53	1.96	0.48	2.25	0.38	2.97	
21	0.77	1.30	0.69	1.46	0.65	1.58	0.54	1.93	0.48	2.21	0.38	2.90	
22	0.78	1.30	0.70	1.45	0.65	1.56	0.54	1.91	0.48	2.18	0.39	2.84	
23	0.78	1.29	0.70	1.44	0.66	1.55	0.55	1.89	0.49	2.15	0.39	2.78	
24	0.78	1.29	0.70	1.44	0.66	1.54	0.55	1.87	0.49	2.12	0.40	2.73	
25	0.78	1.28	0.71	1.43	0.66	1.53	0.56	1.85	0.50	2.10	0.40	2.69	
26	0.79	1.28	0.71	1.42	0.66	1.52	0.56	1.84	0.50	2.07	0.40	2.64	
27	0.79	1.27	0.71	1.42	0.67	1.51	0.56	1.82	0.50	2.05	0.41	2.61	
28	0.79	1.27	0.71	1.41	0.67	1.51	0.56	1.81	0.51	2.03	0.41	2.57	
29	0.79	1.27	0.72	1.40	0.67	1.50	0.57	1.79	0.51	2.01	0.41	2.54	
30	0.79	1.26	0.72	1.40	0.67	1.49	0.57	1.78	0.51	2.00	0.41	2.51	
35	0.80	1.25	0.73	1.38	0.68	1.46	0.58	1.73	0.52	1.93	0.43	2.39	
40	0.81	1.24	0.73	1.36	0.69	1.44	0.59	1.69	0.53	1.88	0.44	2.30	

(v <sub>1</sub> = 45)													
v <sub>2</sub>	50%		67%		75%		90%		95%		99%		
3	0.58	1.93	0.46	2.68	0.41	3.35	0.30	6.48	0.24	10.43	0.17	30.67	
4	0.62	1.75	0.51	2.30	0.45	2.75	0.34	4.69	0.28	6.84	0.20	15.80	
5	0.65	1.64	0.54	2.08	0.49	2.44	0.37	3.84	0.31	5.28	0.23	10.59	
6	0.67	1.57	0.57	1.94	0.51	2.24	0.40	3.35	0.34	4.42	0.25	8.08	
7	0.68	1.52	0.59	1.85	0.53	2.10	0.42	3.02	0.36	3.89	0.27	6.64	
8	0.70	1.48	0.60	1.77	0.55	2.00	0.44	2.80	0.38	3.52	0.29	5.73	
9	0.71	1.45	0.62	1.72	0.57	1.92	0.45	2.63	0.39	3.25	0.30	5.09	
10	0.72	1.42	0.63	1.67	0.58	1.86	0.46	2.50	0.41	3.05	0.31	4.62	
11	0.73	1.40	0.64	1.64	0.59	1.81	0.48	2.39	0.42	2.88	0.32	4.27	
12	0.74	1.38	0.65	1.60	0.60	1.76	0.49	2.30	0.43	2.76	0.33	3.99	
13	0.74	1.37	0.66	1.58	0.61	1.73	0.50	2.23	0.44	2.65	0.34	3.77	
14	0.75	1.36	0.66	1.56	0.62	1.70	0.50	2.17	0.45	2.56	0.35	3.58	
15	0.75	1.34	0.67	1.54	0.62	1.67	0.51	2.12	0.45	2.48	0.36	3.43	
16	0.76	1.33	0.68	1.52	0.63	1.65	0.52	2.07	0.46	2.41	0.37	3.30	
17	0.76	1.32	0.68	1.50	0.64	1.63	0.53	2.03	0.47	2.36	0.37	3.19	
18	0.77	1.32	0.69	1.49	0.64	1.61	0.53	2.00	0.47	2.31	0.38	3.09	
19	0.77	1.31	0.69	1.48	0.65	1.59	0.54	1.97	0.48	2.26	0.38	3.00	
20	0.78	1.30	0.70	1.46	0.65	1.58	0.54	1.94	0.48	2.22	0.39	2.93	
21	0.78	1.30	0.70	1.45	0.65	1.56	0.55	1.91	0.49	2.18	0.39	2.86	
22	0.78	1.29	0.70	1.44	0.66	1.55	0.55	1.89	0.49	2.15	0.40	2.80	
23	0.78	1.28	0.71	1.44	0.66	1.54	0.56	1.87	0.50	2.12	0.40	2.74	
24	0.79	1.28	0.71	1.43	0.67	1.53	0.56	1.85	0.50	2.10	0.41	2.69	
25	0.79	1.27	0.71	1.42	0.67	1.52	0.56	1.83	0.51	2.07	0.41	2.65	
26	0.79	1.27	0.72	1.41	0.67	1.51	0.57	1.82	0.51	2.05	0.41	2.60	
27	0.79	1.27	0.72	1.41	0.67	1.50	0.57	1.80	0.51	2.03	0.42	2.57	
28	0.79	1.26	0.72	1.40	0.68	1.49	0.57	1.79	0.52	2.01	0.42	2.53	
29	0.80	1.26	0.72	1.39	0.68	1.49	0.58	1.77	0.52	1.99	0.42	2.50	
30	0.80	1.26	0.72	1.39	0.68	1.48	0.58	1.76	0.52	1.97	0.43	2.47	
35	0.81	1.24	0.73	1.37	0.69	1.45	0.59	1.71	0.53	1.90	0.44	2.35	
40	0.81	1.23	0.74	1.35	0.70	1.43	0.60	1.67	0.55	1.85	0.45	2.25	
45	0.82	1.22	0.75	1.34	0.71	1.41	0.61	1.64	0.55	1.81	0.46	2.19	

$(\nu_1 = 50)$													
$\nu_2$	50%		67%		75%		90%		95%		99%		
3	0.58	1.93	0.47	2.68	0.41	3.34	0.30	6.46	0.25	10.40	0.18	30.57	
4	0.62	1.75	0.51	2.29	0.46	2.75	0.34	4.67	0.29	6.82	0.21	15.72	
5	0.65	1.64	0.54	2.08	0.49	2.43	0.37	3.83	0.32	5.26	0.23	10.53	
6	0.67	1.57	0.57	1.94	0.52	2.23	0.40	3.33	0.34	4.40	0.26	8.04	
7	0.69	1.52	0.59	1.84	0.54	2.09	0.42	3.01	0.36	3.87	0.28	6.61	
8	0.70	1.48	0.61	1.77	0.55	1.99	0.44	2.78	0.38	3.50	0.29	5.69	
9	0.71	1.44	0.62	1.71	0.57	1.91	0.46	2.61	0.40	3.23	0.31	5.05	
10	0.72	1.42	0.63	1.67	0.58	1.85	0.47	2.48	0.41	3.03	0.32	4.59	
11	0.73	1.40	0.64	1.63	0.59	1.80	0.48	2.38	0.42	2.87	0.33	4.24	
12	0.74	1.38	0.65	1.60	0.60	1.76	0.49	2.29	0.43	2.74	0.34	3.96	
13	0.75	1.36	0.66	1.57	0.61	1.72	0.50	2.22	0.44	2.63	0.35	3.74	
14	0.75	1.35	0.67	1.55	0.62	1.69	0.51	2.16	0.45	2.54	0.36	3.55	
15	0.76	1.34	0.67	1.53	0.63	1.66	0.52	2.10	0.46	2.46	0.37	3.40	
16	0.76	1.33	0.68	1.51	0.63	1.64	0.53	2.06	0.47	2.39	0.37	3.27	
17	0.77	1.32	0.69	1.49	0.64	1.62	0.53	2.02	0.47	2.34	0.38	3.15	
18	0.77	1.31	0.69	1.48	0.65	1.60	0.54	1.98	0.48	2.29	0.39	3.05	
19	0.77	1.30	0.70	1.47	0.65	1.58	0.54	1.95	0.49	2.24	0.39	2.97	
20	0.78	1.30	0.70	1.46	0.66	1.57	0.55	1.92	0.49	2.20	0.40	2.89	
21	0.78	1.29	0.70	1.45	0.66	1.55	0.55	1.90	0.50	2.16	0.40	2.82	
22	0.78	1.28	0.71	1.44	0.66	1.54	0.56	1.87	0.50	2.13	0.41	2.76	
23	0.79	1.28	0.71	1.43	0.67	1.53	0.56	1.85	0.51	2.10	0.41	2.71	
24	0.79	1.27	0.71	1.42	0.67	1.52	0.57	1.83	0.51	2.07	0.42	2.66	
25	0.79	1.27	0.72	1.41	0.67	1.51	0.57	1.82	0.51	2.05	0.42	2.61	
26	0.79	1.27	0.72	1.40	0.68	1.50	0.57	1.80	0.52	2.03	0.42	2.57	
27	0.80	1.26	0.72	1.40	0.68	1.49	0.58	1.78	0.52	2.00	0.43	2.53	
28	0.80	1.26	0.73	1.39	0.68	1.48	0.58	1.77	0.52	1.99	0.43	2.50	
29	0.80	1.25	0.73	1.39	0.69	1.48	0.58	1.76	0.53	1.97	0.43	2.47	
30	0.80	1.25	0.73	1.38	0.69	1.47	0.59	1.74	0.53	1.95	0.44	2.44	
35	0.81	1.24	0.74	1.36	0.70	1.44	0.60	1.69	0.54	1.88	0.45	2.31	
40	0.82	1.23	0.75	1.34	0.71	1.42	0.61	1.65	0.55	1.83	0.46	2.22	
45	0.82	1.22	0.75	1.33	0.71	1.40	0.62	1.62	0.56	1.78	0.47	2.15	
50	0.83	1.21	0.76	1.32	0.72	1.39	0.63	1.60	0.57	1.75	0.48	2.10	

$(\nu_1 = 55)$													
$\nu_2$	50%		67%		75%		90%		95%		99%		
3	0.58	1.93	0.47	2.67	0.41	3.33	0.30	6.44	0.25	10.38	0.18	30.49	
4	0.62	1.74	0.51	2.29	0.46	2.74	0.34	4.66	0.29	6.80	0.21	15.67	
5	0.65	1.64	0.55	2.07	0.49	2.42	0.38	3.81	0.32	5.24	0.24	10.50	
6	0.67	1.56	0.57	1.93	0.52	2.22	0.40	3.32	0.35	4.39	0.26	8.00	
7	0.69	1.51	0.59	1.84	0.54	2.08	0.42	3.00	0.37	3.85	0.28	6.57	
8	0.70	1.47	0.61	1.76	0.56	1.98	0.44	2.77	0.38	3.48	0.30	5.66	
9	0.71	1.44	0.62	1.71	0.57	1.90	0.46	2.60	0.40	3.21	0.31	5.02	
10	0.72	1.42	0.63	1.66	0.59	1.84	0.47	2.47	0.41	3.01	0.32	4.56	
11	0.73	1.39	0.65	1.62	0.60	1.79	0.49	2.36	0.43	2.85	0.33	4.21	
12	0.74	1.38	0.65	1.59	0.61	1.75	0.50	2.28	0.44	2.72	0.35	3.93	
13	0.75	1.36	0.66	1.57	0.62	1.71	0.51	2.21	0.45	2.61	0.36	3.71	
14	0.75	1.35	0.67	1.54	0.62	1.68	0.52	2.14	0.46	2.52	0.36	3.52	
15	0.76	1.34	0.68	1.52	0.63	1.66	0.52	2.09	0.47	2.44	0.37	3.37	
16	0.77	1.32	0.68	1.50	0.64	1.63	0.53	2.05	0.47	2.38	0.38	3.24	
17	0.77	1.32	0.69	1.49	0.64	1.61	0.54	2.01	0.48	2.32	0.39	3.13	
18	0.77	1.31	0.69	1.47	0.65	1.59	0.54	1.97	0.49	2.27	0.39	3.03	
19	0.78	1.30	0.70	1.46	0.65	1.58	0.55	1.94	0.49	2.22	0.40	2.94	
20	0.78	1.29	0.70	1.45	0.66	1.56	0.56	1.91	0.50	2.18	0.40	2.87	
21	0.78	1.29	0.71	1.44	0.66	1.55	0.56	1.89	0.50	2.15	0.41	2.80	
22	0.79	1.28	0.71	1.43	0.67	1.53	0.56	1.86	0.51	2.11	0.41	2.74	
23	0.79	1.28	0.72	1.42	0.67	1.52	0.57	1.84	0.51	2.08	0.42	2.68	
24	0.79	1.27	0.72	1.41	0.68	1.51	0.57	1.82	0.52	2.06	0.42	2.63	
25	0.80	1.27	0.72	1.41	0.68	1.50	0.58	1.80	0.52	2.03	0.43	2.59	
26	0.80	1.26	0.72	1.40	0.68	1.49	0.58	1.79	0.53	2.01	0.43	2.55	
27	0.80	1.26	0.73	1.39	0.69	1.48	0.58	1.77	0.53	1.99	0.43	2.51	
28	0.80	1.25	0.73	1.39	0.69	1.48	0.59	1.76	0.53	1.97	0.44	2.47	
29	0.80	1.25	0.73	1.38	0.69	1.47	0.59	1.74	0.54	1.95	0.44	2.44	
30	0.81	1.25	0.73	1.37	0.69	1.46	0.59	1.73	0.54	1.93	0.44	2.41	
35	0.81	1.23	0.74	1.35	0.70	1.43	0.61	1.68	0.55	1.86	0.46	2.28	
40	0.82	1.22	0.75	1.33	0.71	1.41	0.62	1.64	0.56	1.81	0.47	2.19	
45	0.83	1.21	0.76	1.32	0.72	1.39	0.63	1.61	0.57	1.77	0.48	2.12	
50	0.83	1.21	0.76	1.31	0.73	1.38	0.63	1.58	0.58	1.73	0.49	2.07	
55	0.83	1.20	0.77	1.30	0.73	1.37	0.64	1.56	0.59	1.71	0.49	2.02	

(v <sub>1</sub> = 60)													
v <sub>2</sub>	50%	67%	75%	90%	95%	99%							
3	0.58	1.93	0.47	2.67	0.41	3.33	0.30	6.43	0.25	10.36	0.18	30.43	
4	0.62	1.74	0.51	2.29	0.46	2.73	0.34	4.65	0.29	6.78	0.21	15.63	
5	0.65	1.63	0.55	2.07	0.49	2.42	0.38	3.80	0.32	5.22	0.24	10.45	
6	0.67	1.56	0.57	1.93	0.52	2.22	0.40	3.31	0.35	4.37	0.26	7.97	
7	0.69	1.51	0.59	1.83	0.54	2.08	0.43	2.99	0.37	3.83	0.28	6.54	
8	0.70	1.47	0.61	1.76	0.56	1.98	0.45	2.76	0.39	3.47	0.30	5.63	
9	0.72	1.44	0.62	1.70	0.57	1.90	0.46	2.59	0.40	3.20	0.31	5.00	
10	0.73	1.41	0.64	1.66	0.59	1.84	0.48	2.46	0.42	3.00	0.33	4.53	
11	0.74	1.39	0.65	1.62	0.60	1.79	0.49	2.35	0.43	2.84	0.34	4.18	
12	0.74	1.37	0.66	1.59	0.61	1.74	0.50	2.27	0.44	2.71	0.35	3.91	
13	0.75	1.36	0.67	1.56	0.62	1.71	0.51	2.20	0.45	2.60	0.36	3.68	
14	0.76	1.34	0.67	1.54	0.63	1.68	0.52	2.13	0.46	2.51	0.37	3.50	
15	0.76	1.33	0.68	1.52	0.63	1.65	0.53	2.08	0.47	2.43	0.38	3.35	
16	0.77	1.32	0.69	1.50	0.64	1.63	0.53	2.04	0.48	2.36	0.38	3.22	
17	0.77	1.31	0.69	1.48	0.65	1.60	0.54	2.00	0.48	2.31	0.39	3.10	
18	0.78	1.30	0.70	1.47	0.65	1.59	0.55	1.96	0.49	2.26	0.40	3.01	
19	0.78	1.30	0.70	1.46	0.66	1.57	0.55	1.93	0.50	2.21	0.40	2.92	
20	0.78	1.29	0.71	1.44	0.66	1.55	0.56	1.90	0.50	2.17	0.41	2.84	
21	0.79	1.28	0.71	1.43	0.67	1.54	0.56	1.87	0.51	2.13	0.42	2.78	
22	0.79	1.28	0.71	1.42	0.67	1.53	0.57	1.85	0.51	2.10	0.42	2.71	
23	0.79	1.27	0.72	1.42	0.68	1.52	0.57	1.83	0.52	2.07	0.43	2.66	
24	0.80	1.27	0.72	1.41	0.68	1.50	0.58	1.81	0.52	2.04	0.43	2.61	
25	0.80	1.26	0.73	1.40	0.68	1.49	0.58	1.79	0.53	2.02	0.43	2.56	
26	0.80	1.26	0.73	1.39	0.69	1.49	0.59	1.78	0.53	1.99	0.44	2.52	
27	0.80	1.25	0.73	1.39	0.69	1.48	0.59	1.76	0.53	1.97	0.44	2.48	
28	0.80	1.25	0.73	1.38	0.69	1.47	0.59	1.75	0.54	1.95	0.45	2.45	
29	0.81	1.25	0.74	1.37	0.70	1.46	0.60	1.73	0.54	1.93	0.45	2.42	
30	0.81	1.24	0.74	1.37	0.70	1.45	0.60	1.72	0.54	1.92	0.45	2.39	
35	0.82	1.23	0.75	1.35	0.71	1.43	0.61	1.67	0.56	1.85	0.47	2.26	
40	0.82	1.22	0.76	1.33	0.72	1.40	0.62	1.63	0.57	1.79	0.48	2.17	
45	0.83	1.21	0.76	1.31	0.73	1.38	0.63	1.60	0.58	1.75	0.49	2.10	
50	0.83	1.20	0.77	1.30	0.73	1.37	0.64	1.57	0.59	1.72	0.50	2.05	
55	0.84	1.20	0.77	1.29	0.74	1.36	0.65	1.55	0.59	1.69	0.50	2.00	
60	0.84	1.19	0.78	1.29	0.74	1.35	0.65	1.53	0.60	1.67	0.51	1.96	

Note: If v<sub>2</sub> > v<sub>1</sub> then an interval corresponding to a P% HDR for logF is given by (1/F, 1/F) where (F, F) is the appropriate interval with v<sub>1</sub> and v<sub>2</sub> interchanged.

# Appendix C: R programs

```
#  
# Functions for HDRs and for Behrens' distribution  
#  
hdrsize <- function(p,n,m,distrib){  
  # Uses algorithm described in Jackson (1974)  
  precision <- 0.0001  
  maxiter <- 100  
  pp <- 1  
  if (p > 1) p <- p/100  
  d0 <- dinitial(p,n,m,distrib)  
  d1 <- d0  
  iter <- 0  
  while (prfn(d1,n,m,distrib) > p) d1 <- d0/2  
  while (prfn(d0,n,m,distrib) < p) d0 <- 2*d0  
  while ((abs(pp)>precision)&&(iter<maxiter)){  
    iter <- iter+1  
    dm <- (d0+d1)/2  
    p0 <- prfn(d0,n,m,distrib)  
    pm <- prfn(dm,n,m,distrib)  
    p1 <- prfn(d1,n,m,distrib)  
    if ((p1-p)*(p-p0) > 0) d0 <- dm  
    if ((pm-p)*(p-p0) > 0) d1 <- dm  
    pp <- abs(p1-p0)  
  }  
  return(d0)  
}  
theta0fn <- function(d,n,m,distrib){  
  if (distrib=="chi") return(d/tanh(d/(n-1)))  
  if (distrib=="invchi") return(d/tanh(d/(n+1)))  
  if (distrib=="gamma") return(d/tanh(d/((n-1))))  
  if (distrib=="invgamma") return(d/tanh(d/((n+1)))))
```

```

if (distrib=="gammafromlogs") return(d/tanh(d/n))
if (distrib=="beta")
  return((d^((n-1)/(m-1))-1)/(d^((n-1)/(m-1)+1)-1))
if (distrib=="f")
  return((m/n)*(d-d^((m+2)/(m+n)))/(d^((m+2)/(m+n))-1))
if (distrib=="ffromlogs")
  return((m/n)*(d-d^(m/(m+n)))/(d^(m/(m+n))-1))
}

prfn <- function(d,n,m,distrib){
  if (distrib=="chi")
    return(pchisq(upperfn(d,n,m,distrib)^2,n)-
           pchisq(lowerfn(d,n,m,distrib)^2,n))
  if (distrib=="invchi")
    return(pchisq(1/lowerfn(d,n,m,distrib)^2,n)-
           pchisq(1/upperfn(d,n,m,distrib)^2,n))
  if (distrib=="gamma")
    return(pgamma(upperfn(d,n,m,distrib),n,1)-
           pgamma(lowerfn(d,n,m,distrib),n,1))
  if (distrib=="invgamma")
    return(pgamma(1/lowerfn(d,n,m,distrib),n,1)-
           pgamma(1/upperfn(d,n,m,distrib),n,1))
  if (distrib=="gammafromlogs")
    return(pgamma(upperfn(d,n,m,distrib),n,1)-
           pgamma(lowerfn(d,n,m,distrib),n,1))
  if (distrib=="beta")
    return(pbta(upperfn(d,n,m,distrib),n,m)-
           pbta(lowerfn(d,n,m,distrib),n,m))
  if (distrib=="f")
    return(pf(upperfn(d,n,m,distrib),n,m)-
           pf(lowerfn(d,n,m,distrib),n,m))
  if (distrib=="ffromlogs")
    return(pf(upperfn(d,n,m,distrib),n,m)-
           pf(lowerfn(d,n,m,distrib),n,m))
}

dinitial <- function(p,n,m,distrib){
  if (distrib=="chi")
    return(sqrt(qchisq(1-p/2,n))-sqrt(qchisq(p/2,n))/2)
  if (distrib=="invchi")
    return(1/sqrt(qchisq(p/2,n))-1/sqrt(qchisq(1-p/2,n))/2)
  if (distrib=="gamma")
    return((qgamma(1-p/2,n,1)-qgamma(p/2,n,1))/2)
  if (distrib=="invgamma")
    return((1/qgamma(p/2,n,1)-1/qgamma(1-p/2,n,1))/2)
  if (distrib=="gammafromlogs")
    return((qgamma(1-p/2,n,1)-qgamma(p/2,n,1))/2)
}

```

```

if (distrib=="beta") return(qbeta(1-p/2,n,m)/qbeta(p/2,n,m))
if (distrib=="f") return(qf(1-p/2,n,m)/qf(p/2,n,m))
if (distrib=="ffromlogs") return(qf(1-p/2,n,m)/qf(p/2,n,m))
}
lowerfn <- function(d0,n,m,distrib){
  if (distrib=="chi")
    return(sqrt(theta0fn(d0,n,m,distrib)-d0))
  if (distrib=="invchi")
    return(1/sqrt(theta0fn(d0,n,m,distrib)+d0))
  if (distrib=="gamma")
    return(theta0fn(d0,n,1,distrib)-d0)
  if (distrib=="invgamma")
    return(1/(theta0fn(d0,n,1,distrib)+d0))
  if (distrib=="gammafromlogs")
    return(theta0fn(d0,n,1,distrib)-d0)
  if (distrib=="beta")
    return(theta0fn(d0,n,m,distrib))
  if (distrib=="f")
    return(d0^(-1)*theta0fn(d0,n,m,distrib))
  if (distrib=="ffromlogs")
    return(d0^(-1)*theta0fn(d0,n,m,distrib))
}
upperfn <- function(d0,n,m,distrib){
  if (distrib=="chi")
    return(sqrt(theta0fn(d0,n,m,distrib)+d0))
  if (distrib=="invchi")
    return(1/sqrt(theta0fn(d0,n,m,distrib)-d0))
  if (distrib=="gamma")
    return(theta0fn(d0,n,1,distrib)+d0)
  if (distrib=="invgamma")
    return(1/(theta0fn(d0,n,1,distrib)-d0))
  if (distrib=="gammafromlogs")
    return(theta0fn(d0,n,1,distrib)+d0)
  if (distrib=="beta")
    return(d0*theta0fn(d0,n,m,distrib))
  if (distrib=="f")
    return(theta0fn(d0,n,m,distrib))
  if (distrib=="ffromlogs")
    return(theta0fn(d0,n,m,distrib))
}
hnorm <- function(p,mean=0,sd=1)
  return(c(qnorm((1-p)/2,mean,sd),qnorm((1+p)/2,mean,sd)))
ht <- function(p,df,ncp=0)
  return(c(qt((1-p)/2,df,ncp),qt((1+p)/2,df,ncp)))
hcauchy <- function(p,location=0,scale=1){

```

```

y <- qcauchy((1-p)/2,location,scale,log=FALSE)
z <- qcauchy((1+p)/2,location,scale,log=FALSE)
return(c(y,z))
}
hchi <- function(p,df,log=TRUE,inverse=FALSE) {
  if (log)
    return(sqrt(hchisq(p,df,log=log,inverse=inverse)))
  else{
    if (inverse){
      d0 <- hdrsize(p,df,1,"invchi")
      return(c(lowerfn(d0,df,1,"invchi"),
              upperfn(d0,df,1,"invchi")))
    }
    else{
      d0 <- hdrsize(p,df,1,"chi")
      return(c(lowerfn(d0,df,1,"chi"),
              upperfn(d0,df,1,"chi"))))
    }
  }
}
hf <- function(p,df1,df2,log=TRUE) {
  if (log){
    d0 <- hdrsize(p,df1,df2,"ffromlogs")
    return(c(lowerfn(d0,df1,df2,"ffromlogs"),
            upperfn(d0,df1,df2,"ffromlogs")))
  }
  else{
    d0 <- hdrsize(p,df1,df2,"f")
    return(c(lowerfn(d0,df1,df2,"f"),
            upperfn(d0,df1,df2,"f"))))
  }
}
hgamma <-
function(p,shape,rate=1,scale=1/rate,log=TRUE,inverse=FALSE) {
  if (log){
    if (inverse){
      d0 <- hdrsize(p,shape,rate,"gammafromlogs")
      return(c(1/upperfn(d0,shape,rate,"gammafromlogs"),
              1/lowerfn(d0,shape,rate,"gammafromlogs"))/(4.*scale))
    }
    else{
      d0 <- hdrsize(p,shape,rate,"gammafromlogs")
      return(4.*scale*c(lowerfn(d0,shape,rate,"gammafromlogs"),
                        upperfn(d0,shape,rate,"gammafromlogs")))
    }
  }
}

```

```

}
else{
  if (inverse){
    d0 <- hdrsize(p,shape,rate,"invgamma")
    return(c(lowerfn(d0,shape,rate,"invgamma"),
            upperfn(d0,shape,rate,"invgamma"))/(4*scale))
  }
  else{
  }
  d0 <- hdrsize(p,shape,rate,"gamma")
  return(4*scale*c(lowerfn(d0,shape,rate,"gamma"),
                    upperfn(d0,shape,rate,"gamma")))
}
}

hchisq <- function(p,df,log=TRUE,inverse=FALSE)
  return(hgamma(p,df/2,2,log=log,inverse=inverse))

hbeta <- function(p,shape1,shape2,log=TRUE){
  if (log){
    return(shape1*hf(p,2*shape1,2*shape2) /
           (shape2+shape1*hf(p,2*shape1,2*shape2)))
  }
  else{
    d0 <- hdrsize(p,shape1,shape2,"beta")
    return(c(lowerfn(d0,shape1,shape2,"beta"),
            upperfn(d0,shape1,shape2,"beta")))
  }
}

#
# Behrens' distribution
#
gbehrens <- function(x,z,n,m,phi,degrees=TRUE){
  if (degrees) phi <- pi*phi/180
  k <- 1/(beta(n/2,1/2)*beta(m/2,1/2)*sqrt(n*m))
  return(k*(1+(z*cos(phi)-x*sin(phi))^2/n)^(-(n+1)/2)*
    (1+(z*sin(phi)+x*cos(phi))^2/m)^(-(m+1)/2))
}

dbehrens <- function(z,n,m,phi,degrees=TRUE){
  g <- function(x) gbehrens(x,z,n,m,phi,degrees)
  return(integrate(g,-Inf,Inf)$value)
}

pbehrens <- function(z,n,m,phi,degrees=TRUE){
  # Uses the function adapt from the package adapt
  # Parameter Large needed as adapt will not accept Inf
  library(adapt)
  Large <- 100
}

```

```

f <- function(x) gbehrens(x[1],x[2],n,m,phi,degrees)
if (z==0) y <- 0.5
if (z > 0)
  y <- 0.5+adapt(ndim=2,
    lower=c(0,-Large),upper=c(z,Large),functn=f)$value
if (z < 0)
  y <- 0.5-adapt(ndim=2,
    lower=c(0,-Large),upper=c(-z,Large),functn=f)$value
if (y < 0) y <- 0
if (y > 1) y <- 1
return(y)
}
qbehrens <- function(p,n,m,phi,degrees=TRUE) {
  precision <- 0.01
  x <- qnorm(p)
  while (pbehrns(x,n,m,phi,degrees) < p) x <- x + precision
  p0 <- pbehrns(x-precision,n,m,phi,degrees)
  p1 <- pbehrns(x,n,m,phi,degrees)
  return(x - precision*(p1-p)/(p1-p0))
}
rbehrens <- function(k,n,m,phi,degrees=TRUE) {
  y <- runif(k)
  for (i in 1:k) y[i] <- qbehrens(y[i],n,m,phi)
  return(y)
}
hbehrens <- function(p,n,m,phi,degrees=TRUE) {
  y <- qbehrens((1-p)/2,n,m,phi,degrees)
  z <- qbehrens((1+p)/2,n,m,phi,degrees)
  x <- (z-y)/2
  return(c(-x,x))
}

```

# Appendix D: Further reading

## D.1 Robustness

Although the importance of robustness (or sensitivity analysis) was mentioned at the end of Section 2.3 on several normal means with a normal prior, not much attention has been devoted to this topic in the rest of the book. Some useful references are Berger (1985, Section 4.7), Box and Tiao (1992, Section 3.2 and *passim.*), Hartigan (1983, Chapter ), Kadane (1984) and O'Hagan and Forster (2004, Chapter ).

## D.2 Nonparametric methods

Throughout this book, it is assumed that the data we are analyzing comes from some parametric family, so that the density  $p(x|\theta)$  of any observation  $x$  depends on one or more parameters  $\theta$  (e.g.  $x$  is normal of mean  $\theta$  and known variance). In classical statistics, much attention has been devoted to developing methods which do not make any such assumption, so that you can, for example, say something about the median of a set of observations without assuming that they come from a normal distribution. Some attempts have been made to develop a Bayesian form of nonparametric theory, though this is not easy as it involves setting up a prior distribution over a very large class of densities for the observations. Useful references are Ferguson (1973), Florens *et al.* (1983), Dalal (1980), Hill (1988), Lenk (1991), Ghosh and Ramamoorthi (2003) and Hjort *et al.* (2010). A brief account is given by Müller and Quintana (2004).

## D.3 Multivariate estimation

In order to provide a reasonably simple introduction to Bayesian statistics, avoiding matrix theory as far as possible, the coverage of this book has been restricted largely to cases where only one measurement is taken at a time. Useful references for multivariate Bayesian statistics are Box and Tiao (1992, Chapter ), Zellner (1971, Chapter ), Press (2009) and O'Hagan and Forster (2004, Sections 10.28–10.41).

## D.4 Time series and forecasting

Methods of dealing with time series, that is, random functions of time, constitute an important area of statistics. Important books on this area from the Bayesian standpoint are West and Harrison (1989) and Pole *et al.* (1994). Information about software updates can be found at

<http://www.isds.duke.edu/~mw/bats.html>

A briefer discussion of some of the ideas can be found in Leonard and Hsu (2001, Section 5.3).

## D.5 Sequential methods

Some idea as to how to apply Bayesian methods in cases where observations are collected sequentially through time can be got from Berger (1985, Chapter ) or O'Hagan and Forster (2004, Sections 3.55–3.57).

## D.6 Numerical methods

This is the area in which most progress in Bayesian statistics has been made in recent years. Although Chapter is devoted to numerical methods, a mere sketch of the basic ideas has been given. Very useful texts with a wealth of examples and full programs in WinBUGS available on an associated website are Congdon (2002, 2005, 2006 and 2010). Those seriously interested in the application of Bayesian methods in the real world should consult Tanner (1993), Gelman *et al.* (2004), Carlin and Louis (2008), Gilks *et al.* (1996), French and Smith (1997) and Brooks (1998).

More recently books giving a useful and comprehensible treatment of Bayesian numerical methods using R and WinBUGS include Albert (2009) (a particularly attractive treatment), Marin and Robert (2007), Robert and Casella (2010) and Ntzoufras (2009).

At a much lower level, Albert (1994) is a useful treatment of elementary Bayesian ideas using Minitab.

## D.7 Bayesian networks

References on this topic (on which I am not an expert) include Jensen (1996),

Jensen and Nielson (2010) and Neapolitan (2004).

## D.8 General reading

Apart from Jeffreys (1939, 1948 and 1961), Berger (1985) and Box and Tiao (1992), which have frequently been referred to, some useful references are Lindley (1971a), DeGroot (1970) and Raiffa and Schlaifer (1961). The more recent texts by Bernardo and Smith (1994) and O'Hagan and Forster (2004) are very important and give a good coverage of the Bayesian theory. Some useful coverage of Bayesian methods for the linear model can be found in Broemling (1985). Linear Bayes methods are covered in Goldstein and Wooff (2007). Anyone interested in Bayesian statistics will gain a great deal by reading de Finetti (1972 and 1974–1975) and Savage (1972 and 1981). A useful collection of essays on the foundations of Bayesian statistics is Kyburg and Smokler (1964 and 1980), and a collection of recent influential papers can be found in Polson and Tiao (1995). The Valencia symposia edited by Bernardo *et al.* (1980–2011) and the ‘case studies’ edited by Gatsonis *et al.* (1993–2002) contain a wealth of material. A comparison of Bayesian and other approaches to statistical inference is provided by Barnett (1982). Nice recent textbook treatments at a lower level than this book can be found in Berry (1996) and Bolstad (2007).

A very nice book giving a treatment of Bayesian methods of great interest both to the layman and to the specialist is McGrayne (2011).

# References

- Abramowitz, M., and Stegun, M. A., *Handbook of Mathematical Functions*, Washington, DC: National Bureau of Standards (1964); New York: Dover (1965).
- Aitken, C. G. G. Lies, damned lies and expert witnesses, *Mathematics Today: Bull. Inst. Math. Appl.*, **32** (5/6) (1996), 76–80.
- Aitken, C. G.G., and Taroni, F., *Statistics and the Evaluation of Evidence for Forensic Scientists*, New York: John Wiley & Sons (2004) [1st edn by Aitken alone (1995)].
- Albert, J. H., *Bayesian Computation Using Minitab*, Belmont, CA: Duxbury (1994).
- Albert, J. H., *Bayesian Computation with R* (2nd edn), New York: Springer-Verlag 2009 [1st edn 2007].
- Altham, P. M. E., Exact Bayesian analysis of a  $2 \times 2$  contingency table and Fisher's 'exact' significance test, *J. Roy. Statist. Soc. Ser. B*, **31** (1969), 261–269.
- Arbuthnot, J., An argument for Divine Providence taken from the constant Regularity of the Births of Both Sexes, *Phil. Trans. Roy. Soc. London*, **23** (1710), 186–190 [reprinted in Kendall and Plackett (1977)].
- Armitage, P., Berry, G., and Matthews, J. N. S., *Statistical Methods in Medical Research* (4th edn), Oxford: Blackwells (2001) [1st edn, by Armitage alone (1971); 2nd edn (1987) and 3rd edn (1994) by Armitage and Berry (1987)].
- Arnold, B. C., *Pareto Distributions*, Fairland, MD: International Co-operative Publishing House (1983).
- Aykaç, A., and Brumat, C. (eds), *New Methods in the Applications of Bayesian Methods*, Amsterdam: North-Holland (1977).
- Balding, D. J., and Donnelly, P., Inference in forensic identification, *J. Roy. Statist. Soc. Ser. A*, **158** (1995), 21–53.
- Baird, R. D., *Experimentation: An Introduction to Measurement Theory and Experiment Design*, Englewood Cliffs, MD: Prentice-Hall (1962).
- Balakrishnan, N., Kotz, S., and Johnson, N. L., *Continuous Multivariate Distributions: Models and Applications*, New York: John Wiley & Sons (2012)

[previous edition by Johnson and Kotz alone (1972); this book overlaps with Fang, Kotz and Wang (1989)].

Barnard, G. A., Thomas Bayes's essay towards solving a problem in the doctrine of chances, *Biometrika*, **45** (1958), 293–315 [reprinted in Pearson and Kendall (1970)].

Barnett, V., Comparative Statistical Inference (3rd edn), New York: John Wiley & Sons (1999) [1st edn (1973), 2nd edn (1982)].

Barnett, V. D., Evaluation of the maximum-likelihood estimator where the likelihood equation has multiple roots, *Biometrika*, **53** (1966), 151–165.

Bartlett, M. S., The information available in small samples, *Proc. Cambridge Philos. Soc.*, **32** (1936), 560–566.

Bartlett, M. S., A comment on D. V. Lindley's statistical paradox, *Biometrika*, **44** (1957), 533–534.

Batschelet, E., Circular Statistics in Biology, London: Academic Press (1981).

Bayes, T. R., An essay towards solving a problem in the doctrine of chances, *Phil. Trans. Roy. Soc. London*, **53** (1763), 370–418 [reprinted as part of Barnard (1958) and Pearson and Kendall (1970)]; see also Price (1764).

Beaumont, M. A., Zhang, W., and Balding, D. J., Approximate Bayesian computation in population genetics, *Genetics*, **162** (2002), 2025–2035.

Behrens, W. A., Ein Beitrag zur Fehlensberechnung bei wenigen Beobachtungen, *Landwirtschaftliche Jahrbücher*, **68** (1929), 807–837.

Bellhouse, D. R., The Reverend Thomas Bayes, FRS: A biography to celebrate the tercentenary of his birth (with discussion), *Statistical Science*, **19** (2004), 3–43.

Bellhouse, D. R. *et al.*, Notes about the Rev. Thomas Bayes, *Bull. Inst. Math. Stat.*, **17** (1988), 49, 276–278, 482–483; **19** (1990), 478–479; **20** (1991), 226; **21** (1992), 225–227.

Benford, F., The law of anomalous numbers, *Proc. Amer. Philos. Soc.*, **78** (1938), 551–572.

Berger, J. O., A robust generalized Bayes estimator and confidence region for a multivariate normal mean, *Ann. Statist.*, **8** (1980), 716–761.

Berger, J. O., Statistical Decision Theory and Bayesian Analysis (2nd edn), Berlin: Springer-Verlag (1985) [1st edn published as *Statistical Decision Theory: Foundations, Concepts and Methods*, Berlin: Springer-Verlag (1980)].

- Berger, J. O., and Delampady, M., Testing precise hypotheses (with discussion), *Statistical Science*, **2** (1987), 317–352.
- Berger, J. O., and Wolpert, R. L., *The Likelihood Principle* (2nd edn), Hayward, CA: Institute of Mathematical Statistics (1988) [1st edn (1984)].
- Bernardo, J. M., Reference posterior distributions for Bayesian inference (with discussion), *J. Roy. Statist. Soc. Ser. B*, **41** (1979), 113–147.
- Bernardo, J. M., Bayarri, M. J., Berger, J. O., Dawid, A. P., Heckermann, D., Smith, A. F. M., and West, M. (eds), *Bayesian Statistics 8*, Oxford: Oxford University Press (2007).
- Bernardo, J. M., Bayarri, M. J., Berger, J. O., Dawid, A. P., Heckermann, D., Smith, A. F. M., and West, M. (eds), *Bayesian Statistics 9*, Oxford: Oxford University Press (2011).
- Bernardo, J. M., Berger, J. M., Dawid, A. P., Smith, A. F. M., and DeGroot, M. H. (eds), *Bayesian Statistics 4*, Oxford: Oxford University Press (1992).
- Bernardo, J. M., Berger, J. M., Dawid, A. P., and Smith, A. F. M. (eds), *Bayesian Statistics 5*, Oxford: Oxford University Press (1996).
- Bernardo, J. M., Berger, J. O., Dawid, A. P., and Smith, A. F. M. (eds), *Bayesian Statistics 6*, Oxford: Oxford University Press (1999).
- Bernardo, J. M., Dawid, A. P., Berger, J. O., West, M., Heckermann, D., and Bayarri, M. J. (eds), *Bayesian Statistics 7*, Oxford: Oxford University Press (2003).
- Bernardo, J. M., DeGroot, M. H., Lindley, D. V., and Smith, A. F. M. (eds), *Bayesian Statistics*, Valencia: Valencia University Press (1980).
- Bernardo, J. M., DeGroot, M. H., Lindley, D. V., and Smith, A. F. M. (eds), *Bayesian Statistics 2*, Amsterdam: North-Holland and Valencia: Valencia University Press (1985).
- Bernardo, J. M., DeGroot, M. H., Lindley, D. V., and Smith, A. F. M. (eds), *Bayesian Statistics 3*, Oxford: Oxford University Press (1988).
- Bernardo, J. M., and Smith, A. F. M., *Bayesian Theory*, New York, NY: John Wiley & Sons (1994).
- Berry, D. A., *Statistics: A Bayesian Perspective*, Belmont, CA: Duxbury (1996).
- Besag, J., On the statistical analysis of dirty pictures (with discussion), *J. Roy. Statist. Soc. Ser. B*, **48** (1986), 259–302.
- Birnbaum, A., On the foundations of statistical inference (with discussion), *J.*

- Amer. Statist. Assoc.*, **57** (1962), 269–306.
- Bliss, C. I., The dosage of the dosage-mortality curve, *Annals of Applied Biology*, **22** (1935), 134–167.
- Blum, M. G. B., and François, O., Non-linear regression models for Approximate Bayesian Computation, *Statistics and Computing* **20** (2010), 63–73.
- Bolstad, W. M., Introduction to Bayesian Statistics (2nd edn), Chichester: John Wiley & Sons (2007) [1st edn (2004)].
- Bortkiewicz, L. von, Das Gesetz der Kleinen Zahlenen, Leipzig: Teubner (1898).
- Box, G. E. P., Hunter, W. G., and Hunter, J. S., Statistics for Experimenters, New York: John Wiley & Sons (1978).
- Box, G. E. P., and Tiao, G. C., Bayesian Inference in Statistical Analysis, New York: John Wiley & Sons (1992) [1st edn (1973)].
- Breiman, L., Probability, Reading, MA: Addison-Wesley (1968).
- British Association for the Advancement of Science, Mathematical Tables, Vol. VI: Bessel Functions, Part I, Functions of Order Zero and Unity, Cambridge: Cambridge University Press (1937).
- Broemling, L. D., Bayesian Analysis of Linear Models, Basel: Marcel Dekker (1985).
- Brooks S. P., Markov chain Monte Carlo method and its application, *The Statistician: J. Roy. Statist. Soc. Ser. D*, **47** (1998), 69–100.
- Brooks, S., Gelman, A., Jones, G. L., and Meng, X. L., Handbook of Markov Chain Monte Carlo. Boca Raton, FL: CRC Press (2011).
- Buck, C. E., Cavanagh, W. G., and Litton, C. D., Bayesian Approach to Interpreting Archaeological Data, New York: John Wiley & Sons (1996).
- Calvin, T. W., How and When to Perform Bayesian Acceptance Sampling (ASQC Basic References in Quality Control: Statistical Techniques, Volume 7), Milwaukee, WI: American Society for Quality Control (1984).
- Carlin, B. P., Gelfand, A. E., and Smith, A. F. M., Hierarchical Bayesian analysis of changepoint problems, *Applied Statistics*, **41** (1992), 389–405.
- Carlin, B. P., and Louis, T. A., Bayes and Empirical Bayes Methods for Data Analysis (3rd edn), London: Chapman and Hall (2008) [1st edn (1994), 2nd edn (2000)].

- Casella, G., and George, E., Explaining the Gibbs sampler, *American Statistician*, **46** (1992), 167–174.
- Chatterjee, S. J., Statistical Thought: A Perspective and History, Oxford: Oxford University Press (2003).
- Chen, M.-H., and Shao, Q.-M., Monte Carlo estimation of Bayesian credible intervals and HPD intervals, *Journal of Computational and Graphical Statistics*, **8** (1998), 69–92.
- Chib, S., and Greenberg, E., Bayes inference for regression models with ARMA errors, *Journal of Econometrics*, **64** (1994), 183–206.
- Chib, S., and Greenberg, E., Understanding the Metropolis-Hastings Algorithm, *American Statistician*, **49** (1995), 327–335.
- Cochran, W. G., and Cox, G. M., Experimental Designs (2nd edn), New York: John Wiley & Sons (1957) [1st edn (1950)].
- Congdon, P., Bayesian Statistical Modelling, New York: John Wiley & Sons (2002).
- Congdon, P., Bayesian Models for Categorical Data, New York: John Wiley & Sons (2005).
- Congdon, P., Applied Bayesian Modelling (2nd edn), New York: John Wiley & Sons (2006) [1st edn (2003)].
- Congdon, P., Applied Bayesian Hierarchical Modelling, New York: John Wiley & Sons (2010).
- Corduneanu, A., and Bishop, C. M., Variational Bayesian model selection for mixture distributions, in Jaakola, T., and Richardson, T. (eds), Artificial Intelligence and Statistics, San Mateo, CA: Morgan Kaufmann (2001), pp. 27–34.
- Cornish, E. A., The multivariate t-distribution associated with a set of normal sample deviates, *Austral. J. Phys.*, **7** (1954), 531–542.
- Cornish, E. A., The sampling distribution of statistics derived from the multivariate t-distribution, *Austral. J. Phys.*, **8** (1955), 193–199.
- Cornish, E. A., Published Papers of E. A. Cornish, Adelaide: E. A. Cornish Memorial Appeal, Adelaide (1974).
- Cowles, M. K., and Carlin, B. P., Markov chain Monte Carlo convergence diagnostics: a comparative review, *J. Amer. Statist. Assoc.*, **91** (1996), 883–904.
- Dalal, S. R., Nonparametric Bayes decision theory (with discussion), in

Bernardo *et al.* (1980).

Dale, A., A History of Inverse Probability from Thomas Bayes to Karl Pearson, Berlin: Springer-Verlag (1999) [1st edn (1991)].

Dale, A., Most Honourable Remembrance: The Life and Work of Thomas Bayes, Berlin: Springer-Verlag (2003).

Dalgaard, P., Introductory Statistics with R (2nd edn), Berlin: Springer-Verlag (2008) [1st edn (2002)].

Dalziel, C. F., Lagen, J. B., and Thurston, J. L., Electric shocks, *Trans. IEEE*, **60** (1941), 1073–1079.

David, F. N. Tables of the Correlation Coefficient, Cambridge: Cambridge University Press for Biometrika (1954).

Davis, P. J., and Rabinowitz, P., Methods of Numerical Integration (2nd edn), Orlando, FL: Academic Press (1984) [1st edn (1975)].

Dawid, A. P., The island problem: coherent use of identification evidence, in Freeman and Smith (1994).

Deely, J., and Lindley, D. V., Bayes empirical Bayes, *J. Amer. Statist. Assoc.*, **76** (1981), 833–841.

DeGroot, M. H., Optimal Statistical Decisions, New York: McGraw-Hill (1970).

DeGroot, M. H., Fienberg, S. E., and Kadane, J. B. (eds), Statistics and the Law, New York: John Wiley & Sons (1986).

Dempster, A. P., Laird, N. M., and Rubin, D. B., Maximum likelihood from incomplete data via the *EM* algorithm (with Discussion), *J. Roy. Statist. Soc. Ser. B*, **39** (1977), 1–38.

Dey, D. K., Ghosh, S. K., and Mallick, B. K., Generalized Linear Models: A Bayesian Perspective, Basel: Marcel Dekker (2000).

Di Raimondo, F., *In vitro* and *in vivo* antagonism between vitamins and antibiotics, *Int. Rev. Vitamin Res.*, **23** (1951), 1–12.

Ann. Statist., **7** (1979), 269–281.

Diaconis, P., and Ylvisaker, D., Quantifying prior opinion, in Bernardo *et al.* (1985).

Dobson, A. J., and Barnett, A., An Introduction to Generalized Linear Models (3rd edn), London: Chapman and Hall 2008 [1st edn (1990) and 2nd edn (2002) by Dobson alone].

- Dunnett, C. W., and Sobel, M., A bivariate generalization of Student's t distribution with tables for special cases, *Biometrika*, **41** (1954), 153–176.
- Edwards, A. W. F., The measure of association in a  $2 \times 2$  table, *J. Roy. Statist. Soc. Ser. A*, **126** (1963), 109–113.
- Edwards, A. W. F., Likelihood, Cambridge: Cambridge University Press (1992) [1st edn (1972)].
- Edwards, A. W. F., Bayes, Rev. Thomas, in Dictionary of National Biography: *Missing Persons*, Oxford: Oxford University Press (1993).
- Edwards, A. W. F., Bayes, Rev. Thomas, in Oxford Dictionary of National Biography, Oxford: Oxford University Press (2004).
- Edwards, J., A Treatise on the Integral Calculus, London: Macmillan (1921) [reprinted New York: Chelsea (1955)].
- Edwards, W., Lindman, H., and Savage, L. J., Bayesian statistical inference for psychological research, *Psychological Review*, **70** (1963), 193–242 [reprinted in Luce *et al.* (1965), Kadane (1984), Savage (1981) and Polson and Tiao (1995, Volume I)].
- Efron, B., and Morris, C., Data analysis using Stein's estimator and its generalisations, *J. Amer. Statist. Assoc.*, **70** (1975), 311–319.
- Efron, B., and Morris, C., Stein's paradox in statistics, *Scientific American*, **236** (1977 May), 119–127, 148.
- Eisenhart, C., Hastay, M. W., and Wallis, W. A. (eds), (Selected) Techniques of Statistical Analysis by the Statistical Research Group of Columbia University, New York: McGraw-Hill (1947).
- Evans, G., Practical Numerical Integration, New York: John Wiley & Sons (1993).
- Evans, M., Hastings, N., and Peacock, B., Statistical Distributions, New York: John Wiley & Sons (1993) [1st edn by Hastings and Peacock only (1974)].
- Evans, M., and Swartz, T., Methods for approximating integrals in statistics with special emphasis on Bayesian integration problems (with Discussion), *Statist. Sci.*, **10** (1995), 254–272 and **11** (1996), 54–64.
- Evans, M., and Swartz, T., Approximating Integrals via Monte Carlo and Deterministic Methods, Oxford: Oxford University Press (2000).
- Fan, Y., and Sisson, S. A., Reversible jump MCMC, Chapter 3 in Brooks *et al.* (2011).

- Fang, K.-T., Kotz, S., and Wang, K. W., *Symmetric Multivariate and Related Distributions*, London: Chapman and Hall (1989).
- Fearnhead, P., and Prangle, D., Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation, *J. Roy. Statist. Soc. Ser. B*, **74** (2012), 1–28.
- Feller, W., *An Introduction to Probability Theory and its Applications*, New York: John Wiley & Sons (Vol. 1 1950, 1957, 1968; Vol. 2 1966, 1971).
- Ferguson, T. S., *Mathematical Statistics: A Decision Theoretic Approach*, New York: Academic Press (1967).
- Ferguson, T. S., A Bayesian analysis of some nonparametric problems, *Ann. Statist.*, **1** (1973), 209–230.
- Fienberg, S. E. (ed.), *The Evolving Role of Statistical Assessments as Evidence in the Court*, New York: Springer-Verlag (1989).
- de Finetti, B., La prévision: ses lois logiques, ses sources subjectives, *Ann. Inst. H. Poincaré*, **7** (1937), 86–133 [translated by H. E. Kyberg, Jr, as de Finetti (1964)].
- de Finetti, B., Foresight: its logical laws, its subjective sources, in Kyburg and Smokler (1964) [reprinted in Polson and Tiao (1995, Volume I) and in Kotz and Johnson (1992–1997, Volume I)].
- de Finetti, B., *Probability, Induction and Statistics: The Art of Guessing*, New York: John Wiley & Sons (1972).
- de Finetti, B., *Theory of Probability: A critical introductory treatment* (2 vols.), New York: John Wiley & Sons (1974–1975).
- Fisher, R. A., Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population, *Biometrika*, **10** (1915), 507–521.
- Fisher, R. A., On the ‘probable error’ of a coefficient of correlation deduced from a small sample, *Metron*, **1** (1921), 3–32.
- Fisher, R. A., On the mathematical foundations of theoretical statistics, *Phil. Trans. Roy. Soc. London Ser. A*, **222** (1922), 309–368 [reprinted in Kotz and Johnson (1992–1997, Volume I)].
- Fisher, R. A., On a distribution yielding the error function of several well known statistics, Proc. Internat. Congress of Math., Toronto: Toronto University Press (1924), vol. 2, pp. 805–813.

- Fisher, R. A., Theory of statistical information, *Proc. Cambridge Philos. Soc.*, **22** (1925a), 700–725.
- Fisher, R. A., Statistical Methods for Research Workers, Edinburgh: Oliver & Boyd (1925b) (many subsequent editions).
- Fisher, R. A., The fiducial argument in statistical inference, *Ann. Eugenics*, **6** (1935), 391–398.
- Fisher, R. A., Has Mendel's work been rediscovered?, *Ann. Sci.*, **1** (1936), 115–137.
- Fisher, R. A., On a point raised by M. S. Bartlett in fiducial probability, *Ann. Eugenics*, **7** (1937), 370–375.
- Fisher, R. A., The comparison of samples with possibly unequal variances, *Ann. Eugenics*, **9** (1939), 174–180.
- Fisher, R. A., The analysis of variance with various binomial transformations, *Biometrics* **10** (1954), 130–139.
- Fisher, R. A., Statistical Methods and Scientific Inference (2nd edn), Edinburgh: Oliver & Boyd (1959) [1st edn (1956)].
- Fisher, R. A., Collected Papers of R.A. Fisher (5 vols), edited by J. H. Bennett, Adelaide: University of Adelaide Press (1971–1974).
- Florens, J. P., Mouchart, M., Raoult, J. P., Simar, L., and Smith, A. F. M., Specifying Statistical Models from Parametric to Nonparametric Using Bayesian or Non-Bayesian Approaches (Lecture Notes in Statistics No. 16), Berlin: Springer-Verlag (1983).
- Foreman, L. A., Smith, A. F. M., and Evett, I.W., Bayesian analysis of DNA profiling data in forensic identification applications, *J. Roy. Statist. Soc. Ser. A*, **160** (1997), 429–469.
- Fox, J., An R and S-Plus Companion to Applied Regression, Thousand Oaks, CA: Sage (2002).
- Freeman, P. R., and Smith, A. F. M., Aspects of Uncertainty: A Tribute to D. V. Lindley, New York: John Wiley & Sons (1994).
- French, S., and Smith, J. Q., The Practice of Bayesian Statistics, London: Arnold (1997).
- Gamerman, D., and Lopes, H. F., Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference (2nd edn), Boca Raton, FL: Chapman and Hall/CRC (2006) [1st edn by Gamerman alone (1997)].

Gastwirth, J. L., Statistical Reasoning in Law and Public Policy (2 vols), Boston, MA: Academic Press (1988).

Gatsonis, C., Hodges, J. S., Kass, R. E., and Singpurwalla, N. D. (eds), Case Studies in Bayesian Statistics (Lecture Notes in Statistics, No. 83), Berlin: Springer-Verlag (1993).

Gatsonis, C., Hodges, J. S., Kass, R. E., and Singpurwalla, N. D. (eds), Case Studies in Bayesian Statistics, Volume II (Lecture Notes in Statistics, No. 105), Berlin: Springer-Verlag (1995).

Gatsonis, C., Hodges, J. S., Kass, R. E., McCulloch, R. E., Rossi, P., and Singpurwalla, N. D. (eds), Case Studies in Bayesian Statistics, Volume III (Lecture Notes in Statistics, No. 121), Berlin: Springer-Verlag (1997).

Gatsonis, C., Kass, R. E., Carlin, B., Carriquiry, A., Gelman, A., Verdinelli, I., and West, M. (eds), Case Studies in Bayesian Statistics, Volume IV (Lecture Notes in Statistics, No. 140), Berlin: Springer-Verlag (1999).

Gatsonis, C., Kass, R. E., Carlin, B., Carriquiry, A., Gelman, A., Verdinelli, I., and West, M. (eds), Case Studies in Bayesian Statistics, Volume V (Lecture Notes in Statistics, No. 162), Berlin: Springer-Verlag (2002a).

Gatsonis, C., Kass, R. E., Carriquiry, A., Gelman, A., Higdon, D., Pauder, D. K., and Verdinelli, I. (eds), Case Studies in Bayesian Statistics, Volume VI (Lecture Notes in Statistics, No. 167), Berlin: Springer-Verlag (2002b).

Gaver D., and O'Muircheartaigh I., Robust empirical Bayes analysis of event rates, *Technometrics*, **29** (1) (1987), 1–15.

Gelfand, A. E., Hills, S. E., Racine-Poon, A., and Smith, A. F. M., Illustration of Bayesian inference in normal data models using Gibbs sampling, *J. Amer. Statist. Soc.*, **85** (1990), 972–985.

Gelfand, A. E., and Smith, A. F. M., Sampling-based approaches to calculating marginal densities, *J. Amer. Statist. Assoc.*, **85** (1990), 398–409 [reprinted in Polson and Tiao (1995, Volume II) and in Kotz and Johnson (1992–1997, Volume III)].

Gelfand, I. M., and Fomin, S. V., Calculus of Variations, London: Prentice-Hall (1963).

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B., Bayesian Data Analysis (2nd edn), London: Chapman and Hall (2004) [1st edn (1995)].

Gelman, A., and Rubin, D. B., Inference from iterated simulation using multiple sequences (with discussion), *Statistical Science*, **7** (1992), 457–511.

- Geman, S., and Geman, D., Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images, *Trans. Pattern Analysis and Machine Intelligence*, **6** (1984), 721–742 [reprinted in Polson and Tiao (1995, Volume II) and in Kotz and Johnson (1992–1997, Volume III)].
- Ghosh, J. K., and R.V. Ramamoorthi, R. V., Bayesian Nonparametrics, New York: Springer-Verlag (2003).
- Gilks, W. R., Full conditional distributions, Chapter 5, in Gilks *et al.*, Markov Chain Monte Carlo in Practice, London: Chapman and Hall (1996).
- Gilks, W. R., Clayton, D. G., Spiegelhalter, D. J., Best, N. G., McNeil, A. J., Sharples, L. D., and Kirby, A. J., Modelling complexity: applications of Gibbs sampling in medicine, *J. Roy. Statist. Soc. Ser. B*, **55** (1993), 39–52.
- Gilks, W. R., Richardson, S., and Spiegelhalter, D. J., Markov Chain Monte Carlo in Practice, London: Chapman and Hall (1996).
- Gilks, W. R., and Wild, P., Adaptive rejection sampling for Gibbs sampling, *Applied Statistics*, **41** (1992), 337–348.
- Gill, J., Bayesian Methods: A Social and Behavioural Sciences Approach (2nd edn), Boca Raton, FL: Chapman and Hall/CRC (2007) [1st edn (2002)].
- Godambe, V. P., and Sprott, D. A. (eds), Foundations of Statistical Inference: A Symposium, Toronto: Holt, Rinehart and Winston (1971).
- Goldstein, M., and Wooff, D., Bayes Linear Statistics, Chichester: John Wiley & Sons (2007).
- Good, I. J., Probability and the Weighing of Evidence, London: Griffin (1950).
- Good, I. J., The Estimation of Probabilities: An Essay on Modern Bayesian Methods, Cambridge, MA: MIT Press (1965).
- Good, I. J., Some history of the hierarchical Bayesian methodology, in Bernardo *et al.* (1980) [reprinted as Chapter 9 of Good (1983)].
- Good, I. J., Good Thinking: The Foundations of Probability and its Applications, Minneapolis, MN: University of Minnesota Press (1983).
- Green P. J., Reversible jump Markov chain Monte Carlo computation and Bayesian model determination, *Biometrika*, **82** (4) (1995), 711–732.
- Hald, A., A History of Probability and Statistics and their applications before 1750, New York: John Wiley & Sons (1986).
- Hald, A., A History of Mathematical Statistics from 1750 to 1930, New York: John Wiley & Sons (1998).

- Hald, A., A history of parametric statistical inference from Bernoulli to Fisher, 1713–1935, New York: Springer (2007).
- Haldane, J. B. S., A note on inverse probability, *Proc. Cambridge Philos. Soc.*, **28** (1931), 55–61.
- Hardy, G. H., Littlewood, J. E., and Pólya, G., Inequalities, Cambridge: Cambridge University Press (1952) [1st edn (1934)].
- Härdle, W., Smoothing Techniques with Implementation in S, New York: Springer (1991).
- Harter, H. L., The method of least squares and some alternatives, *Internat. Statist. Review*, **42** (1974), 147–174, 235–264, **43** (1975), 1–44, 125–190, 269–278, and **44** (1976), 113–159.
- Hartigan, J. A., Bayes Theory, Berlin: Springer-Verlag (1983).
- Hastings, W. K., Monte Carlo sampling methods using Markov chains and their applications, *Biometrika* **57** (1970), 97–109 [reprinted in Kotz and Johnson (1992–1997, Volume III)].
- Hill, B. M., On statistical paradoxes and non-conglomerability, in Bernardo *et al.* (1980).
- Hill, B. M., De Finetti's theorem, induction and, or Bayesian nonparametric predictive inference, in Bernardo *et al.* (1988)
- Hjort, N. L., Holmes, C., Müller, P., and Walker, S. G. (eds), Bayesian Nonparametrics, Cambridge: Cambridge University Press (2010).
- Hoerl, A. E., and Kennard, R. W., Ridge regression: biased estimation for nonorthogonal problems, *Techonometrics*, **12** (1970), 55–67, and Ridge regression: applications to nonorthogonal problems, *ibid.*, 69–82.
- Hoerl, A. E., and Kennard, R. W., Ridge regression. In S. Kotz, N. L. Johnson and C. B. Read (eds), Encyclopedia of Statistical Sciences, New York: John Wiley & Sons (1988).
- Holland, G. D., The Reverend Thomas Bayes, F.R.S. (1702–1761), *J. Roy. Statist. Soc. Ser. A*, **125** (1962), 451–461.
- Horn, R. A., and Johnson, C. A., Matrix Analysis, Cambridge: Cambridge University Press (1985).
- Horn, R. A., and Johnson, C. A., Topics in Matrix Analysis, Cambridge: Cambridge University Press (1991).
- Hosmer, D. W., and Lemeshow, S., Applied Logistic Regression (2nd edn), New

- York: John Wiley & Sons (2000) [1st edn (1989)].
- Huzurbazar, V. S., *Sufficient Statistics*, New York: Marcel Dekker (1976).
- Isaacs, G. L., Christ, D. E., Novick, M. R., and Jackson, P. H., *Tables for Bayesian Statisticians*, Ames, IO: Iowa University Press (1974).
- Jackson, P. H., *Formulae for generating highest density credibility regions*, ACT Technical Report No. 20, Iowa City, IO: American College Testing Program (1974).
- James, W., and Stein, C., *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, CA, and Los Angeles, CA: University of California Press (1961), Volume I, pp. 311–319 [reprinted in Kotz and Johnson (1992–1997, Volume I)].
- Jeffreys, H. S., *Theory of Probability* (3rd edn), Oxford: Oxford University Press (1961) [1st edn (1939), 2nd edn (1948)].
- Jensen, F. V., *An Introduction to Bayesian Networks*, London: Taylor and Francis (1996).
- Jensen, F. V., and Nielson, T. D., *Bayesian Networks and Decision Graphs* (2nd edn), New York: Springer (2010) [1st edn (2001)].
- Johnson, N. L., Kemp, A.W., and Kotz, S., *Univariate Discrete Distributions* (3rd edn), New York: John Wiley & Sons (2005) [1st edn published as *Discrete Distributions*, New York: Houghton-Mifflin (1969); 2nd edn (1992)].
- Johnson, N. L., Kotz, S., and Balakrishnan, N., *Continuous Univariate Distributions* (2 vols), New York: John Wiley & Sons (1994–1995) [1st edn published New York: Houghton-Mifflin (1970–1971)].
- Kadane, J. B. (ed.), *Robustness of Bayesian Analyses*, Amsterdam: North-Holland (1984).
- Kale, B. K., On the solution of the likelihood equation by iteration processes, *Biometrika*, **48** (1961), 452–456.
- Kelley, T. L., *Interpretation of Educational Measurements*, Yonkers-on-Hudson, NY: World Book Co. (1927).
- Kendall, M. G., and Plackett, R. L. (eds), *Studies in the History of Probability and Statistics*, Vol. II, London: Griffin (1977).
- Kendall, M. G., Stuart, A., and Ord, J. K., *The Advanced Theory of Statistics*, Vol. I (5th ed.), London: Griffin (1987).
- Kennedy, W. G., and Gentle, J. E., *Statistical Computing*, New York: Marcel

- Dekker (1980).
- Kennett, P., and Ross, C. A., *Geochronology*, London: Longmans (1983).
- Kleinbaum, D. G., *Logistic Regression: A Self-Learning Text*, Berlin: Springer-Verlag (1994).
- Knuth, D. E., *The Art of Computer Programming*, Vol. 2: *Seminumerical Algorithms*, Reading, MA: Addison-Wesley (1981) [1st edn (1969)].
- Kotz, S., and Johnson, N. L. (eds), *Breakthroughs in Statistics* (3 vols), Berlin: Springer-Verlag (1992–1997).
- Kotz, S., and Nadarajan, S., *Multivariate t Distributions*, Cambridge: Cambridge University Press (2004).
- Kotz, S., Read, C. B., Balakrishnan, N, and Vidakovic, B. (eds), *Encyclopedia of Statistical Science* (2nd edn in 16 vols), New York: Wiley-Interscience (2006) [1st edn by Kotz, S., and Johnson, N. L.(eds) (1982–1989)]
- Krause, A., and Olsen, M, *The Basics of S and S-PLUS*, Berlin: Springer-Verlag (2000) [1st edn (1997)].
- Kullback, S., *Information Theory and Statistics*, New York: John Wiley & Sons (1959); New York: Dover (1968).
- Kullback, S., and Leibler, R. A., On information and sufficiency, *Ann. Math. Statist.*, **22** (1951), 79–86.
- Kyburg, H. E., and Smokler, H. E. (eds), *Studies in Subjective Probability*, New York: John Wiley & Sons (1964) [2nd edn (much altered); Melbourne, FA: Krieger (1980)].
- Laplace, P. S., Mémoire sur la probabilité des causes par les évenemens *Mém. de math. et phys. présenté à l'Acad. roy. des sci.*, **6** (1774), 621–686 [reprinted in his *Oeuvres complètes*, **8**, 27–65. An English translation is to be found in Stigler (1986b)].
- Laplace, P. S., *Théorie Analytiques des Probabilités*, Paris: Courcier (1812) [reprinted Brussels: Culture et Civilisation (1967); subsequent edn in 1814 and 1820].
- Lee, P. M., Not so spurious, *Mathematical Gazette*, **75** (1991), 200–201.
- Lehmann, E. L., *Theory of Point Estimation*, New York: John Wiley & Sons (1983).
- Lehmann, E. L., *Testing Statistical Hypotheses* (2nd edn), New York: John Wiley & Sons (1986) [1st edn (1959)].

- Lenk, P. J., Towards a practicable Bayesian nonparametric density estimator, *Biometrika*, **78** (1991), 531–543.
- Leonard, T., and Hsu, J. S. J., Bayesian Methods: An Introduction for Statisticians and Interdisciplinary Researchers, Cambridge: Cambridge University Press (2001).
- Lieberman, G. J., and Owen, D. B., Tables of the Hypergeometric Probability Distribution, Stanford, CA: Stanford University Press (1961).
- Lindgren, B. W., Statistical Theory (4th edn), London: Chapman and Hall (1993) [1st edn (1960), 2nd edn (1962), 3rd edn (1968)].
- Lindley, D. V., On a measure of the information provided by an experiment, *Ann. Math. Statist.*, **27** (1956), 936–1005.
- Lindley, D. V., A statistical paradox, *Biometrika*, **44** (1957), 187–192.
- Lindley, D. V., Introduction to Probability and Statistics from a Bayesian Viewpoint (2 vols—Part I: Probability and Part II: Inference), Cambridge: Cambridge University Press (1965).
- Lindley, D. V., Bayesian least squares, *Bull. Inst. Internat. Statist.*, **43** (2) (1969), 152–153.
- Lindley, D. V., Bayesian Statistics: A Review, Philadelphia, PA: S.I.A.M.—Society for Industrial and Applied Mathematics (1971a).
- Lindley, D. V., The estimation of many parameters (with discussion), in Godambe and Sprott (eds) (1971b).
- Lindley, D. V., A problem in forensic science, *Biometrika*, **64** (1977), 207–213.
- Lindley, D. V., and Scott, W. F., New Cambridge Elementary Statistical Tables, Cambridge: Cambridge University Press (1995) [1st edn (1984)].
- Lindley, D. V., and Smith, A. F. M., Bayes estimates for the linear model (with discussion), *J. Roy. Statist. Soc. Ser. B*, **34** (1972), 1–41 [reprinted in Polson and Tiao (1995, Volume II) and in Kotz and Johnson (1992–1997, Volume III)].
- Liu, Y. S., Peskun's theorem and a modified discrete-state Gibbs sampler, *Biometrika*, **83** (1996), 681–682.
- Liu, Y. S., Markov Chain Strategies in Scientific Computing, Berlin: Springer-Verlag 2001.
- Luce, R. D., Bush, R. B., and Galanter, E., Readings in Mathematical Psychology, Vol. 2, New York: John Wiley & Sons (1965).
- Mardia, K. V., Statistics of Directional Data, New York: Academic Press (1972).

Mardia, K. V., and Jupp, P. E., *Directional Statistics*, New York: John Wiley & Sons (2001).

Marin, J. M., and Robert, C. P., *Bayesian Core: A practical Approach to Computational Bayesian Statistics*, New York: Springer-Verlag (2007).

Maritz, J. S., and Lwin, T., *Empirical Bayes Methods* (2nd edn), London: Methuen (1989) [1st edn by Maritz alone (1970)].

McCullagh, P., and Nelder, J. A., *Generalized Linear Models* (2nd edn), London: Chapman and Hall (1989) [1st edn (1984)].

McGrayne, S. B., *The Theory That Would Not Die: How Bayes' Rule Cracked the Enigma Code, Hunted Down Russian Submarines, and Emerged Triumphant from Two Centuries of Controversy*, New Haven, CT, and London: Yale University Press (2011).

Mendel, G., Versuche über Pflanzen-Hybriden, *Verhandlungen des naturforschenden Vereines in Bürnn*, **4** (1865), 3–47 [translation by E. R. Sherwood in Stern and Sherwood (1966)].

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E., Equation of state calculations by fast computing machines, *J. Chem. Phys.*, **21** (1953), 1087–1092 [reprinted in Kotz and Johnson (1992–1997, Volume III)].

Meyer, D. L., and Collier, R. O. (eds), *Bayesian Statistics*, Itasca, IL: F. E. Peacock (1970).

Meyn, S. P., and Tweedie, R. L., *Markov Chains and Stochastic Stability*, Berlin: Springer-Verlag (1993).

Miller, K. S., *Some Eclectic Matrix Theory*, Malabar, FL: Krieger (1987).

Morris, C., Parametric empirical Bayes confidence sets: theory and applications, *J. Amer. Statist. Assoc.*, **78** (1983), 47–65.

Müller, P., and Quintana, F. A., Nonparametric Bayesian Data Analysis, *Statistical Science*, **19** (2004), 95–110.

Nagel, E., *Principles of the Theory of Probability*, Chicago: University of Chicago Press (1939) [reprinted in Neurath *et al.* (1955)].

Neapolitan, R. E., *Learning Bayesian networks*, Upper Saddle River, NJ: Pearson Prentice Hall (2004).

Neave, H. R., *Statistics Tables for Mathematicians, Engineers, Economists and the Behavioural and Management Sciences*, London: George Allen & Unwin

(1978).

Neurath, O., Carnap, R., and Morris, C. (eds), Foundations of the Unity of Science, Vol. I, Chicago: University of Chicago Press (1955).

Newcomb, S., Note on the frequency of use of the different digits in natural numbers, *Amer. J. Math.*, **4** (1881), 39–40 [reprinted in Stigler (1980)].

Norris, J. R., Markov Chains, Cambridge: Cambridge University Press (1997).

Novick, M. R., and Jackson, P. H., Statistical Methods for Educational and Psychological Research, New York: McGraw-Hill (1974).

Ntzoufras, I., Bayesian Methods Using WinBUGS, Hoboken, NJ: John Wiley & Sons (2009).

Odell, P. L., and Feiveson, A. H., A numerical procedure to generate a sample covariance matrix, *J. Amer. Statist. Assoc.*, **61** (1966), 198–203.

O'Hagan, A., and Forster, J., Kendall's Advanced Theory of Statistics: Volume 2B: Bayesian Inference, London: Arnold (2004) [1st edn by O'Hagan alone (1994)].

Ormerod, J. T., and Wand, M. P., Explaining Variational Approximations, *American Statistician*, **64** (2010), 140–153.

Ó Ruanaidh, J. J. K., and Fitzgerald, W. J., Numerical Bayesian Methods applied to Signal Processing, Berlin: Springer-Verlag (1996).

Patil, V. H., Approximations to the Behrens-Fisher distribution, *Biometrika*, **52** (1965), 267–271.

Pearson, E. S. (ed. Plackett, R. L. and Barnard, H. A.), ‘Student’: A Statistical Biography of William Sealy Gosset, Oxford: University Press (1990).

Pearson, E. S., and Hartley, H. O., Biometrika Tables for Statisticians (2 vols), Cambridge: Cambridge University Press for Biometrika (Vol. I—1954, 1958, 1966; Vol. II—1972).

Pearson, E. S., and Kendall, M. G., Studies in the History of Probability and Statistics, London: Griffin (1970).

Pearson, K., Tables of the Incomplete Gamma Function, Cambridge: Cambridge University Press (1922, 1924).

Pearson, K., Tables of the Incomplete Beta Function, Cambridge: Cambridge University Press (1934, 1968).

Peirce, C. S., The probability of induction, *Popular Science Monthly*, **12** (1878), 705–718 [reprinted in Peirce (1982) and as Arts. 2.669–2.693 of Peirce (1931–

1958]).

Peirce, C. S., Collected Papers, Cambridge, MA: Harvard University Press (1931–1958).

Peirce, C. S., Writings of C. S. Peirce: A Chronological Edition. Volume III: 1872–1878, Bloomington, IN: Indiana University Press (1982).

Pietronero, L., Tosatti, E., Tosatti, V., and Vespignani, A., Explaining the uneven distribution of numbers in nature: the laws of Benford and Zipf, *Physica A*, **293** (2001), 297–304.

Pole, A., West, M., and Harrison, P. J., Applied Bayesian Forecasting and Time Series Analysis, London: Chapman and Hall (1994).

Polson, N., and Tiao, G. C., Bayesian Inference (2 vols) (The International Library of Critical Writings in Econometrics, No. 7), Aldershot: Edward Elgar (1995).

Prentice, R. L., A generalization of the probit and logit model for dose response curves, *Biometrika*, **32** (1976), 761–768.

Press, S. J., Subjective and Objective Bayesian Statistics: Principles, Models and Applications (2nd edn), New York: John Wiley & Sons (2002) [1st edn published as *Bayesian Statistics: Principles, Models and Applications* (1989)].

Press, S. J., Applied Multivariate Analysis: Using Bayesian and Frequentist Measures of Inference (2nd edn), Melbourne, FL: Krieger (2009) [1st edn (1982); earlier version published as *Applied Multivariate Analysis*, New York: Holt, Rinehart, Winston (1972)].

Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P., Numerical Recipes: The Art of Scientific Computing, Cambridge: Cambridge University Press (1986) [further editions and example books in BASIC, FORTRAN, Pascal and C (1986–1993)].

Price, R., A demonstration of the second rule in the essay towards the solution of a problem in the doctrine of chances, *Phil. Trans. Roy. Soc. London*, **54** (1764), 296–325.

Raftery, A. E., and Lewis, S. M., Implementing MCMC, Chapter 7, in Gilks *et al.* (1996).

Raiffa, H., and Schlaifer, R., Applied Statistical Decision Theory, Cambridge, MA: Harvard University Press (1961).

Raimi, R. A., The first digit problem, *Amer. Math. Monthly*, **83** (1976), 531–538.

- Rauch, B., Götsche, M., Brähler, G., and Engel, S., Fact and Fiction in EU-Governmental Economic Data, *German Economic Review*, **12** (2011), 243–255.
- Rao, C. R., Linear Statistical Inference and its Applications (2nd edn), New York: John Wiley & Sons (1973) [1st edn (1965)].
- Rényi, A., Foundations of Probability, San Francisco, CA: Holden-Day (1970).
- Richardson, S., and Green, P. J., On Bayesian analysis of mixtures with an unknown number of components (with discussion), *J. Roy. Stat. Soc. Ser. B*, **59** (1997), 731–792.
- Ripley, B. D., Stochastic Simulation, New York: John Wiley & Sons (1987).
- Robert, C., Convergence control methods for Markov chain Monte Carlo algorithms, *Statist. Sci.*, **10** (1995), 231–253.
- Robert, C., and Casella, G., Introducing Monte Carlo Methods with R, New York: Springer-Verlag (2010).
- Roberts, H. V., Informative stopping rules and inference about population size, *J. Amer. Statist. Assoc.*, **62** (1967), 763–775.
- Robinson, G. K., Properties of Student's t and of the Behrens-Fisher solution to the two means problem, *Ann. Statist.*, **4** (1976), 963–971 and **10** (1982), 321.
- Rothschild, V., and Logothetis, N., Probability Distributions, New York: John Wiley & Sons (1986).
- Rubin, H., Robustness in generalized ridge regression and related topics', in Bernardo *et al.* (1988).
- Savage, L. J., The Foundations of Statistics, New York: John Wiley & Sons (1954); New York: Dover (1972).
- Savage, L. J., The Writings of Leonard Jimmie Savage: A Memorial Selection, Washington, DC: American Statistical Association/Institute of Mathematical Statistics (1981).
- Savage, L. J. *et al.*, The Foundations of Statistical Inference: A Discussion, London: Methuen (1962).
- Scheffé, H., The Analysis of Variance, New York: John Wiley & Sons (1959).
- Schlaifer, R., Introduction to Statistics for Business Decisions, New York: McGraw-Hill (1961).
- Schmitt, S. A., Measuring Uncertainty: An Elementary Introduction to Bayesian Statistics, Reading, MA: Addison-Wesley (1969).

- Seber, G. A. F., and Lee, A. J., *Linear Regression Analysis*, New York: John Wiley & Sons (2003) [1st edn (1977)].
- Shafer, G., Lindley's paradox, *J. Amer. Statist. Assoc.*, **77** (1982), 325–351.
- Shannon, C. E., The mathematical theory of communication, *Bell Syst. Tech. J.*, **27** (1948), 379–423 & 623–656.
- Shannon, C. E., and Weaver, W., *The mathematical theory of communication*, Urbana, IL: University of Illinois Press (1949).
- Silcock, A., *Verse and Worse*, London: Faber & Faber (1952).
- Sisson, S. A., and Fan, Y., Likelihood-free Markov chain Monte Carlo, Chapter 12, in Brooks *et al.* (1980).
- Sisson, S. A., Fan, Y., and Tanaka, M. M., Sequential Monte Carlo without likelihoods, *Proc. Nat. Acad. Sci.*, **104** (6) (2007), 1760–1765 and **106** (39) (2009), 16889.
- Smith, C. A. B., *Biomathematics: The principles of mathematics for students of biological and general science*, Volume 2—Numerical Methods, Matrices, Probability, Statistics (4th edn), London: Griffin (1969) (previous one volume edn 1923, 1935, 1954, the first two by W. M. Feldman).
- Smith, A. F. M., and Roberts, G. O., Bayesian computation via Gibbs sampler and related Markov Chain Monte Carlo methods, *J. Roy. Statist. Soc. B* **55** (1993), 3–24.
- Spencer, J. E., and Largey, A., Geary on inference in multiple regression and on closeness and the taxi problem, *Econ. and Social Rev.*, **24** (3) (1993), 275–295.
- Spiegelhalter, D. J., Best, N. J., Gilks, W. R. and Inskip, H., Hepatitis B: a case study in MCMC methods, Chapter 2, in Gilks *et al.* (1996).
- Sprent, P., *Models in Regression and Related Topics*, London: Methuen (1969).
- Stein, C., Inadmissibility of the usual estimator for the mean of the multivariate normal distribution, *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, CA, and Los Angeles, CA: University of California Press (1956), pp.197–206 [reprinted in Kotz and Johnson (1992–1997, Volume I)].
- Stern, C., and Sherwood, E. R. (eds), *The Origin of Genetics: A Mendel Source Book*, San Francisco: W. H. Freeman (1966).
- Stigler, S., *American Contributions to Mathematical Statistics in the Nineteenth Century* (2 vols), New York: Arno Press (1980).

- Stigler, S., *The History of Statistics: The Measurement of Uncertainty before 1900*, Cambridge, MA: Harvard University Press (1986a).
- Stigler, S. M., *Statistics on the Table: The History of Statistical Concepts and Methods*, Cambridge, MA, and London: Harvard University Press (1999).
- Stigler, S., Laplace's 1774 memoir on inverse probability, *Statistical Science*, **1** (1986b), 359–378.
- 'Student' (W. S. Gosset), The probable error of a correlation coefficient, *Biometrika*, **6** (1908), 1–25.
- 'Student' (W. S. Gosset), *Student's Collected Papers*, Cambridge: Cambridge University Press for Biometrika (1942).
- Tanner, M. A., *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions* (3rd edn), Berlin: Springer (1996) [1st edn (1991), 2nd edn (1993)].
- Tanner, M. A., and Wong, W. H., The calculation of posterior distributions by data augmentation (with discussion), *J. Amer. Statist. Assoc.*, **82** (1987), 528–550 [reprinted in Polson and Tiao (1995, Volume II)].
- Theobald, C. M., Generalizations of mean square error applied to ridge regression, *J. Roy. Statist. Soc. B*, **34** (1974), 103–105.
- Tierney, L., Markov chains for exploring posterior distributions (with discussion), *Annals of Statistics*, **22** (1994), 1701–1762.
- Todhunter, I., *A History of the Mathematical Theory of Probability from the Time of Pascal to That of Laplace*, London: Macmillan (1865) [reprinted New York: Chelsea (1949)].
- Toni, T., and Stumpf, M. P. H., Simulation-based model selection for dynamical systems in systems and population biology, *Bioinformatics*, **26** (1) (2010), 104–110.
- Turchin, V. F., On the Computation of Multidimensional Integrals by the Monte-Carlo Method, *Theory of Probability and its Applications*, **16** (4) (1971), 720–724 [translation of K vychisleniyu mnogomernyx integralov po metodu Monte-Carlo, *Teoriya Veroyatnostej i ee Primeneniya*, **16** (4) (1971), 738–74].
- Turner, P. S., The distribution of l.s.d. and its implications for computer design, *Math. Gazette*, **71** (1987), 26–31.
- Venables, W. N., and Ripley, B. D., *Modern Applied Statistics with S* (4th edn), Berlin: Springer-Verlag 2002 [1st edn (1995), 2nd edn (1997), 3rd edn (1999)].

- von Mises, R., Über die ‘Ganz-zahligkeit’ der Atomgewicht und verwandte Fragen, *Physikal. Z.*, **19** (1918), 490–500.
- von Mises, R., On the correct use of Bayes’ formula, *Ann. Math. Statist.*, **13** (1942), 156–165.
- von Mises, R., Selected Papers (2 vols), Providence, RI: American Mathematical Society (1963–1964).
- von Neumann, J., and Morgenstern, O., Theory of Games and Economic Behaviour, Princeton, NJ: Princeton University Press (1944, 1947, 1953).
- Walther, G., Inference and modelling with log-concave distributions, *Statistical Science*, **24** (2009), 319–327.
- Watkins, P., Story of the W and the Z, Cambridge: Cambridge University Press (1986).
- Weir, C., and Murray, G., Fraud in clinical trials, *Significance*, **8** (2011), 164–168.
- Weisberg, H. L., Bayesian comparison of two ordered multinomial populations, *Biometrics*, **23** (1972), 859–867.
- Weisberg, S., Applied Linear Regression (3rd edn), New York: John Wiley & Sons (2005) [1st edn (1980), 2nd edn (1985)].
- West, M., and Harrison, P. J., Bayesian Forecasting and Dynamic Models, Berlin: Springer (1997) [1st edn (1989)].
- Whitaker's Almanack, London: A. & C. Black (annual).
- Whittaker, E. T., and Robinson, G., The Calculus of Observations (3rd edn), Edinburgh: Blackie (1940) [1st edn (1924), 2nd edn (1926)].
- Whittaker, E. T., and Watson, G. N., A Course of Modern Analysis (4th edn), Cambridge: Cambridge University Press (1927).
- Williams, J. D., The Compleat Strategyst, Being a Primer on the Theory of Games of Strategy (2nd edn), New York: McGraw-Hill (1966) [1st edn (1954)].
- Wilson, E. B., An Introduction to Scientific Research, New York: McGraw-Hill (1952).
- Wishart, J., and Sanders, H. G., Principles and Practice of Field Experimentation (2nd edn), Cambridge: Commonwealth Bureau of Plant Breeding and Genetics (1955) [1st edn (1935)].
- Young, A. S., A Bayesian approach to prediction using polynomials, *Biometrika*, **64** (1977), 309–317.

Zellner, A., An Introduction to Bayesian Inference in Econometrics, New York: John Wiley & Sons (1971).

Zellner, A., Basic Issues in Econometrics, Chicago: University of Chicago Press (1974).

Zellner, A., Maximal data information prior distributions, in Aykaç and Brumat (1977).

Zipf, G. K., The Psycho-biology of Language, Boston, MA: Houghton-Mifflin (1935).