Bayesian Methods for Ecology

Michael A. McCarthy

$$\Pr(H_i|D) = \frac{\Pr(H_i)\times\Pr(D|H_i)}{\sum_j \Pr(H_j)\times\Pr(D|H_j)}$$

CAMBRIDGE

CAMBRIDGE www.cambridge.org/9780521850575

This page intentionally left blank

Bayesian Methods for Ecology

The interest in using Bayesian methods in ecology is increasing, but most ecologists do not know how to carry out the required analyses. This book bridges that gap. It describes Bayesian approaches to analysing averages, frequencies, regression, correlation and analysis of variance in ecology. The book also incorporates case studies to demonstrate mark-recapture analysis, development of population models and the use of subjective judgement. The advantages of Bayesian methods, including the incorporation of any relevant prior information and the ability to assess the evidence in favour of competing hypotheses, are also described here. The analyses described in this book use the freely available software WinBUGS, and there is an accompanying website (http://arcue.botany.unimelb.edu.au/bayes.html) containing the data files and WinBUGS codes that are used in the book. The Bayesian methods described here will be of use to ecologists from the level of upper undergraduate and above.

MICHAEL A. McCARTHY is Senior Ecologist at the Royal Botanical Gardens, Melbourne and Senior Fellow in the School of Botany at the University of Melbourne.

Bayesian Methods for Ecology

MICHAEL A. McCARTHY

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, São Paulo

Cambridge University Press
The Edinburgh Building, Cambridge CB2 8RU, UK
Published in the United States of America by Cambridge University Press, New York
www.cambridge.org
Information on this title: www.cambridge.org/9780521850575

© M. A. McCarthy 2007

This publication is in copyright. Subject to statutory exception and to the provision of relevant collective licensing agreements, no reproduction of any part may take place without the written permission of Cambridge University Press.

First published in print format 2007

ISBN-13 978-0-511-28494-6 eBook (EBL)
ISBN-10 0-511-28494-2 eBook (EBL)

ISBN-13 978-0-521-85057-5 hardback
ISBN-10 0-521-85057-6 hardback

ISBN-13 978-0-521-61559-4 paperback
ISBN-10 0-521-61559-3 paperback

Cambridge University Press has no responsibility for the persistence or accuracy of urls for external or third-party internet websites referred to in this publication, and does not guarantee that any content on such websites is, or will remain, accurate or appropriate.

To Kirsten and Owen

Contents

CASE STUDIES

APPENDICES

Preface

I have three vivid memories about learning statistics as an undergraduate that all involve misconceptions. Firstly, I remember my lecturer telling me that, after obtaining a result that was not statistically significant, I should conclude that timber harvesting did not have an effect (on what, I cannot remember). While the logic was flawed, I have since realized that it is a misconception shared by many ecologists.

My second memory is of reading about Bayesian analyses in journal articles. I wondered what Bayesian methods were, how they differed from the statistical approaches I had been taught (frequentist methods such as null hypothesis testing and construction of confidence intervals), and why I had never heard of them before. On reading the articles, I concluded that Bayesian methods must be hard to do. It turns out that I was incorrect again.

My third memory is that statistics was boring. I was wrong again. I was reasonably good at the mathematics involved, but it was not until I started doing my own data analyses during my Ph.D. that I saw the benefits of using statistics. I began to learn about different ways to do statistics (e.g. likelihood-based methods), and also re-learnt some old topics (e.g. realizing the importance of and learning how to calculate statistical power). For me, statistics and probability continue to be a world of learning.

This book represents a stage in my journey through the world of statistics. It is born out of a frustration with how conventional statistical methods are misused in ecology on a routine basis, and a belief that Bayesian methods are relevant and useful. I hope this book convinces readers of the value of Bayesian methods and helps them learn Bayesian methods more quickly than me.

Approximately five years ago I used null hypothesis significance testing to evaluate the predictions of some models of population viability. An astute reviewer questioned this approach because the models were surely known to be wrong a priori. The reviewer provided a glorious list of quotes that attacked null hypothesis significance testing (not unlike the quotes in Chapter 2). I started thinking about alternatives, leading me to Hilborn and Mangel's (1997) *The Ecological Detective*, and beyond.

The Ecological Detective (Hilborn and Mangel, 1997) is one of the best books available to ecologists for learning about Bayesian methods. However, ecologists wishing to use the suggested methods need at least some skills in computer programming. I intend my book to provide a bridge between a desire to conduct Bayesian analyses and the book by Hilborn and Mangel (1997). WinBUGS code for the analyses in this book is available from http://arcue.botany.unimelb.edu.au/bayes.html.

The bridge is built by using the freely available program WinBUGS (Spiegelhalter *et al.*, 2005; Appendix A) to conduct the vast majority of analyses in this book. I try to start gently, illustrating the most basic analyses, before giving some more complicated examples. More experienced users will find some analyses trivial, and novices may find some examples impenetrable. The aim is to provide a sufficient diversity of examples that the reader will be able to learn how to construct their own statistical models and conduct their own analyses.

This book is not necessarily designed to be read from cover to cover. Read Chapters 1 and 2 if you wish to know more about the differences between Bayesian and frequentist methods. If you just want to learn how to conduct Bayesian analyses, start with Chapter 1, Appendix A, and then move to Chapter 3 or whichever topic is most relevant. As you become more familiar with Bayesian methods, the entire content of the book will become more accessible.

I have many people to thank for their help while writing this book. Ralph Mac Nally and Alan Crowden's suggestion to write this book started the ball rolling. Brendan Wintle has been extremely important as a colleague, a source of advice and insights, and a sounding board for ideas. Kirsten Parris, David Lindenmayer, Jane Elith, Pip Masters, Linda Broome, Tara Martin, Mark McDonnell, Michael Harper, Brendan Wintle, Amy Hahs, Rodney van der Ree and many others have provided data for analysis over the years. I would have learnt very little without them.

This book owes much to the availability of WinBUGS, and I thank the team that developed the software. In particular, David Spiegelhalter,

Andrew Thomas and Bob O'Hara have answered questions and investigated software bugs from time to time.

Hugh Possingham, Mark Burgman, David Lindenmayer and Mark McDonnell have provided opportunities for me to conduct research into risk assessment, and the use of Bayesian methods. They have been influential, as have Fiona Fidler, Neil Thomason and my colleagues listed above. Various funding agencies have supported my research, with the Australian Research Council and The Baker Foundation being particularly generous.

I'd like to thank my parents David and Sue for fostering my interests in mathematics and ecology, and Dr John Gault for his enthusiastic interest in Bayesian issues of a medical nature.

Finally, thank you to everyone who has provided comments, in particular Mark Burgman, Geoff Cumming, Aaron Ellison, Ralph Mac Nally, Kirsten Parris, Gerry Quinn and Julia Stammers, who read most if not all of earlier drafts. Peter Baxter, Barry Brook, Ryan Chisholm, Gareth Davies, Lou Elliott, Fiona Fidler, Bill Langford, Terry Walshe and Nick Williams also provided helpful comments on various sections.

1
Introduction

There is a revolution brewing in ecology. Granted, it is a gentle and slow revolution, but there is growing dissatisfaction with the statistical methods that have been most commonly taught and used in ecology (Hilborn and Mangel, 1997; Wade, 2000; Clark, 2005).[1] One aspect of this revolution is the increasing interest in Bayesian statistics (Fig. 1.1). This book aims to foster the revolution by making Bayesian statistics more accessible to every ecologist.

Ecology is the scientific study of the distribution and abundance of biological organisms, and how their interactions with each other and the environment influence their distribution and abundance (Begon *et al.*, 2005). The discipline depends on the measurement of variables and analysis of relationships between them. Because of the size and complexity of ecological systems, ecological data are almost invariably subject to error. Ecologists use statistical methods to distinguish true responses from error. Statistical methods make the interpretation of data transparent and repeatable, so they play an extremely important role in ecology.

The Bayesian approach is one of a number of ways in which ecologists use data to make inferences about nature. The different approaches are underpinned by fundamentally different philosophies and logic. The appropriateness of different statistical approaches has been fiercely debated in numerous disciplines but ecologists are only now becoming aware of this controversy. This occurs at least in part because the majority of statistical books read by ecologists propound conventional

[1] The conventional statistical methods are known as frequentist statistics and include null hypothesis significance testing (NHST) and construction of confidence intervals. NHST attracts the most criticism. See Chapter 2 for more details of these methods.

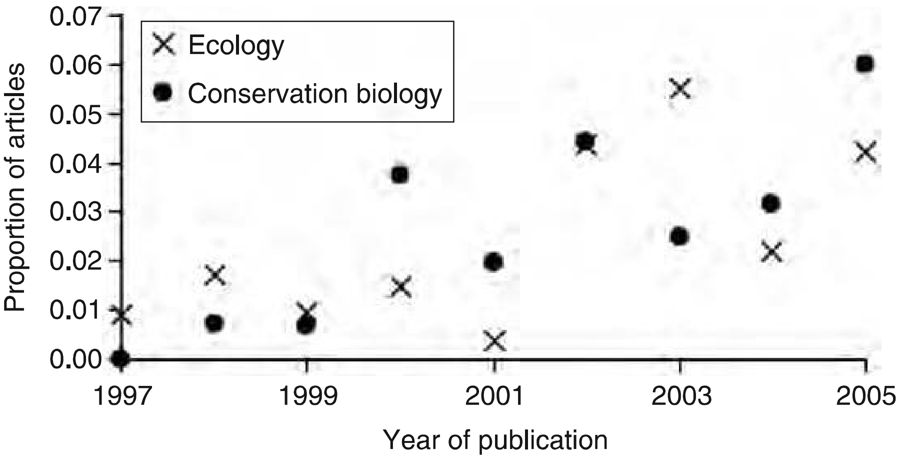

Fig. 1.1 The proportion of articles in the journals *Ecology* and *Conservation Biology* that refer to 'Bayes' or 'Bayesian'.

statistics, ignore criticisms of these methods and do not acknowledge that there are alternatives (Fowler *et al.*, 1998; Sokal and Rohlf, 1995; Underwood, 1997; Zar, 1999). Those that do address the controversy usually aim to change the status quo (Hilborn and Mangel, 1997; Burnham and Anderson, 2002), although there are exceptions (Quinn and Keough, 2002; Gotelli and Ellison, 2004).

The Bayesian approach is used relatively rarely (Fig. 1.1), so why should it interest ecologists? There are several reasons but two are particularly relevant ones. Firstly, Bayesian methods are fully consistent with mathematical logic, while conventional statistics are only logical when making probabilistic statements about data, not hypotheses (Cox, 1946; Berger and Berry, 1988; Jaynes, 2003). Bayesian methods can be used to make probabilistic predictions about the state of the world, while conventional statistics are restricted to statements about long-run averages obtained from hypothetical replicates of sampled data.

Secondly, relevant prior information can be incorporated naturally into Bayesian analyses by specifying the appropriate prior probabilities for the parameters. In contrast, conventional statistical methods are forced to ignore any relevant information other than that contained in the data. Difficulties with Bayesian methods and other benefits are discussed more fully in Chapter 2 and throughout this book.

Bayesian statistics are founded on the work of the Reverend Thomas Bayes, who lived and died in eighteenth century England (Box 1.1). Bayesian methods explicitly recognize and combine four

Box 1.1
The Reverend Thomas Bayes, FRS

Very little is known about Thomas Bayes. The portrait above (O'Donnell, 1936) may be of Bayes, but no other portraits are known (Bellhouse, 2004). Even the year (1701 or 1702) and place of his birth (London or Hertfordshire, England) are uncertain (Dale, 1999). There are few records to indicate the nature of his early schooling, but he is known to have studied divinity and mathematics at the University of Edinburgh. He was ordained as a Presbyterian minister by 1728. He was elected as a Fellow of the Royal Society in 1742 but it was not until after his death in 1761 that his most famous contribution, his essay in the *Philosophical Transactions* of the Royal Society of London, was published (Bayes, 1763). In that essay, Bayes described his theory of probability and presented what is now known as Bayes' rule (or Bayes' theorem), establishing the basis of Bayesian statistics.

components of knowledge. Prior knowledge and new data are combined using a model to produce posterior knowledge.[2] These four components may be represented as:

$$\text{prior} + \text{data} \xrightarrow{\text{model}} \text{posterior}$$

It is common in everyday life to combine prior information and new data to update knowledge. We might hear a weather forecast that the chance of rain is small. However, if we stepped outside and saw dark

[2] Prior and posterior refer to before and after considering the data.

clouds looming above us, most people would think that the risk of rain was higher than previously believed. In contrast, our expectation of a fine day would be reinforced by a sunny sky. Thus, both the prior information (the weather forecast) and the data (the current state of the weather) influence our newly updated belief in the prospects of rain.

Our updated belief in the chance of rain (the posterior) will depend on the relative weight we place on the prior information compared to the new data and the magnitude of the difference between the two pieces of information. In this case the 'model' is contained within our understanding of the weather. Our thought processes combine the prior information, data, and model to update our belief that it will rain. Bayesian statistics provide a logically consistent, objective and repeatable method for combining prior information with data to produce the posterior, rather than the subjective judgement that most people would use when stepping outside.

Before considering the benefits and limitations of Bayesian methods and its alternatives in Chapter 2, I will illustrate the use of the different statistical approaches with two examples. These highlight how Bayesian methods provide answers to the kinds of questions that ecologists ask, and how they can usefully incorporate prior information.

Example 1: Logic in determining the presence or absence of a species

Consider an ecologist who surveys ponds in a city for frogs. On her first visit to a pond, she searches the edge and listens for frog calls over a 20-minute period. The southern brown tree frog (*Litoria ewingii*) is the most common species in her study area, but it is not found on this particular visit (Fig. 1.2). However, the researcher would not be particularly surprised that the species was not detected because she knows from experience that when surveying ponds, southern brown tree frogs are detected on only 80% of visits when they are in fact present. Given this information, what can she conclude about whether the southern brown tree frog is present at the site or not?

The question about the presence of a species is a simple example of those asked by ecologists. We assume that there is a particular true state of nature and we hope to use scientific methods to determine a reasonable approximation of the truth. However, the probability that a species is

Fig. 1.2 The southern brown tree frog *Litoria ewingii*, a common species in the ponds of Melbourne, Victoria. Photograph by Nick Clemann.

present at a site is rarely calculated by ecologists, although it should be a fundamental part of any field study that depends on knowing where a species does and does not occur. This probability is not calculated partly because the statistical methods used by most ecologists are not well-suited to this question. I will examine three different approaches to answering this question and demonstrate that a satisfactory answer requires Bayesian methods.

Frequentist approaches

Conventional approaches to data analysis in ecology estimate the likelihood of observing the data (and more extreme data in the case of null hypothesis testing). These approaches are referred to as frequentist methods because they are based on the expected frequency that such data would be observed if the same procedure of data collection and analysis was implemented many times. Frequentist methods focus on the frequency with which the observed data are likely to be obtained from hypothetical replicates of sampling.

There are numerous types of frequentist statistics that are used in ecology, including null hypothesis significance testing and information-theoretic methods. These are applied below to the question about whether southern brown tree frogs are present at the pond.

Null hypothesis significance testing

The first statistical approach to answering the question is null hypothesis significance testing. The null hypothesis for this first case might be that the southern brown tree frog is absent from the site. The researcher then seeks to disprove the null hypothesis with the collection of data. The single piece of data in this case is that the frog was not detected. The researcher then asks: ‘What is the probability of obtaining this result if the null hypothesis were true?’[3] This probability is the *p*-value of the significance test. If the *p*-value is sufficiently small (conventionally if less than 0.05), it means that the data (or more extreme data) would be unlikely to occur if the null hypothesis is true. If the *p*-value is small, then we assume that the data are inconsistent with the null hypothesis, which is then rejected in favour of the alternative.

In the case of the frog survey, the *p*-value is equal to 1.0. This is calculated as the probability that we would fail to record the frog (i.e. obtain the observed data) if it is absent (i.e. if the null hypothesis is true). The high *p*-value means that the researcher fails to reject the null hypothesis that the frog is absent.

The other possible null hypothesis is that the frog is present at the site. In this case, the probability of obtaining the data is equal to 0.2 (one minus the probability of detecting the species if present) given that the null hypothesis is true. Thus, the *p*-value is 0.2, and using a conventional cut-off of 0.05, the researcher would have a non-significant result. The researcher would fail to reject the null hypothesis that the southern brown tree frog was present.

It is surprising (to some people) that the two different null hypotheses can produce different results. The conclusion about whether the species is present or absent simply depends on which null hypothesis we choose. The source of this surprise is our failure to consider statistical power, which I will return to in Chapter 2.

Another possible source of surprise is that the *p*-value does not necessarily provide a reliable indicator of the support for the null hypotheses. For example, the *p*-value is equal to 1.0 for the null hypothesis that the frog is absent. This is the largest possible *p*-value, but it is still not proof that the null hypothesis is true. If we continued to return to the

[3] In actual fact, a null hypothesis significance test asks what is the probability of obtaining the data *or a more extreme result*. However, in this case, a more extreme result is not possible; it is not possible to fail to detect the frog more than once with one visit, so the *p*-value is simply the probability of observing the data.

same pond and failed to find the frog, the *p*-value would remain equal to 1.0, insensitive to the accumulation of evidence that the frog is absent. This apparent discrepancy occurs because frequentist methods in general and *p*-values in particular do not provide direct statements about the reliability of hypotheses (Berger and Sellke, 1987; Berger and Berry, 1988). They provide direct information about the frequency of occurrence of data, which only gives indirect support for or against the hypotheses. In this way, frequentist methods are only partially consistent with mathematical logic, being confined to statements about data but not directly about hypotheses (Berger and Sellke, 1987; Jaynes, 2003).

Information theoretic methods

An information theoretic approach based on 'likelihood' is an alternative frequentist method to null hypothesis significance testing. It evaluates the consistency of the data with multiple competing hypotheses (Burnham and Anderson, 2002). In the current example, there are only two possible hypotheses: the frog is absent (H_a) and the frog is present (H_p). Likelihood-based methods ask: 'What is the probability of observing the data under each of the competing hypotheses?' In this example it is the probability of not detecting the species during a visit to a site.

Unlike null hypothesis testing, likelihood-based methods, including information-theoretic methods, do not consider the possibility of more extreme (unobserved) data. The likelihood for a given hypothesis can be calculated as the probability of obtaining the data given that the hypothesis is true.[4] Despite the implication of its name, the likelihood of a hypothesis is not the same as the probability that the hypothesis is true.

Under the first hypothesis (the frog is absent), the probability of observing the data ($\Pr(D \mid H_\mathrm{a})$) is equal to 1. Under the second hypothesis (the frog is present) the probability ($\Pr(D \mid H_\mathrm{p})$) is 0.2. Information-theoretic methods then determine the amount of evidence in favour of these two hypotheses by examining the ratio of these values (Burnham and Anderson, 2002).[5] These ratios may be interpreted by rules of thumb (see also Chapter 4). Using the criteria of Burnham and Anderson (2002),

[4] The likelihood need only be proportional to the probability of obtaining the data, not strictly equal to it. Terms that do not include the data or the parameters being estimated can be ignored because they will cancel out of the subsequent calculations.

[5] Information-theoretic methods are modified by the number of parameters that are estimated with the data. In this case, the parameter of the analyses (the detection rate) is not estimated with the data, so the number of estimated parameters is zero.

we might conclude that the southern brown tree frog is ‘considerably less’ likely to be present than it is to be absent ($\Pr(D \mid H_p)/\Pr(D \mid H_a) = 1/5$).

Bayesian methods

Frequentist methods are in general not well-suited to the species detection problem because they are strictly limited to assessing long-run averages rather than predicting individual observations (Quinn and Keough, 2002). This is revealing; frequentist methods are not strictly suitable for predicting whether a species is absent from a particular site when it has not been seen. Such a problem is fundamental in ecology, which relies on knowing the distribution of species. In contrast, the species detection problem can be tackled using Bayesian methods.

Bayesian methods are similar to likelihood-based methods, but also incorporate prior information using what is known as ‘prior probabilities’. Bayesian methods update estimates of the evidence in favour of the different hypotheses by combining the prior probabilities and the probabilities of obtaining the data under each of the hypotheses. The probability that a hypothesis is true increases if the data support it more than the competing hypotheses.

Why might the prior information be useful? If the researcher visited a pond that appeared to have excellent habitat for southern brown tree frogs (e.g. a large well-vegetated pond in a large leafy garden), then a failure to detect the species on a single visit would not necessarily make the researcher believe that the frog was absent. However, if the researcher visited a pond that was very unlikely to contain the frog (e.g. a concrete fountain in the middle of an asphalt car park), a single failure to detect the frog might be enough to convince the researcher that the southern brown tree frog did not occur at the pond. Frequentist methods cannot incorporate such prior information, but it is integral to Bayesian methods.

Another key difference between Bayesian methods and frequentist methods is that instead of asking: ‘What is the probability of observing the data given that the various hypotheses are true?’ Bayesian methods ask:

What is the probability of the hypotheses being true given the observed data?

At face value, this is a better approach for our problem because we are interested in the truth of the hypotheses (the frog’s presence or absence at the site) rather than the probability of obtaining the observed data given different possible truths.

In practice, Bayesian methods differ from likelihood methods by weighting the likelihood values by the prior probabilities to obtain posterior probabilities. I will use the two symbols $\Pr(H_a)$ and $\Pr(H_p)$ to represent the prior probabilities. Therefore, the likelihood for the presence of the frog given that it was not seen (0.2) is weighted by $\Pr(H_p)$ and the likelihood for the absence of the frog (1.0) is weighted by $\Pr(H_a)$. Thus, the posterior probability of presence is a function of the prior probability $\Pr(H_p)$, the data (the frog was not seen) and the model, which describes how the data were generated conditional on the presence or absence of the frog. Now we must determine a coherent scheme for determining the values for the prior probabilities $\Pr(H_p)$ and $\Pr(H_a)$. This incorporation of prior information is one of the unique aspects of Bayesian statistics. It also generates the most controversy.

Both hypotheses might be equally likely (prior to observing the data) if half the sites in the study area were occupied by southern brown tree frogs (Parris unpublished data). In this case, $\Pr(H_a) = 0.5$, as does $\Pr(H_p)$. With these priors, the probability of the southern brown tree frog being absent will be proportional to $0.5 \times 1.0 = 0.5$, and the probability of it being present will be proportional to $0.5 \times 0.2 = 0.1$.

The posterior probabilities must sum to one, so these proportional values (0.5 and 0.1) can be converted to posterior probabilities by dividing by their sum $(0.5 + 0.1 = 0.6)$. Therefore, the probability of the frog being present is $1/6$ $(= 0.1/0.6)$, and the probability of absence is $5/6$ $(= 0.5/0.6)$. So, with equal prior probabilities $(\Pr(H_a) = \Pr(H_p) = 0.5)$, we would conclude that the presence of the frog is five times less probable than the absence of the frog because the ratio $(\Pr(H_p \mid D)/\Pr(H_a \mid D))$ equals $1/5$. You may have noticed that this result is numerically identical to the likelihood-based result. I will return to this point later.

A different prior could have been chosen for the analysis. A statistical model predicts the probability of occupancy of ponds by southern brown tree frogs based on the level of urbanization (measured by road density), characteristics of the vegetation, and the size of the pond (based on Parris 2006.). If the pond represented relatively high-quality habitat, with a predicted probability of occupancy of 0.75, then the probability of the frog being present will be proportional to $0.75 \times 0.2 = 0.15$ and the probability of absence will be proportional to $(1 - 0.75) \times 1.0 = 0.25$. With these priors, the probability of the frog being present is equal to $3/8$ $(= 0.15/(0.15 + 0.25))$, and the probability of absence is $5/8$ $(= 0.25/(0.15 + 0.25))$.

The incorporation of prior information (the presence of good quality habitat) increases the probability that the pond is occupied by southern brown tree frogs compared to when the prior information is ignored (0.375 versus 0.167). The actual occupancy has not changed at all – the pond is still either occupied or not. What has changed is the researcher's belief in whether the pond is occupied. These Bayesian analyses may be formalized using Bayes' rule, which, following a short introduction to conditional probability (Box 1.2), is given in Box 1.3.

Box 1.2

Conditional probability

Bayes' rule is based on conditional probability. Consider two events: event C and event D. We are interested in the probability of event C occurring given event D has occurred. I will write this probability using the symbol $\Pr(C \mid D)$, and introduce three more symbols:

$\Pr(C)$ – the probability of event C occurring;

$\Pr(D)$ – the probability of event D occurring; and

$\Pr(C \text{ and } D)$ – the probability of both events occurring together.

Conditional probability theory tells us that:

$$\Pr(C \text{ and } D) = Pr(D) \times \Pr(C \mid D),$$

which in words is: the probability of events C and D both occurring is equal to the probability of event C occurring given that event D has occurred multiplied by the probability of event D occurring (independent of event C). The | symbol means 'given the truth or occurrence of'.

The above can be rearranged to give:

$$\Pr(C \mid D) = \Pr(C \text{ and } D) / \Pr(D).$$

For example, *Pfiesteria*, a toxic alga is present in samples with probability 0.03 (Stow and Borsuk 2003). *Pfiesteria* is a subset of *Pfiesteria*-like organisms (PLOs), the latter being present in samples with probability 0.35. Therefore, we can calculate the conditional probability that *Pfiesteria* is present given that PLOs are present:

$$\begin{aligned}\Pr(\textit{Pfiesteria} \mid \text{PLO}) &= \Pr(\textit{Pfiesteria} \text{ and PLO}) / \Pr(\text{PLO}) \\ &= 0.03/0.35 = 0.086.\end{aligned}$$

Box 1.3

Bayes' rule for a finite number of hypotheses

Conditional probability (Box 1.2) states that for two events C and D:

$$\Pr(C \text{ and } D) = \Pr(D) \times \Pr(C \mid D).$$

C and D are simply labels for events (outcomes) that can be swapped arbitrarily, so the following is also true:

$$\Pr(D \text{ and } C) = \Pr(C) \times \Pr(D \mid C).$$

These two equivalent expressions for $\Pr(C \text{ and } D)$ can be set equal to each other:

$$\Pr(D) \times \Pr(C \mid D) = \Pr(C) \times \Pr(D \mid C).$$

It is then straightforward to obtain:

$$\Pr(C \mid D) = \Pr(C) \times \Pr(D \mid C)/\Pr(D).$$

Let us assume that event C is that a particular hypothesis is true, and event D is the occurrence of the data. Then, the posterior probability that the frog is absent given the data ($\Pr(H_\mathrm{a} \mid D)$) is:

$$\Pr(H_\mathrm{a}|D) = \Pr(H_\mathrm{a}) \times \Pr(D|H_\mathrm{a})/\Pr(D).$$

The various components of the equation are the prior probability that the frog is absent ($\Pr(H_\mathrm{a})$), the probability of obtaining the data given that it is absent ($\Pr(D \mid H_\mathrm{a})$, which is the likelihood), and the probability of obtaining the data independent of the hypothesis being considered ($\Pr(D)$).

The probability of obtaining the data (the frog was not detected) given H_a is true (the frog is absent) was provided when using the likelihood-based methods:

$$\Pr(D \mid H_\mathrm{a}) = 1.0.$$

Similarly, given the presence of the frog:

$$\Pr(D \mid H_\mathrm{p}) = 0.2.$$

The value of $\Pr(D)$ is the same regardless of the hypothesis being considered (H_p the frog is present, or H_a the frog is absent), so it simply acts as a scaling constant. Therefore, $\Pr(H_\mathrm{a} \mid D)$ is proportional to $\Pr(H_\mathrm{a}) \times \Pr(D \mid H_\mathrm{a})$, and $\Pr(H_\mathrm{p} \mid D)$ is proportional to

$\Pr(H_p) \times \Pr(D \mid H_p)$, with both expressions having the same constant of proportionality ($1/\Pr(D)$).

$\Pr(D)$ is calculated as the sum of the values $\Pr(H) \times \Pr(D \mid H)$ under all hypotheses. When prior probabilities are equal ($\Pr(H_a) = \Pr(H_p) = 0.5$):

$$\begin{aligned}\Pr(D) &= [\Pr(H_a) \times \Pr(D \mid H_a)] + [\Pr(H_p) \times \Pr(D \mid H_p)] \\ &= (0.5 \times 1) + (0.5 \times 0.2) = 0.6.\end{aligned}$$

Therefore, the posterior probabilities are 5/6 (0.5/0.6) for the absence of the frog, and 1/6 (0.1/0.6) for the presence of the frog.

So, for a finite number of hypotheses, Bayes' rule states that the probability of the hypothesis given the data is calculated using the prior probabilities of the different hypotheses ($\Pr(H_j)$) and the probability of obtaining the data given the hypotheses ($\Pr(D \mid H_j)$):

$$\Pr(H_i \mid D) = \frac{\Pr(H_i) \times \Pr(D \mid H_i)}{\sum_j \Pr(H_j) \times \Pr(D \mid H_j)}$$

This expression uses the mathematical notation for summation $\sum$.

If on the other hand, the pond had poor habitat for southern brown tree frogs, the prior probability of presence might be 0.1. Thus, $\Pr(H_p) = 0.1$ and $\Pr(H_a) = 0.9$. As before, $\Pr(D \mid H_p) = 0.2$ and $\Pr(D \mid H_a) = 1.0$. Note that the values for the priors but not the likelihoods have changed. Using Bayes' rule (Box 1.3), the posterior probability of presence is:

$$\begin{aligned}\Pr(H_p \mid D) &= \Pr(H_p) \times \Pr(D \mid H_p) / [\Pr(H_p) \times \Pr(D \mid H_p) + \Pr(H_a) \times \\ &\quad \Pr(D \mid H_a)] \\ &= 0.1 \times 0.2 / [0.1 \times 0.2 + 0.9 \times 1.0] \\ &= 0.022\end{aligned}$$

Therefore, there is only a small chance that the frog is at the site if it has poor habitat and the species is not detected on a single visit.

Bayesian methods use probability distributions to describe uncertainty in the parameters being estimated (see Appendix B for more background on probability distributions). Probability distributions are used for both priors and posteriors. The frog surveying problem has two possible outcomes; the frog is either present or absent. Such a binary outcome (e.g. presence/absence, heads/tails, increasing/decreasing) can be represented by a Bernoulli probability distribution, which is a special case of the binomial distribution with a sample size of one. Bernoulli random variables take a value of one (representing the presence of the frog) with a probability equal to p and a value of zero (representing the absence of the frog) with probability 1-p. Therefore, uncertainty about the presence of the frog at the pond can be represented as a Bernoulli random variable in which the probability of presence is equal to p.

It is important to note that a probability distribution is used to represent the *uncertainty* about the presence of the frog. The frog is assumed to be actually present or absent at the site, and the distribution is used to represent the probability that it is present. There appears to be misunderstanding among at least some ecologists that Bayesian parameters do not have fixed values, but change randomly from one measurement to another. Although such models can be accommodated within Bayesian analyses (e.g. by using hierarchical models, Box 3.6), parameters are usually assumed to have fixed values. The prior and posterior distributions are used to represent the uncertainty about the estimate of the parameters.

I have illustrated three components of a Bayesian analysis: priors, data and posteriors. I have not explicitly stated the model, which is the fourth aspect I mentioned in the introduction. The model in the above example is relatively simple and is the same as was used in the frequentist analyses. It can be stated as: 'the detection of the southern brown tree frog during the survey occurs randomly with a probability (p_{detect}) that depends on whether the pond is occupied ($p_{\text{detect}} = 0.8$) or not ($p_{\text{detect}} = 0.0$)'.

This model may be written algebraically as:

$$p_{\text{detect}} = 0.8 \times \textit{present}$$
$$\textit{detected} \sim \text{Bernoulli}(p_{\text{detect}}).$$

The second expression says that the variable called '*detected*' is a Bernoulli random variable. A value of one for '*detected*' indicates that the frog was detected and a zero indicates it was not. The probability of detection is equal to p_{detect}, and is given in the first equation. It depends on whether the frog is present at the site ($\textit{present} = 1$, $p_{\text{detect}} = 0.8$) or absent ($\textit{present} = 0$, $p_{\text{detect}} = 0.0$).

Random sampling from the posterior distribution using WinBUGS

This Bayesian analysis can also be implemented in the freely available software package WinBUGS (Spiegelhalter *et al*., 2005). Appendix A provides information about obtaining the program WinBUGS and a tutorial on its use. I will use WinBUGS throughout the book, so it is worth investing some time in understanding it. Readers who are unfamiliar with WinBUGS should study Appendix A now, before continuing with the rest of the book.

The acronym WinBUGS is based on the original program BUGS (Bayesian inference Using Gibbs Sampling), but is now designed to run under the Microsoft Windows operating system (hence the Win prefix). WinBUGS works by randomly sampling the parameters used in Bayesian models from their appropriate posterior distributions. Because the posterior distribution for the example of detecting southern brown treefrogs can be calculated (Box 1.3), it is not necessary to use WinBUGS in this case. However, for many problems it is difficult or impossible to calculate the posterior distribution, but samples from it can be obtained relatively easily using WinBUGS or other MCMC software. If a sufficiently large number of replicates are taken, the form of the posterior distribution can be determined and its parameters, such as the mean, standard deviation, and percentiles, can be estimated.

WinBUGS takes samples from the posterior distribution by using 'Markov chain Monte Carlo' (MCMC) methods. 'Monte Carlo' implies random sampling, referring to roulette wheels and other games of chance. 'Markov chain' refers to the method of generating the random samples. A series of random numbers in which the value of each is conditional on the previous number is known as a Markov chain. MCMC algorithms are constructed in such a way that the samples from the Markov chain are equivalent to samples from the required posterior distribution (see Appendix C).

The advantage of using Markov chains for sampling from the posterior distribution is that it is not necessary to calculate the value of the denominator in Bayes' rule. The calculation is avoided because each successive sample depends on the ratio of two posterior probabilities that share the same denominator, which then cancels (Appendix C). This simplifies matters, because the Bayesian analysis only requires the product of the prior probability and the likelihood of the data.

If each sample depends on the value of the previous sample, successive values drawn from the Markov chain may be correlated. Correlations between the samples have some important consequences. The first is that the initial values that are used in the Markov chain may influence the results until a sufficiently large number of samples is generated. After this time, the 'memory' of the initial values is sufficiently small and the samples will be drawn from the posterior distribution (Box 1.4). Because of the potential for dependence on the initial values, their possible influence is

Box 1.4

The burn-in when sampling from Markov chains

It can take thousands of iterations for some Markov chains to converge to the posterior distribution, while others converge immediately. Therefore, it is necessary to check convergence, and discard the initial samples from a Markov chain until convergence is achieved. These discarded values are referred to as a 'burn-in'.

There are several ways to check for convergence. One of the simplest is to plot the sampled values versus the iteration number. In the example in Fig 1.3, the initial value is approximately 1200, changing to values in the approximate range 100 to 400 after five samples. The values continue to be around 100 to 400 indefinitely, suggesting that the chain has reached what is known as its stationary distribution. The Markov chain is constructed in such a way for Bayesian analyses that this stationary distribution is the posterior distribution (Appendix C).

A further check for stationarity is to initiate the Markov chain with a second set of initial values. The stationary distribution will be insensitive to the initial values. If two chains with different initial values converge, then it suggests that both chains have reached their stationary distribution. There are formal methods for checking the convergence of a pair of Markov chains, such as the Gelman-Rubin statistic (Brooks and Gelman, 1998), which compares the variation of the samples within chains and the variation of the samples when the chains are pooled. Initially, the pooled variation will be greater than the average variation of the two chains and then become equal as the chains converge. Additionally, the level of variation both within and between chains should stabilize with convergence.

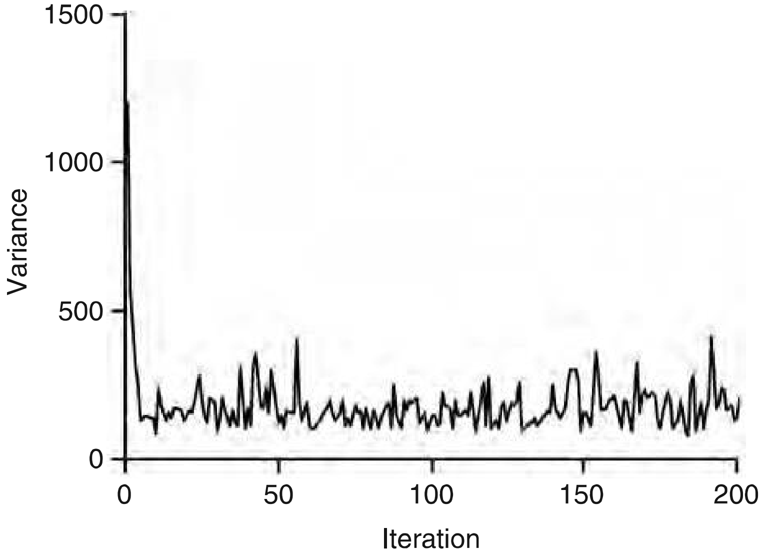

Fig. 1.3 The first 200 samples of the variance of the number of trees in a remnant for the model in Box 3.2.

examined and it may be necessary to discard some of the initial samples (perhaps the first few thousand or more) as a 'burn in' (Box 1.4).

A second consequence of any correlation is that, compared to an uncorrelated sample, each additional sample contains only a fraction of the information about the posterior distribution. Because of this, a large number of samples may be required to obtain a sufficiently precise sample if there is strong correlation between samples. Although the presence of correlation in the Markov chain reduces the efficiency of the sampling algorithm, it does not preclude the use of Markov chain methods. The reduced efficiency is simply the cost to be paid when it is not possible to obtain an analytical solution for the posterior distribution. Gilks *et al.* (1996) provides further information about Markov chain Monte Carlo methods.

The frog surveying problem in WinBUGS

Code for analysing the frog surveying problem in WinBUGS is given in Box 1.5. A Bayesian model specified in WinBUGS has the four components of a Bayesian analysis:

- prior distributions for the parameters being estimated;
- data;
- a model that relates the parameters to the data; and
- the posterior distributions for the parameters.

Box 1.5

WinBUGS code for determining the presence of a species

The frog surveying problem involves determining whether the species is present at a site given that it was not detected during a survey. In WinBUGS, the code works by specifying the prior for the probability of presence and the model, which describes how the parameter of interest (the presence of the frog) is related to the data. Pseudo-code for this problem would be:

1. Specify the prior probability of presence;
2. Specify that the frog is either present or absent with a particular probability;
3. Calculate the probability of detecting the species given that it is either present (probability of detection = 0.8) or absent (probability of detection = 0.0);
4. Cpecify that the detection of the frog or failure to detect the frog (the data) arises randomly, depending on the probability of detection.

Steps 1–2 specify the prior for the presence of the frog. Steps 3–4 specify the model, describing how the data (the observation of an absence in this case) are related to the presence of the frog, which is the parameter being estimated.

The WinBUGS code for the frog surveying problem is written below.

```
model
{
  prior <- 0.5                         # the prior
                                         probability of
                                         presence
  present ~ dbern(prior)               # actual presence
                                         drawn from a
                                         Bernoulli
                                         dist'n
  prob_detect <- 0.8*present           # prob of
                                         detection depends
                                         on
                                         presence/absence
  detected ~ dbern(prob_detect) # actual detection
                                  occurs with
                                  random variation
}
list(detected = 0)              # the data - the frog
                                 was not detected
```

In this model we are interested in determining whether the frog is present (represented by the variable `present`). The variable prior is the prior probability of the frog being present. The prior probability of the frog being absent is therefore 1−`prior`. The actual presence at the site is determined randomly, by drawing from a Bernoulli distribution; a value of one indicates the frog is present and zero indicates the frog is absent. Therefore, the first two lines define the expected presence of the frog prior to the collection of the data.

The next two lines describe the model of how the data were collected. If present, the probability of detecting the frog (`prob_detect`) is equal to 0.8, and it will equal zero if it is absent. The fourth line then states that the data are assumed to occur randomly, again drawn from a Bernoulli distribution, with the probability of detecting the frog on a single visit being equal to `prob_detect`, and the probability of not detecting the frog being equal to `1−prob_detect`.

The observed data (written in the line `list(detected = 0)`) then influence the values of the variable `present`, through the application of Bayes' rule within WinBUGS (Box 1.3). Values of the variable `present` are sampled by WinBUGS such that they are drawn as random samples from its posterior distribution. Sampling in this way is called Monte Carlo sampling. It is a relatively common method of analysing probabilistic models (Box 1.6). If enough samples are taken, the probability of the frog being present can be estimated by the proportion of times that the variable `present` equals one. This proportion equals the mean of the variable `present`.

Sampling 100 000 times from this model in WinBUGS (after ignoring the first 10 000 samples) leads to a mean value of `present` of 0.17, which is equivalent to 1/6, as determined analytically. This is our estimate of the posterior probability that the site is occupied given the prior and the data. Changing the value of `prior` to 0.75 leads to a mean value of `present` that is equal to 0.38 (again based on 100 000 samples), which is equivalent to 3/8 as determined analytically.

The results in WinBUGS are not exact because of random sampling error. If we took more samples in WinBUGS, the results would be closer to the truth. For example, the posterior probability of presence equals 0.3754 if half a million samples are taken when the prior for this value is 0.75. It is not precisely the same as the true answer (0.3750), but the answer in WinBUGS will continue to become more precise as more samples are taken.

Box 1.6
Monte Carlo methods

Monte Carlo methods use simulation to estimate the probability of occurrence of uncertain events. For example, consider a five-card poker hand. We could use probability theory to work out the chance of obtaining a flush (five cards of the same suit). The probability is equal to the probability that the second, third, fourth and fifth cards are the same suit as the first. For a 52-card deck, this is equal to:

$$(12/51) \times (11/50) \times (10/49) \times (9/48) = 0.00033$$

We could also work out this probability with a Monte Carlo method by dealing, shuffling, and re-dealing and calculating the proportion of times that a flush appears. If we did this ten times, we might get one flush (if we were lucky). Based on these results (one occurrence out of ten deals), we might estimate that the probability of a flush is 0.1. This is an imprecise estimate. Obtaining more samples increases the precision. If we dealt the cards 10 000 times, we might get three flushes, implying that the probability of a flush is 0.0003. This is better, but still not perfect; we could deal the cards several million times and get an even more precise estimate of the probability.

Of course, it is laborious to deal the cards that many times. An efficient alternative might be for a machine to deal the cards for us. Such a task might be suitable for computers, because they specialize in repetitive tasks. However, instead of dealing a physical deck of cards, the computer could use its circuitry to generate ‘random’ numbers that have the same statistical properties as the cards. In this case, thousands of samples can be generated very quickly by randomly generating integers between 1 and 52 (representing the 52 possible cards) with equal probability.

This virtual random sampling is the same sort of process that is used by WinBUGS. It generates samples that have the same statistical properties as the posterior distribution. The samples generated by WinBUGS can then be analysed to estimate the statistical properties of the posterior distribution such as its mean and percentiles.

WinBUGS code includes the prior for the parameters, but most of the code is usually the model, which describes how the data are related to the parameters. The posterior is then generated by WinBUGS with Monte Carlo sampling (Box 1.6; Appendix C).

The advantage of using a Monte Carlo approach is that it is able to sample from the posterior distribution without analysts having to do the various calculations themselves. In the frog surveying problem, the calculations done by hand are relatively easy. In the few cases where the calculations can be done by hand, they are usually more difficult, and in most other cases they are impossible.

Monte Carlo methods have another appealing property. Even relatively complex statistical analyses (e.g. regression analysis) do not require WinBUGS code that is much more complex than that presented in Box 1.5. Once familiar with relatively simple analyses, it is not much more difficult to write code for more complex analyses.

Example 2: Estimation of a mean

The second example of Bayesian analysis involves estimating the average diameter of trees in a remnant patch of eucalypt forest (Harper *et al.*, 2005). The size of trees is important when studying, for example, nutrient dynamics, provision of habitat for animals, production of nectar, mitigation of temperature extremes, and amelioration of pollution (Bormann and Likens, 1979; Attiwill and Leeper, 1987; Huang, 1987; McPherson *et al.*, 1998; Brack, 2002; Gibbons and Lindenmayer, 2002; Brereton *et al.*, 2004).

The mean diameter of trees could conceivably take any value between zero and some large number. Therefore, the hypotheses are not discrete. There are an infinite number of hypotheses, represented by any conceivable value for the mean diameter of the trees. Bayesian methods are able to accommodate these sorts of cases where hypotheses are distributed along a continuum, by using continuous rather than discrete probability distributions to represent uncertainty in the variables. The only modification to Bayes' rule is how the constant of proportionality is calculated (Box 1.7).

Assume that a researcher has measured the diameter of 10 randomly selected trees. In analysing the data, the researcher must choose the prior probability distribution for the parameters being estimated. Although the mean size of trees could be conceivably any positive number, the researcher has previously measured more than 2500 trees in 43 other

Box 1.7

Bayes' rule for continuous hypotheses

In the case of continuous hypotheses, continuous probability distributions are used to represent different possible values for parameters. Bayes' rule is then expressed as:

$$\Pr(H \mid D) = \frac{\Pr(H) \times \Pr(D \mid H)}{\int_0^\infty \Pr(x) \times \Pr(D \mid x) dx},$$

where H represents a particular value for the parameter. The integral in the denominator substitutes for the summation in the discrete case, and the limits of the integration are over all the possible values of the parameter (x), which in this case is assumed to be positive. This integral makes Bayesian methods difficult to conduct analytically, because in most cases it cannot be determined.

Readers who are uncomfortable with mathematics may look at the above equation and decide that they can never solve those sorts of problems and decide that Bayesian methods are too hard. The complexity of the equation should not be discouraging because in most cases it is impossible to solve, regardless of a person's mathematical skills. Fortunately, software is available so users do not need to evaluate or even construct the integral.

As with the case when there were a finite number of hypotheses (Box 1.3), the denominator simply acts as a scaling constant, because it is the same for all possible values for the parameter H. As with discrete hypotheses, the posterior probability is simply proportional to the prior probability ($\Pr(H)$) multiplied by the likelihood ($\Pr(D \mid H)$). The main analytical task of Bayesian analyses is to determine the constant of proportionality.

remnants (Harper *et al.*, 2005). After measuring so many trees in the study area, he has a good idea about likely values for the mean diameter of trees in the previously unmeasured remnant. Frequentist analyses do not permit this additional information to be used in determining the mean diameter of trees in the new remnant, but a Bayesian analysis does.

Based on data from the other 43 remnants, the mean diameter of trees in remnants is 53 cm and the mean varies among remnants with a standard deviation of approximately 5 cm. Assuming the mean diameter of trees follows a normal distribution, we would expect approximately 95% of remnants to have a mean tree diameter that is within 1.96 standard deviations of the overall average. Therefore, prior to collecting the data there is a 95% chance that the mean diameter of trees in the new remnant will be between approximately 43 and 63 cm. This prior reflects the researchers' expectation of the mean size of trees in a newly measured remnant based on his previous experience in the study area. A plot of the prior shows the range of likely values (Fig. 1.4).

The Bayesian solution for the normal mean

In the simplest case, and to make the analysis comparable to a traditional frequentist analysis, we will assume that the diameter of trees within the

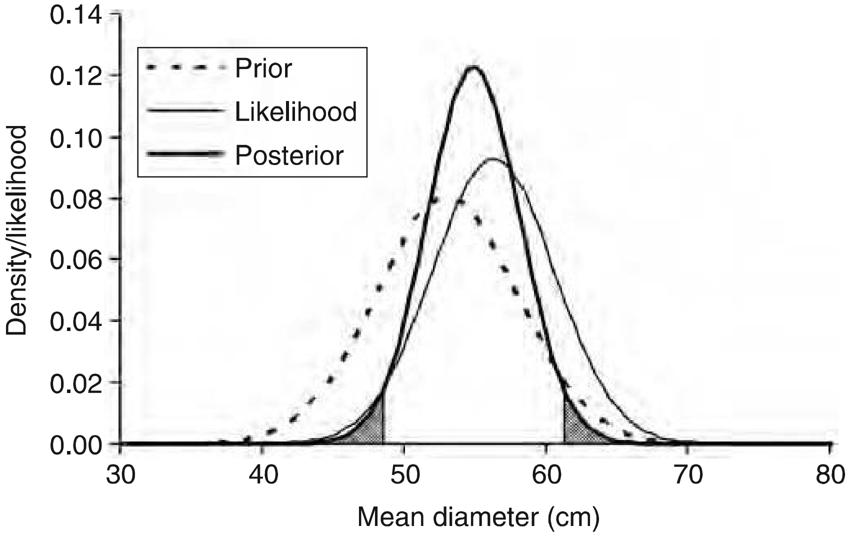

Fig. 1.4 The prior and posterior density functions and likelihood for the mean diameter of trees in a remnant, based on a sample of ten trees. The posterior would equal the likelihood if the prior was uninformative. The posterior is more precise than both the prior and the likelihood function because the posterior combines the information in both. The limits of the 95% credible interval of the posterior have 2.5% of the area under the curve in each tail (shaded).

remnant follows a normal distribution. In the case where the data and the prior both have normal distributions, Bayes' rule (Box 1.7) provides an analytical solution for the posterior distribution. However, analytical solutions are available for only a handful of Bayesian models, so I will first illustrate this example using WinBUGS (Box 1.8). It is simply a matter of specifying a prior distribution for the mean of the diameter

Box 1.8
Estimating a mean for a normal model using WinBUGS

In estimating the mean diameter of trees, the prior has a mean of 53 cm and a standard deviation of 5 cm. In WinBUGS, the width of a normal distribution is expressed using the precision (1/variance = 1/sd^2), which in this case is equal to 0.04 (1/25 = 1/5^2).

In this example, the variance of the data is assumed to be known, making it equivalent to using a z-value rather than a t-value in a frequentist analysis. However, uncertainty in estimating the precision of the data can be included easily in the WinBUGS analysis (Chapter 3).

The pseudo-code for the WinBUGS analysis is:

1. Specify the prior for the mean diameter of trees in the remnant as being normally distributed with a mean of 53 and precision of 0.04 (standard deviation of 5).
2. Calculate the standard deviation of the data.
3. Specify the precision of the data as the inverse of the variance of the diameter of trees in the remnant (the variance equals 184.9 in this example).
4. For each of the ten trees that were measured, assume that their diameter is drawn from a normal distribution with the mean and precision as specified in steps 1 and 3.

The WinBUGS code is:

```
model
{
  m ~ dnorm(53, 0.04)       # prior for mean
  stdev <- sd(Y[])          # calculate std deviation
                              of data
  prec <- 1/(stdev*stdev) # precision of the data =
                            1/variance
  for (i in 1:10)         # for each of the ten
                            trees ...
  {
    Y[i] ~ dnorm(m, prec) # diameter drawn from a
                            normal distribution
  }
}
list(Y = c(42, 43, 58, 70, 47, 51, 85, 63, 58, 46))
```

The ‘for loop’, designated by the line `for (i in 1:10)` and subsequent line within the curly brackets, is equivalent to ten lines of code, one for each of the ten trees, i.e. `Y[1] ~ dnorm(m, prec)`, up to `Y[10] ~ dnorm(m, prec)`. It is shorthand to replace repetitive sections of code.

The data are provided in the line:

```
list(Y = c(42, 43, 58, 70, 47, 51, 85, 63, 58, 46))
```

The ‘c’ before the brackets indicates that the following data are concatenated (linked together) into the one variable, with the first variable represented by `Y[1]`, the second by `Y[2]`, etc.

This analysis also requires that the user specifies an initial value of `mean` for the Markov chain (Box 1.4). The choice is not important because the chain converges quickly to the posterior distribution in this case, and could be generated randomly. However, in some cases the speed of convergence is increased if the Markov chain is initiated with values that are close to the posterior distribution, so the following arbitrary value was used:

```
list(m = 55)       # an arbitrary initial value
```

After discarding the first 1000 samples as a burn-in, 100 000 samples were generated in WinBUGS. Of these samples, 2.5% were less than 48.5 and 2.5% were more than 61.3. Therefore the 95% credible interval is 48.5–61.3 cm for the mean diameter of trees in the remnant. This result is insensitive to the choice of the initial value.

of trees in the remnant, and then constructing a model in which the measured diameters are drawn from a distribution with that mean. The posterior distribution calculated in WinBUGS is the same as that obtained using the analytical solution (Box 1.9).

Confidence intervals and credible intervals

A frequentist analysis would ignore the prior information and simply use the mean of the data and the standard error ($=\sqrt{(184.9/10)} = 4.3$),

Box 1.9

Estimating a mean for a normal model analytically

When the data and prior have normal distributions, the posterior distribution also has a normal distribution, the mean and variance of which depends, not surprisingly, on the mean and variance of the prior. The posterior distribution also depends on the sample size, mean and variance of the data. The mean and variance of the posterior can be calculated from the following formulae (Gelman *et al.* 2004):

$$\mu_{\text{post}} = \frac{\mu_{\text{prior}}/\sigma^2_{\text{prior}} + \mu_{\text{data}}n/\sigma^2_{\text{data}}}{1/\sigma^2_{\text{prior}} + n/\sigma^2_{\text{data}}}, \text{ and}$$

$$\sigma^2_{\text{post}} = \frac{\sigma^2_{\text{prior}}\sigma^2_{\text{data}}/n}{\sigma^2_{\text{data}}/n + \sigma^2_{\text{prior}}},$$

where n is equal to the sample size, σ^2_{prior}, σ^2_{data}, and σ^2_{prior} are the variances of the prior, data and posterior, and μ_{prior}, μ_{data} and μ_{post} are the means of the distributions.

These formulae provide useful insights into Bayesian statistics. The mean of the posterior is a weighted average of the means of the prior and data. The weights are the precisions of the prior ($1/\sigma^2_{\text{prior}}$) and the data ($n/\sigma^2_{\text{data}}$). The influence of the data and prior on the posterior mean depends on which is more informative. When there are no data ($n = 0$), the mean of the posterior is equal to the mean of the prior. When the variance of the prior is very large, $1/\sigma^2_{\text{prior}}$ approaches zero and the mean of the posterior will be close to the mean of the data. The prior is said to be uninformative when the posterior is influenced exclusively by the data. This is achieved by using a prior with a large variance.

The variance of the posterior has similar properties, but these are most obvious when its formula is re-arranged to be expressed as the inverse of the variance:

$$\frac{1}{\sigma^2_{\text{post}}} = \frac{n}{\sigma^2_{\text{data}}} + \frac{1}{\sigma^2_{\text{prior}}}$$

The inverse of the variance measures precision. Large values for the precision mean the variance is small. The quantity n/σ^2_{data} is the inverse of the standard error squared, and it measures precision in an ordinary frequentist analysis. Therefore, the precision of the posterior is simply equal to the precision based on the data (the inverse of the standard error squared) plus the precision of the prior. The precision of an estimate is increased by using prior information.

The diameter measurements of ten trees in the new remnant (42, 43, 58, 70, 47, 51, 85, 63, 58, 46 cm) have a mean of 56.3 and variance of 184.9. Given the prior has a mean and variance of 53 and 25, the posterior distribution for the mean diameter of trees in the new remnant has the following mean and variance:

$$\mu_{\text{post}} = \frac{53/25 + 56.3 \times 10/184.9}{1/25 + 10/184.9} = 54.9$$

$$\sigma^2_{\text{post}} = \frac{25 \times 184.9/10}{184.9/10 + 25} = 10.6$$

The standard deviation of the posterior is 3.26 cm ($\sqrt{}10.6$). Therefore, there is an approximate 95% chance that the mean diameter of trees in the park is between 48.5 cm and 61.3 cm (the mean of the posterior plus or minus 1.96 times the standard deviation of the posterior) (Fig. 1.4). This 95% credible interval is the same as that obtained from WinBUGS (Box 1.8).

leading to a 95% confidence interval of 47.9–64.7 cm ($56.3 \pm 1.96 \times 4.3$). This is the same as the credible interval that was obtained when using a Bayesian analysis with an uninformative prior (Box 1.10).

Bayesian credible intervals and frequentist confidence are usually numerically identical if the Bayesian prior is uninformative. An

Box 1.10

Estimating the mean of a normal model with an uninformative prior

An uninformative prior for the mean diameter of trees can be specified by using the following line of code for the prior instead of the one in Box 1.8:

```
mean ~ dnorm(0, 1.0E-6)     # wide prior for mean
```

This is a very wide normal distribution with a mean of zero and a standard deviation of 1000. Therefore, mean diameters between, for example, zero and 200 cm have approximately the same prior probability. If this uninformative prior is used, the posterior distribution for the mean diameter of trees in the remnant has a mean of 56.3 and 95% credible interval of 47.8–64.7, numerically equivalent to the 95% confidence interval of a frequentist analysis.

uninformative prior is one in which the data (via the likelihood, which is $\Pr(D \mid H)$ in Bayes' rule) dominates the posterior. This is achieved by using a prior with a large variance. A large variance permits the parameter to be drawn from a wide range of possible values and the prior probabilities of all reasonable parameter values are approximately equal. When the prior distribution is uninformative, the posterior distribution has the same form as the likelihood (Fig. 1.5). The likelihood and posterior have different forms when the prior is informative (Fig. 1.4).

The posterior distribution is less precise, and hence the credible interval is wider, if the prior information is ignored (Fig. 1.5). Ignoring the prior information would imply that the researcher believed that the remnant could have any mean diameter prior to collecting the data. Such a belief would be inconsistent with the researchers' previous experience in the study area, which provides useful data on the range of likely results.

Although frequentist confidence intervals and Bayesian credible intervals may appear similar, they are in fact different. For a 95% Bayesian credible interval, there is a 95% chance that the true value of the parameter will be within the interval. Ecologists are often interested in this kind of interval because they want to know the chance that the true value of the parameter is within a specified range. Such an answer requires the use of Bayesian credible intervals.

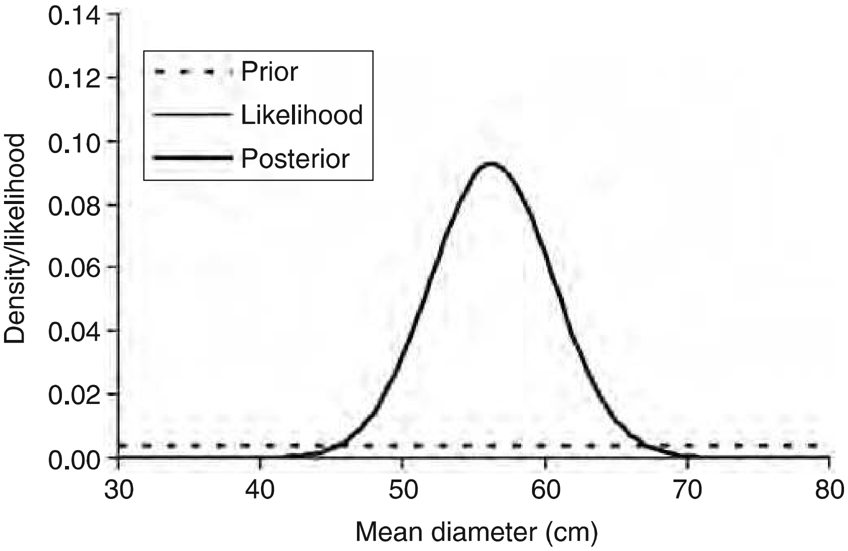

Fig. 1.5 The prior and posterior density functions and likelihood for the mean diameter of trees in a remnant, based on a sample of ten trees and using an uninformative prior. The prior distribution (drawn on an arbitrary scale to assist comparison) has a mean of zero and standard deviation of 1000 making mean diameters between 30 and 80 cm all equally likely a priori. The likelihood and the posterior are indistinguishable. The posterior is less precise than in Fig. 1.4 because the prior is uninformative.

In contrast, a 95% frequentist confidence interval does not contain the true parameter with 95% probability. Instead, it is based on the concept of an infinite number of samples. If I repeat the data collection an infinitely large number of times and construct 95% confidence intervals for the mean for each set of data, 95% of these confidence intervals would encompass the true mean.

This different meaning of confidence and credible intervals is not just semantic. In some circumstances, it can lead to numerical differences even when the credible interval is based on an uninformative prior. For example, in estimating a fail-safe period from three observations of failure times (12, 14 and 16), Jaynes (1976) shows that the shortest possible 90% confidence interval is 12.1–13.8. This interval does not contain the true fail-safe period, which must be less than the smallest observed lifespan (12). This result is not an error. The method of calculating 90% confidence intervals will produce intervals that enclose the true value of the parameter 90% of time. However, the true value might surely lie outside any single interval, as in this example.

In contrast, the Bayesian analysis with an uninformative prior arrives at a sensible conclusion; the shortest possible 90% credible interval is 11.2–12.0 (Jaynes, 1976). When the intervals are the same, the choice of

Bayesian or frequentist methods does not matter. However, when the intervals are different, only Bayesian methods provide logical results (Jaynes, 1976).

Concluding remarks

In introducing Bayesian methods, this chapter made two important points. Firstly, Bayesian methods can answer questions that are relevant to ecologists, such as: 'What is the probability that this hypothesis is true?' and 'What is the probability that a parameter will take values within a specified interval?' Secondly, relevant prior information can also be incorporated into Bayesian analyses to improve the precision of estimates.

Bayes' rule is the basis of Bayesian methods. It is derived as a simple expression of conditional probability. The rule specifies how prior information and data are combined using a model to arrive at the posterior state of knowledge. Both the prior and posterior states of knowledge are represented as probability distributions. The posterior probability simply equals the prior probability multiplied by the likelihood of the data and a scaling constant. Bayesian methods become difficult because the scaling constant is usually hard to calculate analytically. However, recent numerical methods such as Markov chain Monte Carlo make Bayesian methods accessible to all scientists.

Frequentist confidence intervals and Bayesian credible intervals will usually be numerically equivalent if uninformative priors are used. In this way Bayesian methods provide a numerical generalization of frequentist methods. They also do so in such a way that probabilistic statements about the state of nature are mathematically logical. The next chapter provides a more thorough comparison of different statistical schools and examines their various strengths and weaknesses.

2
Critiques of statistical methods

Introduction

Statistics in the discipline of ecology is dominated by null hypothesis significance testing. Apart from the construction of confidence intervals, it is almost the only statistical method taught in ecology at the undergraduate level. In leading ecological and conservation journals, such as *Conservation Biology*, *Biological Conservation*, *Ecology* and the *Journal of Wildlife Management*, null hypothesis testing has been used in approximately 90% of articles between 1978 and 2001 (Anderson *et al.*, 2000; Fidler *et al.*, 2004), although this proportion was only 80% in 2005 (Fidler, 2005). Since 1980, there have been several thousand null hypothesis tests (on average) reported each year in *Ecology* (Anderson *et al.*, 2000), a further illustration of the dominance of this method. In comparison, only about 5% of ecological articles refer to Bayesian methods and even fewer use them (Fig. 1.1).

Despite its dominance, null hypothesis significance testing has ardent critics. There are alternatives but their use is controversial. In this chapter, I review three different methods of statistical analysis that are used in ecology (see also Oakes, 1986). These are null hypothesis significance testing, information-theoretic methods, and Bayesian methods. Readers will not be surprised, given the topic of this book, that I believe there are clear advantages in using Bayesian methods, although not necessarily to the total exclusion of others. However, I will first present an example that further illustrates some of the differences and similarities of the statistical methods.

Sex ratio of koalas

The following example illustrates how results of Bayesian and likelihood-based methods can differ from those obtained using null hypothesis significance testing.[1] Consider a researcher who is studying the population ecology of koalas (*Phascolarctos cinereus*) with a particular interest in the sex ratio of pouch young of mothers in poor physical condition. Assume that the researcher samples 12 female koalas in poor condition each with an offspring in its pouch (pouch young). Three of the offspring are male and nine are female. Based on this study, what can we say about the sex ratio of koalas produced by females in poor physical condition?

Null hypothesis significance testing

A reasonable null hypothesis in this case might be that the number of male and female offspring would be equal. Under this hypothesis, the sex ratio (the proportion of males in the population of pouch young) would be 0.5. However, Trivers and Willard (1973) suggest that a female-biased sex ratio would be expected in animals with poor physical condition. Thus, a reasonable alternative hypothesis is that the sex ratio is less than 0.5.

The data could have been obtained in at least two ways. Firstly, the researcher could have decided to sample 12 koalas with offspring, in which case her data are the number of males (three males). Alternatively, she may have sampled koalas until three males had been obtained, in which case her data would be the number of female koalas until the third male was encountered (nine females).[2] Regardless of the sampling strategy, the data are equivalent (three males and nine females); the only difference is the stopping rule for her sampling strategy (sample until 12 individuals, or sample until three males are obtained).

The null hypothesis tests under these two stopping rules are described in Box 2.1. Under the first stopping rule the *p*-value is 0.073, so using the

[1] This example is based on a thought experiment conducted by Lindley and Phillips (1976). It has been modified from tossing of coins for an ecological audience (Johnson, 1999).

[2] This second sampling strategy may seem odd at first, but it may occur in reality when a researcher has multiple questions. For example, the study might be mainly based on researching male offspring, with the calculation of the sex ratio as a secondary interest. Additionally, the sample sizes may be regarded by some as very small (three males or 12 offspring), but I chose them for illustration because this simplifies the mathematics.

Box 2.1

Null hypothesis tests for a proportion

In the first case, 12 offspring are sampled so that the number of males can take any number between zero and 12. If the sex ratio (the proportion of the population of pouch young that are male) were *r*, the chance of getting 0 males would equal the probability of the first being male (*r*), multiplied by the probability of the second being male (*r*), etc. Thus, the probability of all 12 koalas being male equals r^{12}.

The probability of one koala being male and the other 11 being female is equal to the probability that the first is male and all the others are female $r(1\text{-}r)^{11}$, plus the probability that the second is male and all others are female $r(1\text{-}r)^{11}$, etc. Thus, the probability that only one of the 12 is male is equal to $12r(1\text{-}r)^{11}$.

It turns out that the sampling in this case can be described by the binomial distribution (Appendix B, Johnson *et al.*, 1992, Fowler *et al.*, 1998), which states that the probability of there being *x* males in a sample of 12 is given by:

$$\Pr(\text{males} = x) = \frac{12!}{(12 - x)!x!} r^x (1 - r)^{12-x},$$

where $x!$ ('*x* factorial') equals $1 \times 2 \times 3 \times \ldots \times x$, with $0! = 1$.

The *p*-value for the null hypothesis $r = 0.5$ is the probability of getting three males or a more extreme result (in this case, fewer than three males) from a sample of 12. Thus:

$$P_1 = \Pr(\text{males} = 3) + \Pr(\text{males} = 2) + \Pr(\text{males} = 1) + \Pr(\text{males} = 0).$$

After substituting the binomial probabilities, one obtains $P_1 = 0.073$.

Thus, using the conventional 'cut-off' (type I error rate) of 0.05, we would conclude that the sex ratio *is not* significantly (in a frequentist sense) less than 0.5.

What if the researcher used the different sampling strategy in which she records the sex of koala pouch young until three males have been sampled? The probability of there being no females in the sample is equal to the probability that the first three pouch young are male (r^3).

One female will be sampled if one of the first three pouch young is female while the fourth is male at which point sampling will cease. Thus, the probability that there is one female in the sample is equal to the probability that two of the first three pouch young are male (and one is female) ($= (3!/1! \times 2!)r^2(1-r)$ from the binomial distribution) multiplied by the probability that the fourth is also male (r), leading to $3r^3(1-r)$.

The probability that there are two females in the sample is equal to the probability that two of the first four pouch young are male (and two are female) ($= (4!/2! \times 2!)r^2(1-r)^2$) multiplied by probability that the fifth is male (r), leading to $6r^3(1-r)^2$. This can be continued indefinitely for any number of females.

More generally, the number of females until the *i*th (in this case third) male is described by the negative binomial distribution (Appendix B; Johnson *et al.*, 1992; Fowler *et al.*, 1998). Its probabilities are given by:

$$\Pr(\text{females} = x) = \frac{(3 + x - 1)!}{x!2!} r^3(1-r)^x.$$

Using this sampling design, the *p*-value is equal to the probability of sampling nine or more females before the third male, and it is given by:

$$P_2 = \Pr(\text{females} = 9) + \Pr(\text{females} = 10) + \Pr(\text{females} = 11) + \cdots$$

The series continues indefinitely (until 'females' equals infinity), because there is the (small) possibility that three males will not be sampled even after sampling many animals. Substituting the negative binomial probabilities into the above equation leads to $P_2 = 0.033$.

If we again used the conventional type I error rate of 0.05, we would conclude that the sex ratio is significantly (in a frequentist sense) less than 0.5.

usual type-I error rate of 0.05 we would not reject the null hypothesis that the sex ratio is equal to 0.5. The *p*-value is 0.033 under the second stopping rule, so we would accept the alternative hypothesis that the sex ratio is less than 0.5.

The two different stopping rules for the sampling strategies lead to different conclusions about the null hypothesis, even though the actual data are identical. Therefore, our decision about the sex ratio of koala pouch young would be determined not only by the data we collected but by how we decided to stop sampling. The difference occurs because the null hypothesis test depends on the data *and more extreme* (but unobserved) values. The stopping rule does not influence the results if the data are analysed using Bayesian or information-theoretic methods because their results are not conditioned on unobserved data (Lindley and Phillips, 1976; Berger and Berry, 1988). I will illustrate the information-theoretic and Bayesian solutions below.

Information-theoretic methods

Information-theoretic methods use maximum likelihood estimation to determine parameter values. Maximum likelihood methods can estimate the sex ratio and place confidence intervals around the estimate (Edwards, 1992; Hilborn and Mangel, 1997). Maximum likelihood methods are so named because the best estimate is the one for which the probability of obtaining the observed data is maximized. Under the binomial model, the probability of obtaining the observed data (three males) is:

$$\Pr(\text{males} = x) = \frac{12!}{(12 - x)!x!} r^x (1 - r)^{12-x} = \frac{12!}{9!3!} r^3 (1 - r)^9.$$

This expression is maximized when the term $L = r^3(1-r)^9$ is maximized. This occurs when the sex ratio r is equal to 0.25, for which $L_{\max} = 0.001173$.

Under the negative binomial model, the probability of obtaining the observed data (nine females) is:

$$\Pr(\text{females} = x) = \frac{(3 + x - 1)!}{x!2!} r^3 (1 - r)^x = \frac{11!}{9!2!} r^3 (1 - r)^9.$$

This is also maximized when the term $r^3(1-r)^9$ is maximized, illustrating that methods based on maximum likelihood are not influenced by the stopping rule.

Approximate confidence intervals can be placed on the maximum likelihood estimate of the sex ratio r by finding values of r such that L is equal to $L_{\max}\exp(-\chi/2)$, where χ is a value from the appropriate chi-squared (χ^2) distribution (Edwards, 1992; Hilborn and Mangel, 1997). For a 95% confidence interval $\chi^2 = 3.84$, which corresponds to a tail probability of 0.05 for a chi-squared distribution with one degree of freedom. Values of r for which L is equal to $L_{\max}\exp(-3.84/2) = 0.000172$ are 0.069 and 0.528. These define the limits of the 95% confidence interval for the sex ratio.

The Bayesian method

This problem can be analysed using a Bayesian method in WinBUGS (Box 2.2). Regardless of the stopping rule that is used, the probability density function for the sex ratio of offspring is the same. Therefore, only the data (and our prior) influences the estimate of the sex ratio of pouch young in koalas, not the choice of when to stop sampling. Unlike null hypothesis testing, Bayesian methods are based only on the observed data not unobserved (more extreme) values.

The mean of the posterior distribution for the sex ratio is 0.286, and the 95% credible interval is 0.091–0.537. Thus, it is likely that the sex ratio is less than 0.5, but it may in fact not be. Note that the 95% credible interval is similar to the confidence interval constructed with maximum likelihood estimation, but is different because the chi-squared value in the likelihood method requires a large-sample approximation.

This example illustrates that null hypothesis significance testing and Bayesian methods can lead to different conclusions. Additionally, when an uninformative prior is used, estimates based on Bayesian and information-theoretic methods are similar (see also Chapter 1). However, informative priors increase the precision of Bayesian estimates (e.g. Fig. 1.4). The strengths and weakness of the different statistical methods are described in more detail in the following sections.

Null hypothesis significance testing

Null hypothesis testing works in a series of steps.

1. A null hypothesis is defined, along with a single alternative hypothesis.
2. Data are collected.
3. The analyst calculates the probability of collecting the data or more extreme data given that the null hypothesis is true. This probability is the *p*-value.
4. If this probability is sufficiently small, then the analyst concludes that the data are unusual given the null hypothesis. The almost universal convention is to use an arbitrary cut-off of 0.05. If the *p*-value is less than 0.05, then the analyst concludes that the null hypothesis is unlikely to be true, and the alternative hypothesis is accepted.

Box 2.2
Bayesian analysis of a proportion

The first sampling strategy (sample 12 koalas) can be analysed from a Bayesian perspective in WinBUGS with the following code:

```
model
{
  x ~ dbin(r, 12)  # data sampled binomially with n = 12
  r ~ dunif(0, 1)# prior for the sex ratio of pouch
                   young
}
list(x = 3)        # 3 males sampled
```

The data are given by x, and r is the sex ratio being estimated. We assume that the data are drawn from a binomial distribution with a sample size of 12 (see Box 2.1). For simplicity, I have chosen a uniform distribution for the sex ratio as an uninformative prior. This ignores the fact that a sex ratio equal to zero or one is very unlikely to occur in any mammal species. I could use data on the sex ratio of offspring in other mammals or other koala populations to generate a more reasonable prior.

Sampling 100 000 times in WinBUGS (after discarding the first 10 000 samples) provides the posterior distribution (Fig. 2.1). The mode of the distribution is 0.25 and the median is 0.275. The mean of the distribution is 0.286, with a 95% credible interval of 0.091–0.537. The posterior distribution indicates that the data are not entirely inconsistent with a sex ratio of 0.5, but it is likely that the sex ratio is less than 0.5, in accordance with the Trivers and Willard (1973) hypothesis.

The second sampling strategy (sample until three males have been recorded) can also be implemented in WinBUGS. In this case the value for the number of females before the third male is encountered is drawn from a negative binomial distribution (see Box 2.1 and Appendix B). The WinBUGS code is:

```
model
{
  x ~ dnegbin(r, 3)# number of females sampled neg.
                     binomially
  r ~ dunif(0, 1)# prior for the sex ratio of pouch
                   young
}
list(x = 9)       # 9 females sampled before the 3rd
                   male
```

Again, sampling from WinBUGS provides the posterior distribution. In the negative binomial case, the result is identical to the sampling strategy that used the binomial model (Fig. 2.1). Therefore, only the data (and our prior) would influence the estimate of the sex ratio of pouch young in koalas, not the choice of how to stop sampling.

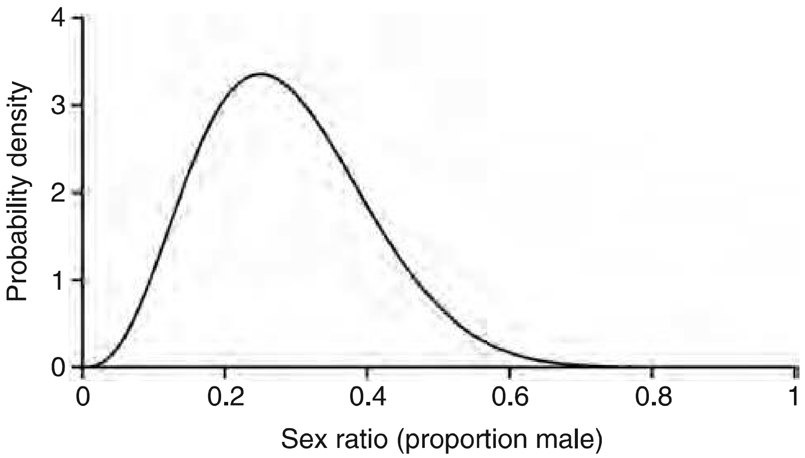

Fig. 2.1 Posterior probability density function for the sex ratio of koalas, based on a sample of three males and nine females and a uniform prior between 0 and 1.

5. If the probability is not below the critical level, then the analyst fails to reject the null hypothesis. By implication, the null hypothesis is 'accepted', but it is not proved because null hypothesis significance testing can only falsify hypotheses.

Define the null hypothesis and its alternative

The null hypothesis is a statement about the state of the system, often expressed in terms of parameter values. For example, an arbitrarily chosen null hypothesis[3] is that the Shannon-Weiner index of plant species diversity is the same in salt, brackish and fresh water marshes (Mullan Crain *et al.*, 2004). An alternative hypothesis is also chosen, which will be accepted if the null is rejected. In this example, the alternative hypothesis is that the plant species diversity is different in marshes of different salinity.

The choice of a useful null hypothesis is important. Ideally, the null hypothesis should be such that its rejection will have important logical consequences that lead to better ecological understanding (Underwood, 1997). However, ecologists routinely use nil nulls (predicting no effect or no difference) that are very unlikely to be correct (Johnson, 1995; Anderson *et al.*, 2000). These hypotheses are also referred to as false or trivial null hypotheses, or silly nulls (Stephens *et al.*, 2005). Anderson *et al.* (2000) reported that 90% of ecological studies use silly nulls.

Silly nulls take forms such as 'the survival of juveniles and adults is the same', 'there is no relationship between two variables of interest', or 'the growth rate of individuals is the same' (Anderson *et al.*, 2000). While studies that include silly nulls can provide useful scientific information (e.g. by demonstrating the size of effects), the rejection of a trivial null hypothesis is largely worthless because it was not a reasonable proposition in the first place.

The above null hypothesis of Mullan Crain *et al.* (2004) could be viewed as trivial. A priori we would expect that the diversity index for plants in marshes will vary with the salt content of water. As ecologists, we know fresh and salt water marshes would contain different plant species, and will therefore almost certainly have different diversity indices. A fundamentally more interesting question might be about how the diversity index changes across the salt gradient. Mullan Crain *et al.* (2004) do address this, but as it is not concerned with null hypothesis testing I will return to it later.

Why do ecologists use nil nulls so frequently when their rejection is usually uninformative? Why do we bother trying to reject literally thousands if not millions of hypotheses each year that are probably false? Perhaps because it is difficult to construct null hypotheses with a non-zero effect. For example, in the study of Mullan Crain *et al.* (2004), an alternative null hypothesis is difficult to formulate a priori. Although we can be reasonably sure that a difference would be expected, it is difficult to specify a precise prediction for a non-nil hypothesis. A theory

[3] This example was chosen as the first null hypothesis encountered after a randomly selected page of the 2004 volume of the journal *Ecology*.

that could predict species diversity of marshes as a function of the salt content of the water would provide a reasonable null hypothesis. Rejection of the null in this case would be very interesting, because it would tell us that the theory is lacking.

Such falsification of a well-reasoned hypothesis is a potentially powerful aspect of null hypothesis testing. However, ecological theory is not sufficiently precise that exact null hypotheses (other than nil nulls) can be constructed routinely. There are some exceptions, such as allometric models that predict particular scaling exponents (West *et al.*, 1997). However, the prevalence of nil nulls suggests that similarly precise predictions for non-nil nulls are rare in ecology. Although there is a large amount of data available to ecologists, such data can at best be used to make uncertain (probabilistic) predictions. Null hypothesis testing, like other frequentist methods, is not suitable for evaluating predictions that are imprecise.

Ecologists may also use null hypothesis testing through an adherence to Popperian falsification (e.g. Underwood, 1997). However, the rejection of a trivial null hypothesis fails to meet Popperian, or any other well-known philosophical criteria for good scientific practice. Further, Popperian falsification can be achieved without null hypothesis significance testing. If null hypothesis testing is to be used successfully, ecologists need to use logical null hypotheses. The evidence demonstrates that this does not occur despite continued criticism of the use of null hypothesis testing in ecology (Johnson, 1995; Anderson *et al.*, 2000; Fidler *et al.*, 2006).

Collect data

There is little that is controversial when it comes to collecting data for null hypothesis testing. It is assumed that the subjects, quadrats or other units of sampling are selected at random, while accounting for any underlying stratification or structure in the data during the analysis. Similar or identical assumptions apply to any statistical method. Readers should refer to literature on experimental design for further information (Underwood, 1997; Quinn and Keough, 2002).

Calculate the *p*-value

The *p*-value of null hypothesis testing is equal to the probability of obtaining the observed data *or more extreme data* if the null hypothesis is true. For example, consider the null hypothesis that the exponent of the scaling relationship between metabolic rate and body size is 0.75. Then, we collect some data on metabolic rate and body size and estimate the value of the exponent as 0.77, leading to a difference between the null hypothesis and the estimate of 0.02. However, given the variation expected in the data, a difference this large might be expected just by chance. The *p*-value is the probability of getting a difference this big or bigger if the null hypothesis is true.

Critics of null hypothesis testing ask: ‘Why should data that have never been observed (e.g. the occurrence of an exponent greater than 0.77) influence our inference about the validity of the null hypothesis?’ This seems to be a reasonable concern. It is easy to construct examples in which the observed data are impossible if the null hypothesis is true, but where the *p*-value is not zero because more extreme data are possible (e.g. a null hypothesis of an odd number of breeding birds in a monogamous species).

In practice, most null hypotheses predict unimodal distributions for the data, with the most common form being a normal distribution or a similar distribution derived from the normal (e.g. t or chi-squared distribution). As a result, there is usually a monotonic relationship between the probability of obtaining the observed data and the *p*-value. As the probability of observing the data increases, so too does the *p*-value. Therefore, the influence of the ‘unobserved results’ is usually small. For the example in Box 2.1, the different stopping rule led to different interpretations of what constituted more extreme data. The subsequent difference in the *p*-value was relatively small (0.033 versus 0.073), although large enough that the result of the hypothesis test was affected in this case.

Reject the null hypothesis if the *p*-value is small

A small *p*-value indicates that the observed data would be unlikely to occur if the null hypothesis were true. This then provides evidence against the null hypothesis, and it will be rejected and the alternative hypothesis accepted if the *p*-value is sufficiently small. The logic of this process is not entirely straightforward and is often misinterpreted. The most common misinterpretation is that the *p*-value is the probability of the null hypothesis being true, given the data. This misinterpretation is even shared by some of those who teach the method (Haller and Krauss, 2002). However, it is actually the converse of this; it is the probability of data (or more extreme data) given that the null hypothesis is true (Berger and Sellke, 1987; Ellison, 1996).

The distinction between the two probabilities can be illustrated with an example of probability with which readers will be familiar. Consider the null hypothesis that I am rolling a fair six-sided die. Let the observed data be that a value of one is rolled. The *p*-value for this outcome is 0.167 (1/6). Now consider the inverse of this problem: if a one is rolled, what is the probability that I am using a fair six-sided die? It is definitely not 0.167. The probability that I am using a fair die would depend on whether I own and use biased dice. Your belief in whether I am using a fair die has more to do with the perception of my character than the result of a single throw of the die.

However, if I continued to roll ones on subsequent throws, you would be rightly suspicious. The important point is that the probability of obtaining the data and the probability of a hypothesis being true are not the same, although there is a relationship between the two. This relationship is defined by Bayes' rule (Box 1.3).

Null hypotheses are routinely rejected when the *p*-value is less than an arbitrary value of 0.05. This choice has virtually no basis in logic. It is simply a number that ensures that correct null hypotheses would be rejected only 5% of the time in the long run. This rate of rejection of correct null hypotheses is the type I error rate. The type II error rate is the proportion of times that false null hypotheses would not be rejected. Power is equal to one minus the type II error rate. It is the proportion of times that false null hypotheses are rejected.

Ideally, the probability of making a poor decision with null hypothesis testing should decline to zero as the sample size increases. In fact, it would be possible to ensure this is the case if statistical power was considered by ensuring that both the type I and type II error rates were reduced towards zero as the sample size increased. However, by having a slavish adherence to the threshold of 0.05, ecologists set a limit such that even the largest studies will lead to erroneous conclusions about true null hypotheses 5% of the time.

When the null hypothesis is rejected, we accept the alternative hypothesis without explicitly considering how well it matches the data. Unless the alternative hypothesis is constructed with care and is a reasonable choice, we run the risk of accepting a hypothesis that is even more implausible (given the data) than the null we just rejected. Box 2.3 provides an example where the apparent rejection of a hypothesis and acceptance of an unlikely alternative has caused considerable trouble. The same data are analysed using information theoretic methods and Bayesian methods. Only Bayesian methods arrive at the correct conclusion in this case.

Box 2.3

Null hypotheses in the courts

Two sons of Sally Clark, a London lawyer, died while very young about a year apart and both in mysterious circumstances. In 1998, seven months after the second death, Sally Clark was charged with murder. She was eventually tried, found guilty, and sentenced to two life terms of imprisonment in 1999.

Part of the evidence presented in her trial was that the probability of two children dying of cot death in the one family was vanishingly small (quoted in court as one in 73 million). This is essentially a *p*-value: the probability of the obtaining the data (two children dying) given the null hypothesis (Sally Clark was innocent). Since this probability is so small, the null hypothesis of innocence could be rejected and the alternative hypothesis (that Sally Clark murdered her children) accepted. As claimed by the prosecutor, two cot deaths were 'beyond coincidence'. Of course, this acceptance of the alternative hypothesis ignores whether the available evidence supports it and whether it is reasonable in the first place.

The application of null hypothesis testing in this case gets it alarmingly wrong. Despite the vanishingly small *p*-value, evidence came to light that demonstrated that Sally Clark was unlikely to have killed her two sons, and after spending more than three years in prison she was released. A small *p*-value does not necessarily mean that the alternative hypothesis is true.

Fail to reject the null when the *p*-value is large

If the *p*-value is large, it would be nice to be able to conclude that the null hypothesis is true. Large *p*-values can occur if the null hypothesis is true or close enough to being true. However, they can also occur if the study is not sufficiently well-designed to have a reasonable chance of generating a low *p*-value if an important difference from the null hypothesis actually exists. Even a *p*-value of 1.0, which is the highest value it can possibly be, does not necessarily provide strong evidence that the null is true because large *p*-values can also be obtained if the null is false but the study is poorly designed.

The quality of a study is measured by its statistical power, and *p*-values need to be interpreted in its light. Power is the probability of obtaining a statistically significant result given that the null hypothesis is not true. Statistical power can help determine necessary sample sizes and assist the planning of data collection and subsequent analysis. The only problem is that this is rarely done in ecology. Power is almost never reported by ecologists, but in approximately half of all cases authors interpret their non-significant results as evidence that the null hypothesis is true (Peterman, 1990; Taylor and Gerrodette, 1993; Johnson, 1999; Anderson *et al*., 2000; Fidler *et al*., 2004). This is despite the fact that power must be known if we are to interpret the importance of non-significant results (Fidler *et al*., 2004).

When calculating power it is necessary to specify both the difference one wishes to detect and the variance of the data. Both values can be difficult to determine, but any calculation of power is conditional on the values that are used. Smaller differences may go undetected, and power will be less than expected if the variance is underestimated.

Summary of null hypothesis testing

There are several problems with the use of null hypothesis significance testing in ecology. These problems are mainly due to how the method is implemented, rather than the basis of the method. In summary, errors in the use of null hypothesis testing include:

1. using silly null hypotheses;
2. believing that the *p*-value is the probability that the null hypothesis is true;
3. interpreting large *p*-values as evidence that the null hypothesis is true (a sub-set of point 2);
4. ignoring statistical power (related to point 3);
5. following the almost universal convention to use a type I error rate of 0.05, despite power being ignored, so the type-II error rate is unknown; and
6. ignoring the size of effects being estimated and/or the evidence in favour of competing hypotheses when *p*-values are cited in results.

There are two problems with the actual basis of null hypothesis testing:

1. Data that were never observed or cannot be obtained influence the results (e.g. Box 2.1) because the *p*-value is based on data that

are more extreme than those observed, as well as the observed data; and

2. Evidence in support of the alternative hypothesis is ignored in the decision about whether to reject the null hypothesis in favour of the alternative.

In practice, these two problems need not have dire consequences for null hypothesis testing. If the null and alternative hypotheses are both reasonable (i.e. there has been logical and thoughtful development of the hypotheses), then the *p*-value provides a measure of the evidence in support of the two possible hypotheses, although it tends to overstate the evidence against the null (Berger and Sellke, 1987; see also Chapter 4). Despite warnings, silly nulls are common in ecology, and ecologists routinely ignore power while interpreting non-significant results as evidence that the null hypothesis is true. The same errors and efforts to correct them are repeated in other disciplines (Fidler *et al.*, 2004). The evidence suggests that null hypothesis testing is used poorly. Because of these repeated problems, there has been ongoing and ardent criticism of null hypothesis testing (Parkhurst, 1997; see also http://www.warnercnr.colostate.edu/∼anderson/null.html):

> Clark (1963) '... no longer a sound or fruitful basis for statistical investigation'
> Bakan (1966) '... essential mindlessness in the conduct of research.'
> Deming (1975) '... small wonder that students have trouble understanding hypothesis tests. They may be trying to think.'
> Carver (1978) '... significance testing should be eliminated; it is not only useless, it is also harmful...'
> Cohen (1994) '... hypothesis testing does not tell us what we want to know... out of desperation, we nevertheless believe that it does.'
> Rozeboom (1997) 'Null hypothesis significance testing is surely the most bone-headedly misguided procedure ever institutionalised in the rote training of scientists.'

I recommend that ecologists largely stop using it in favour of the methods discussed in the remainder of this chapter. One of these methods is Bayesian statistics. Given the problems with null hypothesis testing and its prevalence in ecology, one might ask how the discipline has managed to progress (Dennis, 1996). I believe part of the answer is that many ecologists do more than just null hypothesis testing when analysing their data. They also estimate the size of effects that they are studying. This answers more relevant questions such as 'what is the magnitude of the difference?' rather than 'is there a difference?'

This approach to data analysis is considered more fully at the end of the chapter.

Information-theoretic methods

Information-theoretic methods work in a series of steps:

1. A set of candidate models are selected that represent different hypotheses for explaining reality.
2. Data are collected.
3. The data are used to assess the relative support for the different models, by estimating the amount of information lost when using each. The best model is selected as the one that is estimated to lose the least amount of information.
4. Any required predictions are made using an average of the models that is weighted towards those that are estimated to lose less information.

Select a set of candidate models

One of the main tenets of information theoretic methods is to select a set of possible models (hypotheses) for explaining reality. Information theoretic methods are not constrained to examining only two possible hypotheses as required for null hypothesis testing. An arbitrary number of hypotheses can be examined simultaneously, but Burnham and Anderson (2002) recommend careful selection of the hypotheses. Each hypothesis is represented as a statistical model. The statistical models link the data that are to be collected to various parameter values. The models are selected with the knowledge that most, if not all models in ecology will be imperfect. The aim is to find the most parsimonious model or set of models.

An example will illustrate the construction of possible hypotheses and associated models. Grand *et al.* (1998; see also Anderson *et al.*, 2000) were interested in the effect of lead poisoning on female spectacled eider. Data were available from two sites and the birds were classified as either having been exposed to lead or not based on blood analysis. The five possible hypotheses were:

1. Survival depended on lead exposure but did not vary among sites.
2. Survival depended on both lead exposure and site, with an additive effect.

3. Survival depended on both lead exposure and site, with an interaction between the two (i.e. the effect of lead varied among the sites).
4. Survival did not depend on lead exposure but varied among the sites.
5. Survival did not depend on the site or lead exposure.

Grand *et al.* (1998) constructed statistical models for each of these hypotheses. The models related the observations of each individual over three years (the capture/recapture history) to the survival rates, and the survival rates were functions of the relevant explanatory variables. The results are presented later in this section.

Advocates of information-theoretic methods are some of the firmest critics of the (mis)use of null hypothesis testing in ecology, in particular the use of silly nulls (Johnson, 1999; Anderson *et al.*, 2000). Of course, information-theoretic methods are not immune to silly hypotheses. In the above example, it could be argued that any model that did not include an effect of lead on mortality can be discounted as unlikely a priori. Similarly, one could argue that there must be at least some difference in mortality of spectacled eiders among sites given that the birds will be exposed to different conditions (hunters, predators, food, etc.). Further, one would expect that the effect of lead poisoning on mortality would depend on the site, with no two sites having perfectly identical responses. So, we can claim a priori that model 3 is our best model and that the other models are silly. Any number of other models could be added to the list, such as those in which annual mortality varies as a function of possible weather variables.

Of course, the production of an unlimited number of increasingly complex models is counter to the aim of parsimony (finding the simplest model that still fits the data reasonably well). Advocates of null hypothesis significance testing might argue that by using silly nulls, they are doing something very similar; only including detail when there is evidence that the extra detail is justified. The difference is that information-theoretic methods are based on an aim to minimize the loss of information, whereas decisions about including parameters based on null hypothesis testing depend on the type I error rate, which is usually set at the arbitrary value 0.05.

Collect data

Again, the collection of data is a largely uncontroversial step, with principles of randomization and replication being important, as well as attempts to minimize biases and imprecision. In the spectacled eider example, the researchers marked individuals and constructed a re-sighting history that recorded whether each bird was recorded in subsequent years.

Calculate the relative amount of information lost by each model

The concept of information loss is easy to envisage for digital images. If an image is made up of many pixels, it will tend to be a good reproduction of the original scene. However, as the number of pixels decreases, the image will become less clear as the pixels become larger, and greater amounts of information (detail) will be lost. Although no digital image will provide a perfect representation of the original scene, the various images that are available can be ranked on the basis of their clarity, which measures the relative amount of information lost or gained by using one image compared to another. Similarly, some ecological models will lose more or less detail when trying to represent reality by having different levels of bias and precision.

Akaike (1973) identified the relationship between a formal measure of the information content of a model (Kullback-Leibler information) and values of the maximum likelihood or deviance that are commonly used in statistics (Anderson *et al.*, 2000). This led to Aikake's information criterion (AIC) which is an estimate of the relative Kullback-Leibler information of a model. AIC is calculated from the minimum deviance of the model (a measure of fit) and the number of estimated parameters (a measure of complexity). Poorer fitting models and more complex models lead to greater AIC values. Chapter 4 provides more information about likelihood, deviance, and AIC.

The best model, of those being considered, is the one that is expected to lose the least amount of information (i.e. has the lowest AIC value). Embedded in the calculation of AIC values is the concept of parsimony. Using AIC to select the best model involves a trade-off between model fit and complexity, with more complex models being selected only if they provide a sufficiently superior fit (see Chapter 4).

The differences in AIC among models are more important than the actual values. Differences are usually expressed relative to the model with the smallest AIC value ($\Delta_i = \text{AIC}_i - \text{AIC}_{\text{min}}$). For the spectacled eider example, the model with an effect of only lead was the best model, while the model with additive effects of site and lead was the second best model (Table 2.1)

Table 2.1. *Differences in AIC values (Δ_i) between the best model (model 1) and the other models. Model weights (w_i) were calculated using the relative AIC values (Δ_i). The effect size and standard error is based on the maximum likelihood estimate for the effect of lead for that model.*

Model	Δ_i	w_i	Effect size (s.e.)
1. Lead effect	0.00	0.673	0.337 (0.105)
2. Additive lead and site effects	2.07	0.239	0.335 (0.148)
3. Interactive lead and site effects	4.11	0.086	0.330 (0.216)
4. Site effect (no lead effect)	14.25	0.001	-
5. No site or lead effect	12.71	0.001	-

The AIC differences of each model (Δ_i) can be converted to Akaike weights (w_i) that measure the likelihood of the data given the model. When there are R candidate models, the Akaike weights are

$$w_i = \exp(-\Delta_i/2)/\sum_{r=1}^{R}\exp(-\Delta_r/2).$$

This equation simply means that the Akaike weights are obtained by transforming the AIC differences (Δ_i) using $\exp(-\Delta_i/2)$ and then re-scaling the subsequent values so that they sum to 1. Anderson *et al.* (2000) interpret these weights as approximate probabilities that the model (of those in the candidate set) is the Kullback-Leibler best model, i.e. that of the models considered it minimizes the loss of information. This means that the Akaike weights provide a measure of evidence in favour of each of the candidate models provided we have a priori reasons to believe that the models are equally reasonable. However, Box 2.4 illustrates that a priori evidence matters. Burnham and Anderson (2002, pp. 302–5) discuss more fully the relationship between Akaike weights and model probabilities.

Information theoretic methods also permit evaluation of the evidence that lead influences the mortality of spectacled eider. This is achieved by summing the Akaike weights for the models that include an effect of lead (0.998). Thus, there is very strong evidence that lead affects survival, because the models that do not include lead as an effect have very low weights.

Box 2.4

Information theoretic methods in the courts

An information theoretic approach to the evidence in Sally Clark's case arrives at a similar conclusion to null hypothesis testing (Box 2.3). In this case, we have basically two hypotheses: Sally Clark murdered her children or did not murder her children (ignoring the chance that she murdered only one of them). We can calculate the likelihood of the data (her two children died) under the two hypotheses. For the first, the probability of her two children dying given that she murdered them is clearly 1. The probability of two deaths under the second hypothesis that she is innocent can be calculated from data on cot deaths in the UK. This value is approximately one in 300 000 (not one in 73 million as quoted in court because cot deaths are unlikely to be independent events within families; Hill, 2004). We can calculate AIC values for these two hypotheses. Because the data (two sons dying) are not used to estimate the parameters of the models, $K=0$ and the AIC values are simply equal to the deviance (−2 In(likelihood)):

Hypothesis	Likelihood	AIC	w_i
Children murdered	1.000	0	~1.0
Children not murdered	3.4×10^{-6}	25.2	3.4×10^{-6}

Faced with this analysis, things are not looking up for Sally Clark. The interpretation of Akaike weights (w_i) as model probabilities require us to conclude that it is likely that she murdered her two sons. However, this analysis neglects a vital piece of information, which is that parents only very rarely kill their children (Box 2.5).

Average across models

Information-theoretic methods also provide a basis for including uncertainty about the best model in assessments of effect sizes. For example, model 1 predicts that the presence of lead reduces survival by 0.337 with a standard error of 0.105 (Table 2.1). However, the other models predict slightly different effects. For example, the standard error

Box 2.5
Bayesian methods in the courts

Using the Bayesian method, the prior probability that Sally Clark murdered her two sons can be estimated from rates of infanticide in the United Kingdom (Hill, 2004). Most parents do not murder their children, so the rate of murdering two children is very low (the probability is approximately one in 2.7 million). Thus, we have the prior probability that Sally Clark murdered her two sons (0.00000037), and the probabilities of her two sons dying given the two hypotheses (1 in 300 000 if she did not murder them and 1 if she did), so we can calculate the posterior probability:

$$\begin{aligned}&\text{Pr(Sally Clark murdered her sons given the data)}\\&= 3.7 \times 10^{-7} \times 1/[3.7 \times 10^{-7} \times 1 + (1 - 3.7 \times 10^{-7}) \times 1/300000]\\&= 0.1.\end{aligned}$$

Therefore, it is approximately ten times more likely than not (given the two deaths) that Sally Clark is innocent (Hill, 2004; see also Bondi, 2004 and Joyce, 2002). Of course, other evidence could be brought to bear on this case. Firstly, the rate of infanticide among parents is much lower than the figure used if those parents do not have a history of violence towards their children, as is the case with Sally Clark. At the same time, the rate of cot death is also lower for such families (Hill, 2004). Secondly, medical evidence, some of which only came to light on appeal, increases the likelihood of death by natural causes. The first son was found to have a respiratory infection and the second a bacterial infection, both of which were likely causes of death. Thankfully, given the evidence, Sally Clark appealed her conviction and is now free after spending more than three years in jail. However, ecologists continue to fall into the same trap that appears to have contributed to the conviction of Sally Clark, which is known as the prosecutor's fallacy. This is the mistaken belief that a low probability of obtaining data given a hypothesis means that the alternative hypothesis is likely to be true.

for the effect of lead for model 3, which included the interaction term, is almost twice that of model 1 (Table 2.1).

By using model averaging, it is possible to calculate an effect of lead that accounts for uncertainty in the choice of the best model. Model averaging weights the estimated effect by the Akaike weights. Additionally, the standard error of the model-averaged predictions is a function of the within-model variation (i.e. the standard error of the prediction for each model), the between-model variation (i.e. the differences in the predictions among the different models) and the Akaike weights (see Burnham and Anderson, 2002 for details). In this example, the model-averaged prediction is that lead reduces survival by 0.335 with a standard error of 0.125.

The ability to consider more than one model when making inferences is one of the strengths of information theoretic methods. The chief advantage is recognizing that there is usually some uncertainty about which of the candidate models best describes the data. It is risky to put all one's eggs in one basket (a single model) when other plausible models might make different predictions. More detail on using multi-model inference in ecology can be found in Burnham and Anderson (2002).

Summary of information-theoretic methods

Information theoretic methods have three advantages over null hypothesis testing:

1. They are not influenced by extreme unobserved data.
2. In evaluating a hypothesis, the relative evidence in favour of the different hypotheses is assessed simultaneously while null hypothesis testing can lead to acceptance of the alternative without directly assessing evidence in its favour.
3. They permit simultaneous assessment of multiple hypotheses rather than being confined to pair-wise comparisons. Inference about the magnitude of effects can be based on the relative evidence in favour of these different hypotheses.

It has been argued that information-theoretic methods overcome the problems of null hypothesis testing. However, many of the problems of null hypothesis testing lie in its use rather than the method itself. It is entirely possible that similar errors of use may arise when using information theoretic methods (or other approaches to statistics such as

Bayesian methods). For example, the following possible errors that might arise when using this method are largely analogous to the errors that occur with the misuse of null hypothesis testing:

1. Using silly hypotheses;
2. Believing that the Akaike weight is the probability that the hypothesis is true;
3. Choosing the best model (that with the smallest AIC value) and ignoring other possible models with similar AIC values;
4. Not assessing the ability of study designs to distinguish between different models a priori;
5. Using arbitrary thresholds for differences between AIC values to decide whether a model is considered further or not;
6. Ignoring the size of effects being estimated when deciding which model is most parsimonious.

It remains to be seen whether these errors or others become common in ecology. Proponents of information-theoretic methods would argue that such errors are the fault of the user, not of the method and that the method has ways of dealing with them. A similar defence can be mounted for most of the criticisms of null hypothesis testing. Therefore, I believe that whether information theoretic methods are better than null hypothesis testing will depend on how they are used; whether people make fewer errors of interpretation and implementation when using one or the other. This is largely a question of cognition, depending on how well the different methods are taught and understood, the quality of the available software, etc.

In the next section, I describe Bayesian methods that have some clear advantages over null hypothesis testing and information theoretic methods. Bayesian methods introduce some extra difficulties, but most of these are easy to overcome.

Bayesian methods

Perhaps the main defining feature of Bayesian methods is calculation of the probability of a hypothesis being true. These hypotheses can be discrete (e.g. the frog surveying problem in Chapter 1) or continuous (e.g. when estimating a mean, Box 1.8). While both null hypothesis testing and information theoretic methods might seem to measure the reliability of different hypotheses given the data (with *p*-values or

Akaike weights), they actually represent the probability of obtaining the data given the hypotheses.

The steps to conducting a Bayesian analysis are:

1. A set of candidate models are selected that represent different hypotheses for explaining reality.
2. Prior probabilities are assigned to these different models.
3. Data are collected.
4. Bayes' rule is used to combine the prior probabilities with the information contained in the new data to generate the posterior predictions.

Select a set of candidate models

This is essentially the same step as used in information theoretic methods, with it being possible to use any number of competing models. The same criticisms apply. While critics point out that null hypothesis testing can lead to the use of silly nulls, there is nothing to stop silly hypotheses being used with information theoretic or Bayesian methods. Perhaps one advantage of Bayesian methods is that users are forced to establish prior probabilities for the competing models. Therefore, silly hypotheses may be noted and assigned small prior probabilities. However, how does one assign these probabilities?

Assign prior probabilities

The frog surveying problem (Chapter 1) provides an example of assigning priors to the different hypotheses. The two hypotheses are that the southern brown tree frog is present or absent from a surveyed site. If we had no previous information, then we might conclude there is nothing to choose between the two hypotheses before collecting data and assign equal prior probabilities to each. Such use of uniform priors is common in the face of ignorance.

However, if we know from previous surveys that the species is found in a particular fraction of ponds in the region, then we could use that fraction as the probability that the frog is present. The probability of the frog being absent is simply one minus the probability that it is present.

Finally, we might have a model for predicting the probability that the frog is present at ponds based on their characteristics, in which case this could be used as the prior. Each of these three cases reflects different levels of prior information. The first represents ignorance, the second the mean rate of occurrence of the frog within ponds, and the third how the rate of occurrence varies among ponds of different types. The addition of prior information in this way influences the results. If the pond has a high prior probability of the frog being present, then a single survey in which it is not seen would not be enough for us to be reasonably sure it is absent unless our ability to detect the frog was very good.

Using a uniform distribution to represent ignorance makes sense in some ways, but is problematic in others. Consider the case where we wish to determine the proportion of individuals that belong to each species in an African national park. Among the herbivores, we might be interested in the proportion of individuals that are zebras, wildebeest or some other species. By using the uniform distribution to represent ignorance, we would assign a one-third probability to each of these three classes (zebras, wildebeest, other). However, the probability of one-third is simply an artefact of our classification. The zebra would have had a probability of one-quarter if we had included gazelles as an additional class. Therefore, representing ignorance is not always straightforward.

Problems of representing ignorance also arise when specifying priors for continuous hypotheses (see also Box 3.12). For example, we may wish to estimate the density of territories of a species that are adjacent but non-overlapping. We could assign the prior distribution as uniform between 0.1 and 1.0 territories per ha if we were confident that the density was somewhere within that range but unsure of the actual value. This prior implies that the probability of the density being less than 0.2 territories per ha is 0.111 (0.1/0.9).

Alternatively, we could specify that the area of each territory is between 1 and 10 ha. This is equivalent to our limits for density (1.0 and 0.1 territories per ha, respectively). If we used a uniform distribution between 1 and 10 ha for territory size, the probability of the territory size being more than 5 ha is 0.555 (5/9), which is five times the probability calculated above for the equivalent density (0.2 territories per ha). Thus, we seem to have proved that $0.111 = 0.555$.

The difference arises because the units of the two approaches are not linearly related, so the probabilities are not conserved when they are transformed (in this case by inversion). The prior distributions are very different (Fig. 2.2). This effect of the scale of measurement is not unique to Bayesian analyses. For example, a frequentist confidence interval based on territory size would not be equivalent to a confidence interval based on densities of territories per ha.

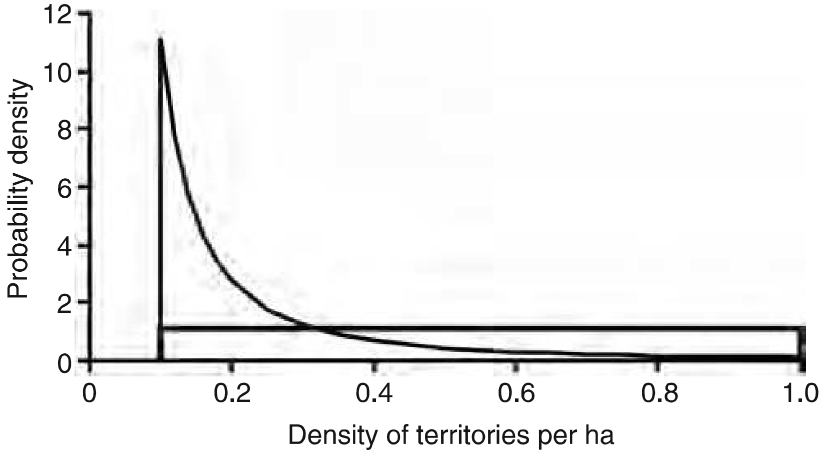

Fig. 2.2 Prior distributions for the density of territories assuming that the density is uniformly distributed between 0.1 and 1.0 territories per ha (the uniform distribution) and assuming that the size of territories (the inverse of density) is uniformly distributed between 1 and 10 ha (the sharply peaked distribution). The probability of territory density being less than 0.2 per ha (the areas under the curves to the left of 0.2) is very different for the two priors.

One of the difficulties in establishing prior probabilities is that humans tend to judge them poorly (Tversky and Kahneman, 1974; Kahneman *et al.*, 1982; Ayton and Wright, 1994; Gigerenzer and Hoffrage, 1995; Anderson, 1998; Burgman, 2005). Construction of priors by using subjective judgement is likely to depend on a range of personal attributes, how the problem is presented, motivational biases and advocacy (Anderson, 1998; Burgman, 2005). Experts are not immune to these frailties of human nature (Burgman, 2005; see also Chapter 10).

Even when there are data for constructing priors, some subjective judgement is required to determine how the prior information is represented as a probability distribution. Frequentist methods are not free of subjective judgement because they also depend on judgements about the questions to be examined, how the data are collected, the variables to be analysed, and the statistical methods and models that are used (Howson and Urbach, 1991).

It could be argued that although science is not free of subjectivity (Burgman, 2005), it should seek to minimize it (Dennis, 1996). How can Bayesian methods be used reliably and convincingly in the face of subjectivity? One approach is to be as careful, rigorous and convincing in the choice of prior as in the collection of data. This book contains various examples of using previous results and data to construct priors. Furthermore, there will be many cases where the choice of prior has virtually no effect on the results.

However, there will still be cases where uncertainty in the choice of prior remains. This uncertainty can be regarded as an honest incorporation of subjectivity in science (Berger and Berry, 1988, Howson and Urbach, 1991). The role of science is to ensure that this opinion is updated logically as evidence accumulates. Bayes' rule ensures that beliefs are updated logically, with differing opinions converging as data are collected (Cox, 1946; Howson and Urbach, 1991; Crome *et al.*, 1996; Jaynes, 2003).

There are extensions to Bayesian methods for dealing with uncertainty in the choice of priors. The methods, lumped under the title of 'robust Bayesian analysis', can also deal with uncertainty in the models used to represent the hypotheses. They involve, for example, placing bounds on the possible parameters or distributions for priors and likelihoods and, therefore, bounds on the possible posterior distributions (e.g. Berger, 1985; Walley, 1991; Ferson, 2005). Robust Bayesian methods are not without controversy, and they usually add to the computational burden. In providing an introduction to Bayesian methods for ecology, I will only touch on them briefly in Chapter 10. Interested readers are referred to Berger (1985) and Ferson (2005).

Although the prior can pose difficulties for Bayesian methods, it is in fact one of its strengths. Ecologists, in the discussion sections of journal articles, routinely consider their results in the light of previous studies. Bayesian methods provide a formal basis for these comparisons through the use of priors. Scientists may be forced to be more rigorous and less subjective when using priors to represent previous work than when simply using their judgement to make comparisons. Bayes' rule provides the means of incorporating previous findings into the formal interpretation of new data.

Use Bayes' rule to combine the prior and the data

Bayes' rule states that the probability of a hypothesis given data (the posterior) is proportional to the product of the prior and the probability of the data given the hypothesis. The constant of proportionality is given by a sum (for discrete hypotheses) or an integral (for continuous hypotheses).

There is little that is controversial about Bayes' rule itself. Given a prior probability, some data and a set of hypotheses, it provides the updated belief in the hypotheses. Bayes' rule provides a logical means (some claim the only logical means, Jaynes, 2003) of updating belief (Box 2.5).

Given that having a posterior belief in hypotheses requires a prior belief, Bayesian methods are required if we wish to use data to assign degrees of belief to hypotheses (Cox, 1946; Jeffreys, 1961; Jaynes, 2003). Bayesian methods are required even when it is difficult to construct the priors. As mentioned previously, the main controversy with Bayesian methods involves how these priors are constructed. While critics of Bayesian methods point to difficulties of establishing priors, proponents are uncomfortable about ignoring relevant prior information if it exists.

In Bayesian statistics, probability distributions for the prior and posterior distribution represent uncertainty about a parameter value. Because of this use of probability distributions, some authors refer to Bayesian parameters as random variables (Dennis, 1996; Ellison, 2004). However, this does not necessarily mean that the true value for a parameter is assumed to vary randomly from one measurement to another (Clark, 2005). The parameter might have a fixed but unknown value, which can only be expressed probabilistically. The probability distribution represents the uncertainty about the parameter, describing which values are more or less probable. As more (unbiased) data are collected, the posterior distribution becomes more concentrated on the true value for the parameter.

Arbitrarily complex statistical models can be analysed using Bayesian methods. For example, the numerical procedures for most analyses of variance require that the variance of the data for each level of a factor is identical; Bayesian analyses can easily handle cases where the variances are different. Similarly, variances can be assumed to change across the range of the data for regression analyses, rather than assuming the variance is constant. Another example is that it is relatively straightforward to introduce hierarchical effects (Clark, 2005), making it much easier to deal with problems such as pseudo-replication. Therefore, rather than making the study design conform to the required analysis, Bayesian data analyses can be made to conform to the study design. This opens the possibility of combining data from multiple sources and using a wider range of statistical models.

Nevertheless, some computational limitations of Bayesian methods remain. Computationally intensive methods (e.g. multivariate factor analyses) can take a long time to analyse with Bayesian methods, and may not be feasible if the required mathematical functions are not contained in the available software. Although these issues limit some current applications, they will most likely be surmounted with further software development.

Estimating effect sizes

One of the main types of questions asked by ecologists is how big is the effect or what is the nature of the relationship between variables? For example, we might ask how has the population size changed over time or what is the strength of the relationship between these ecological variables? Bayesian and information theoretic methods, and with slight modification the basic statistical machinery that is used to calculate *p*-values, can be used to answer these questions. So, rather than asking: 'Does plant diversity of marshes vary with salinity?' we could ask: 'How does plant diversity of marshes vary with salinity?' The former question was answered by Mullan Crain *et al.* (2004) using null hypothesis testing ($p < 0.0001$).

Mullan Crain *et al.* (2004) also reported the answer to the second question—species diversity increased from salt to brackish to fresh marshes. The estimated Shannon diversity indices were 0.12 ± 0.013 (mean ± s.e.) for salt, 0.22 ± 0.015 for brackish and 0.39 ± 0.016 for fresh. By providing standard errors, readers can construct confidence intervals for the estimated plant species diversity and consider the magnitude of the differences. Such considerations show that there is a trend in plant species diversity, with the diversity index in freshwater being approximately three times that of saltwater. This result is clearly more informative than 'salinity affects plant species diversity', which we expected to be true prior to the data collection and analysis. However, we might not have known the nature of those differences.

There is virtually no disagreement that the estimated size of effects and a measure of the precision of the estimate should be provided for any statistical analysis (Fidler *et al.*, 2004). The Ecological Society of America encourages this practice in their guide to authors who wish to publish in their journals. Despite this encouragement, it is not routinely practised by ecologists or by researchers in some other disciplines (Fidler *et al.*, 2004).

Frequentist confidence intervals and Bayesian credible intervals can be used to represent the precision of an estimate.[4] Such intervals can be used to determine whether an estimated effect is likely to be ecologically important. For example, McCarthy and Parris (2004) presented Bayesian 95% credible intervals for the effect of clipping toes from frogs on return

[4] This representation of precision is based on the length of the interval, rather than the earlier definition of precision as the inverse of the variance.

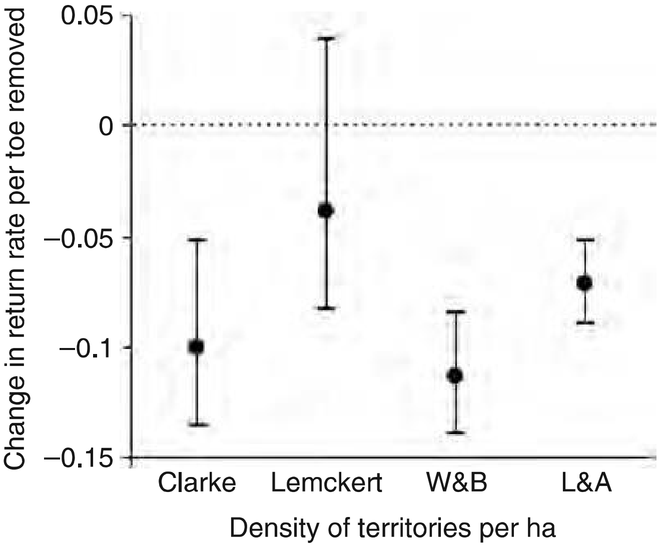

Fig. 2.3 Estimated effect of toe clipping on return rates of frogs for four different studies. The bars represent 95% Bayesian credible intervals and the circles are the means of the predicted effect. Negative values indicate an adverse effect (from McCarthy and Parris, 2004). See Chapter 8 for more details.

rates of marked animals (Fig. 2.3). The results demonstrate that toe clipping almost certainly reduces return rates in three of the four studies examined because the intervals are less than zero. Because the confidence and credible intervals in this case are numerically similar, this is equivalent to obtaining a statistically significant result, rejecting the null hypothesis that there is no effect of toe clipping (represented by the dashed line at zero).

In one study, the credible interval is rather wide and does encompass zero. This is equivalent to not rejecting the null hypothesis. However, the results can also be compared to values that might be deemed ecologically important. A value of −0.03 might be regarded as ecologically important, because given that it is not unusual to clip three toes from frogs in mark-recapture studies, the actual reduction in return rate would be ~10% per frog. If 10% of frogs are not recaptured because of the marking method, this might lead to unacceptable bias in the results as well as impacts on the population if the toe clipping is causing mortality.

By inspecting the credible intervals (Fig. 2.3), we can determine whether the results are consistent with an ecologically important effect such as −0.03. In three of the four studies, we can be confident that effects are at least this large. In the other study, there is a reasonably large chance that the reduction is greater than 0.03 per toe, but it is possible that the effect is not this large.

This illustrates one of the advantages of using intervals; the results can be compared easily to values that are ecologically meaningful, rather than only focusing on statistical significance. Although ecological importance can be examined with null hypothesis testing (e.g. by using a null hypothesis that is ecologically important), the prevalence of silly nulls means that this is rarely done.

An additional advantage of using confidence or credible intervals is that the concept of power is communicated by the width of the interval. A wide interval (relative to the size of the difference we wish to detect) means that the study has low precision—it is equivalent to having low power in null hypothesis testing. In the toe clipping example, a study would need to have a narrow confidence interval centred on a value close to (or greater than) zero to show that the adverse effect of toe clipping was not large.

A focus on ecological importance can be difficult, because we often do not know what constitutes an important difference. In these cases, one might argue that null hypothesis testing should then rely on determining the presence or absence of any difference, so a null hypothesis of no difference is appropriate. However, if we cannot establish whether an observed difference is important or not, should we be testing for any particular difference in the first place? Analyses that quantify the magnitude of effects, precision of their estimates, and the relationships among different variables would be more appropriate in these circumstances. These studies could then help us to identify ecologically important results.

In most cases, the choice between Bayesian credible intervals and frequentist confidence intervals is not important because the results are numerical similar. However, for some statistical models or when prior information is available, the two approaches will generate different answers (Jaynes, 1976; see also Chapter 1). In such circumstances, the frequentist confidence intervals perform poorly compared with the Bayesian credible intervals (Jaynes, 1976).

There are numerous calls for estimating effect sizes in ecology and other disciplines (Anderson *et al.*, 2000; Fidler *et al.*, 2004; Ziliak and McCloskey, 2004). As discussed above, there are several advantages of using intervals. Additionally, the results are much more useful for meta-analyses (Arnquist and Wooster, 1995; Gurevitch and Hedges, 2001), and the ecological importance of the results is more readily apparent. From merely selfish perspectives, this should encourage authors and editors to use intervals in their manuscripts because their work will be cited more frequently if the results are clearer and more appropriate for meta-analysis. These points relate to the progress of science, in which evidence accumulates over time. Both meta-analyses that rely on estimating effect sizes and Bayesian analyses are cumulative over studies, whereas null hypothesis testing is not.

Concluding remarks

Frequentist and Bayesian methods of statistical analysis differ in how they treat the notion of probability. Bayesian methods use probabilities to assign degrees of belief to hypotheses or parameter values. In contrast, frequentist methods (null hypothesis testing and information theoretic methods) are confined to stating the frequency with which data would be collected given hypothetical replicate sampling and specified hypotheses being true. Given the disagreement about which approach to statistics is preferred, the relative merits of the different methods are clearly a matter of opinion. My preference is for Bayesian methods because I believe ecologists are usually attempting to assign degrees of belief in parameter values, models or hypotheses more generally (Table 2.2). Ecologists regard the truth as uncertain and attempt to use science to gain an improved understanding of the truth. Such an approach is consistent with Bayesian statistics.

Many of the criticisms of the different statistical methods are directed at the use of the methods, rather than their underlying basis. Null hypothesis significance testing is criticized because of its

Table 2.2. *Benefits and limitations of Bayesian statistics (adapted from O'Hagan and Luce, 2003).*

Benefits	Limitations
Allows for intuitive interpretation	Introduces an element of subjectivity (although treating it explicitly rather than ignoring it may be a benefit)
Uses prior information	There are difficulties in constructing priors
Addresses a greater range of problems	Bayesian methods are not commonly taught
Allows complex models to be analysed easily	
Accommodates decision making	
Use all the information transparently	

widespread misuse. However, it also has logical shortfalls because unobserved data influence the results, and acceptance of the alternative hypothesis does not depend on how well the evidence supports it. Frequentist methods in general are forced to ignore any relevant prior information. Additionally, they are not well-suited to decisions about individual cases, being restricted to assessing long-run frequencies obtained from hypothetical samples. Bayesian methods are criticized because it can be difficult to determine how prior information should be incorporated into analyses.

Despite the differences, Bayesian and frequentist methods often generate numerically similar answers, especially when estimating parameters and prior information is uninformative. In these circumstances, the best approach will be largely determined by which is most easily and successfully taught, learnt, and executed. Therefore, the success of the different methods lies firmly in the realm of cognitive psychology not just statistics. However, Bayesian methods have the distinct advantage that when the numerical results differ, the Bayesian methods are invariably correct (Jaynes, 1976).