CHAPTER

1

Introduction

OUTLINE

WinBUGS (Gilks et al., 1994; Spiegelhalter et al., 2003; Lunn et al., 2009) is a general-purpose software program to fit statistical models under the Bayesian approach to statistics. That is, statistical inference is based on the posterior distribution, which expresses all that is known about the parameters of a statistical model, given the data and existing knowledge. In recent years, the Bayesian paradigm has gained tremendous momentum

Introduction to WinBUGS for Ecologists
DOI: 10.1016/B978-0-12-378605-0.00001-6

in statistics and its applications, including ecology, so it is natural to wonder about the reasons for this.

1.1 ADVANTAGES OF THE BAYESIAN APPROACH TO STATISTICS

Key assets of the Bayesian approach and of the associated computational methods include the following:

1.1.1 Numerical Tractability

Many statistical models are currently too complex to be fitted using classical statistical methods, but they can be fitted using Bayesian computational methods (Link et al., 2002). However, it may be reassuring that, in many cases, Bayesian inference gives answers that numerically closely match those obtained by classical methods.

1.1.2 Absence of Asymptotics

Asymptotically, that is, for a "large" sample, classical inference based on maximum likelihood (ML) is unbiased, i.e., in the long run right on target. However, for finite sample sizes, *i.e., for your data set*, ML may well be biased (Le Cam, 1990). Similarly, standard errors and confidence intervals are valid only for "large" samples. Statisticians never say what "large" exactly means, but you can be assured that typical ecological data sets aren't large. In contrast, Bayesian inference is *exact* for any sample size. This issue is not widely understood by ecological practitioners of statistics but may be particularly interesting for ecologists since our data sets are typically small to very small.

1.1.3 Ease of Error Propagation

In classical statistics, computing the uncertainty of functions of random variables such as parameters is not straightforward and involves approximations such as the delta method (Williams et al., 2002). For instance, consider obtaining an estimate for a population growth rate ($\hat{r}$) that is composed of two estimates of population size in subsequent years ($\hat{N}_1$, $\hat{N}_2$). We have $\hat{N}_1$ and $\hat{N}_2$ and we want $\hat{r}$: what should we do? Getting the point estimate of $\hat{r}$ is easy, but what about its standard error? In a Bayesian analysis with Markov chain Monte Carlo, estimating such, and much more complex, derived quantities including their uncertainty is trivial once we have a random sample from the posterior distribution of their constituent parts, such as $\hat{N}_1$ and $\hat{N}_2$ in our example.

1.1.4 Formal Framework for Combining Information

By basing inference on both what we knew before (the prior) and what we see now (the data at hand), and using solely the laws of probability for that combination, Bayesian statistics provides a formal mechanism for introducing external knowledge into an analysis. This may greatly increase the precision of the estimates (McCarthy and Masters, 2005); some parameters may only become estimable through precisely this combination of information.

Using existing information also appears a very sensible thing to do: after all, only rarely don't we know anything at all about the likely magnitude of an estimated parameter. For instance, when estimating the annual survival rate in a population of some large bird species such as a condor, we would be rather surprised to find it to be less than, say, 0.9. Values of less than, say, 0.5 would appear downright impossible. However, in classical statistics, by not using any existing information, we effectively say that the survival rate in that population could be just as well 0.1 as 0.9, or even 0 or 1. This is not really a sensible attitude since every population ecologists knows very well *a priori* that no condor population would ever survive for very long with a survival rate of 0.1. In classical statistics, we always feign total ignorance about the system under study when we analyze it.

However, within some limits, it is also possible to specify ignorance in a Bayesian analysis. That is, also under the Bayesian paradigm, we can base our inference on the observed data alone and thereby obtain inferences that are typically very similar numerically to those obtained in a classical analysis.

1.1.5 Intuitive Appeal

The *interpretation of probability* in the Bayesian paradigm is much more intuitive than in the classical statistical framework; in particular, we directly calculate the probability that a parameter has a certain value rather than the probability of obtaining a certain kind of data set, given some Null hypothesis. Hence, popular statements such as "I am 99% sure that …" are only possible in a Bayesian mode of inference, but they are impossible in principle under the classical mode of inference. This is because, in the Bayesian approach, a probability statement is made about a parameter, whereas in the classical approach, it is about a data set.

Furthermore, by drawing conclusions based on a combination of what we knew before (the prior, or the "experience" part of learning) and what we see now (the likelihood, or the "current observation" part of learning), Bayesian statistics represent a *mathematical formalization of the learning process*, i.e., of how we all deal with and process information in science as well as in our daily life.

1.1.6 Coherence and Intellectual Beauty

The entire Bayesian theory of statistics is based on just three axioms of probability (Lindley, 1983, 2006). This contrasts with classical statistics that Bayesians are so fond to criticize for being a patchwork of theory and *ad hoc* amendments containing plenty of internal contradictions.

1.2 SO WHY THEN ISN'T EVERYONE A BAYESIAN?

Given all the advantages of the Bayesian approach to statistics just mentioned, it may come as a surprise that currently almost all ecologists still use classical statistics. Why is this?

Of course, there is some resistance to the Bayesian philosophy with its perceived subjectivity of prior choice and the challenge of avoiding to, unknowingly, inject information into an analysis via the priors, see Chapter 2. However, arguably, the lack of a much more widespread adoption of Bayesian methods in ecology has mostly practical reasons.

First, a Bayesian treatment shines most for complex models, which may not even be fit in a frequentist mode of inference (Link et al., 2002). Hence, until very recently, most applications of Bayesian statistics featured rather complex statistical models. These are neither the easiest to understand in the first place, nor may they be relevant to the majority of ecologists. Second, typical introductory books on Bayesian statistics are written in what is fairly heavy mathematics to most ecologists. Hence, getting to the entry point of the Bayesian world of statistics has been very difficult for many ecologists. Third, Bayesian philosophy and computational methods are not usually taught at universities. Finally, and perhaps most importantly, the practical implementation of a Bayesian analysis has typically involved custom-written code in general-purpose computer languages such as Fortran or C++. Therefore, for someone lacking a solid knowledge in statistics and computing, Bayesian analyses were essentially out of reach.

1.3 WinBUGS

This last point has radically changed with the advent of WinBUGS (Lunn et al., 2009). Arguably, WinBUGS is the only software that allows an average numerate ecologist to conduct his own Bayesian analyses of realistically complex, customized statistical models. By customized I mean that one is not constrained to run only those models that a program lets you select by clicking on a button. However, although WinBUGS has been and is increasingly being used in ecology, the paucity of really accessible and attractive introductions to WinBUGS for ecologists is a surprise (but see McCarthy, 2007). I believe that this is the main reason for why WinBUGS isn't even more widely used in ecology.

1.4 WHY THIS BOOK?

This book aims at filling this gap by gently introducing ecologists to WinBUGS for exactly those methods they use most often, i.e., the linear, generalized linear, linear mixed, and generalized linear mixed model (GLMM). Table 1.1 shows how the three latter model classes are all generalizations of the simple Normal linear model (LM) in the top left cell of the body of the table. They extend the Normal model to contain either more than a single random process (represented by the residual in the Normal LM) and/or to exponential family distributions other than the Normal, e.g., Poisson and Binomial. Alternatively, starting from the GLMM in the bottom right cell, the other three model classes can be viewed as special cases obtained by imposing restrictions on a general GLMM.

These four model classes form the core of modern applied statistics. However, even though many ecologists will have applied them often using click-and-point programs or even statistics packages with a programming language such as GenStat, R, or SAS, I dare express doubts whether they all really always understand the models they have fitted. Having to specify a model in the elementary way that one has to in WinBUGS will prove to greatly enhance your understanding of these models, whether you fit them by some sort of likelihood analysis (e.g., ML or restricted maximum likelihood [REML]) or in a Bayesian analysis.

Apart from the gentle and nonmathematical presentation by examples, the unique selling points of this book, which distinguish it from others, are

TABLE 1.1 Classification of Some Core Models Used for Applied Statistical Analysis

	Single Random Process	**Two or More Random Processes**
Normal response	Linear model (LM)	Linear mixed model (LMM)
Exponential family response	Generalized linear model (GLM)	Generalized linear mixed model (GLMM)

the full integration of all WinBUGS analyses into program R, the parallel presentation of classical and Bayesian analyses of all models and the use of simulated data sets. Next, I briefly expand on each of these points.

1.4.1 This Is Also an R Book

One key feature of this book as an introduction to WinBUGS is that we conduct all analyses in WinBUGS fully integrated within program R (R Development Core Team, 2007). R has become the *lingua franca* of modern statistical computing and conducting your Bayesian analysis in WinBUGS from within an R session has great practical benefits. Moreover, we also see how to conduct all analyses using common R functions such as `lm()`, `glm()`, and `glmer()`. This has the added bonus that this book will be useful to you even if you only want to learn to understand and fit the models in Table 1 in a classical statistical setting.

1.4.2 Juxtaposition of Classical and Bayesian Analyses

Another key feature is the juxtaposition of analyses using the classical methods provided for in program R (mostly ML) and the analyses of the same models in a Bayesian mode of inference using WinBUGS. Thus, with the exception of Chapters 20 and 21, we fit every model in both the classical and the Bayesian mode of inference. I have two reasons for creating parallel examples. First, this should increase your confidence into the "new" (Bayesian) solutions since with vague priors they give numerically very similar answers as the "old" solutions (e.g., ML). Second, the analysis of a single model by both classical and Bayesian methods should help to demystify Bayesian analysis. One sometimes reads statements like "we used a Bayesian model," or "perhaps a Bayesian model should be tried on this difficult problem." This is nonsense! Since any model exists independently of the method we choose to analyze it. For instance, the linear regression model is not Bayesian or non-Bayesian; rather, this model may be *analyzed* in a Bayesian or in a frequentist mode of inference. Even that class of models which has come to be seen as almost synonymous with Bayesian inference, hierarchical models which specify a hierarchy of stochastic processes, is not intrinsically Bayesian; rather, hierarchical models can be analyzed by frequentist (de Valpine and Hastings, 2002; Lee et al., 2006; de Valpine, 2009; Ponciano et al., 2009) or by Bayesian methods (Link and Sauer, 2002; Sauer and Link, 2002; Wikle, 2003; Clark et al., 2005). Indeed, many statisticians now use the two modes of inference quite opportunistically (Royle and Dorazio, 2006, 2008). Thus, the juxtaposition of classical and Bayesian analysis of the same models should make it very clear that a model is one thing and its analysis another and that there really is no such thing as a "Bayesian model."

1.4.3 The Power of Simulating Data

A third key feature of this book is the use of simulated data sets throughout (except for one data set used repeatedly in the exercises). At first, this may seem artificial, and I have no doubts that some readers may be disinterested in an analysis when a problem is perceived as "unreal." However, I would claim that several very important benefits accrue from the use of simulated data sets, especially in an introductory book:

1. For simulated data, truth is known. That is, estimates obtained in the analysis of a model can be compared with what we know they should be in the long-run average.
2. When coding an analysis in WinBUGS, especially in more complex cases but even for simpler ones, it is very easy to make mistakes. Ensuring that an analysis recovers estimates that resemble the known input values used to generate a data set can be an important check that it has been coded correctly.
3. It has been said that one of the most difficult, but absolutely necessary statistical concepts to grasp is that of the sampling variation of an estimator. For nonstatisticians, I don't see any other way to grasp the meaning of sampling variation other than literally experiencing it by repeatedly simulating data under the same model, analyzing them, and seeing how estimates differ randomly from one sample to the next: this variation is exactly what the standard error of an estimate quantifies. In real life, one typically only ever observes a single realization (i.e., data set) from the stochastic system about which one wants to make an inference in a statistical analysis. Hence, for ecologists it may be hard to make the connection with the concept of repeated samples from a system, when all we have is a single data set (and related to that, to understand the difference between a standard deviation and a standard error).
4. Simulating data can be used to study the long-run average characteristics of estimates, given a certain kind of data set, by repeating the same data generation-data analysis cycle many times. In this way, the (frequentist) operating characteristics of an estimator (bias, or "is it on target on average?"; efficiency, or "how far away from the target is the individual estimate on average?") can be studied by packaging both the simulation and the analysis into a loop and comparing the distribution of the resulting estimates to the known truth. Further, required sample sizes to obtain a desired level of precision can be investigated, as can issues of parameter estimability. All this can be done for exactly the specifications of one's data set, e.g., replicate data sets can be generated and analyzed with sample size and parameter values identical to those in one's real data set to get an impression, say, of the precision of the estimates that one is likely to

obtain. This is also the idea behind posterior predictive checks of goodness-of-fit, where the "natural" lack of fit for a model is studied using ideal data sets and then compared with the lack of fit observed for the actual data set (see Section 8.4.2).

5. Simulated data sets can be used to study effects of assumption violations. All models embody a set of assumptions that will be violated to some degree. Whether this has serious consequences for those estimates one is particularly interested in, can be studied using simulation.
6. Finally, and perhaps most importantly, I would claim that the ultimate proof that one has really understood the analysis of a statistical model is when one is able to simulate a data set under that very model. Analyzing data is a little like fixing a motorbike but in reverse: it consists of breaking a data set into its parts (e.g., covariate effects and variances), whereas fixing a bike means putting all the parts of a bike into the right place. One way to convince yourself that you really understand how a bike works is to first dismantle and then reassemble it again to a functioning vehicle. Similarly, for data analysis, by first assembling a data set and then breaking it apart into recognizable parts by analyzing it, you can prove to yourself that you really understand the analysis.

In summary, I believe that the value of simulation for analysis and understanding of complex stochastic systems can hardly be overstated. On a personal note, what has helped me most to understand nonnormal GLMs or mixed models, apart from having to specify them in the intuitive BUGS language, was to simulate the associated data sets in program R, which is great for simulating data.

Finally, I hope that the slightly artificial flavor of my data sets is more than made up for by their nice ecological setting and the attractive organisms we pretend to be studying. I imagine that many ecologists will by far prefer learning about new statistical methods using artificial *ecological* data sets than using real, but "boring" data sets from the political, social, economical, or medical sciences, as one has to do in many excellent introductory books.

1.5 WHAT THIS BOOK IS NOT ABOUT: THEORY OF BAYESIAN STATISTICS AND COMPUTATION

The theory of Bayesian inference is treated only very cursorily in this book (see Chapter 2). Other authors have done this admirably, and I refer you to them. Texts that should be accessible to ecologists include

Ellison (1996), Wade (2000), Link et al. (2002), Bernardo (2003), Brooks (2003), Gelman et al. (2004), Woodworth (2004), McCarthy (2007), Royle and Dorazio (2008), King et al. (2009), and Link and Barker (2010).

Furthermore, I don't dwell on explaining Markov chain Monte Carlo (MCMC) or Gibbs sampling, the computational methods most frequently used to fit models in the Bayesian framework. Arguably, a deep understanding of the details of MCMC is not required for an ecologist to conduct an adequate Bayesian analysis using WinBUGS. After all, very few ecologists who nowadays fit a GLM or a mixed model understand the (possibly restricted) likelihood function or the algorithms used to find its maximum. (Or can you explain the Newton–Raphson algorithm? And how about iteratively reweighted least squares?) Rather, by using WinBUGS we are going to experience some of the key features of MCMC. This includes the chain's initial transient behavior, the resultant need for visual or numerical assessment of convergence that leads to discarding of initial ("burn-in") parts of a chain, and the fact that successive iterations are not independent. If you want to read more on Bayesian computation, most of the above references may serve as an entry point to a rich literature.

1.6 FURTHER READING

If you seriously consider going Bayesian for your statistical modeling, you will probably want to purchase more than a single book. McCarthy (2007) is an accessible introduction to WinBUGS for beginners, although it presents WinBUGS only as a standalone application (i.e., not run from R) and the coverage of model classes dealt with is somewhat more limited. Gelman and Hill (2007) is an excellent textbook on linear, generalized, and mixed (generalized) linear models fit in both the classical and the Bayesian mode of inference and using both R and WinBUGS. Thus, its concept is somewhat similar to that of this book, though it does not feature the rigorous juxtaposition of both kinds of analysis. All examples are from the social and political sciences, which will perhaps not particularly interest an ecologist. However, the book contains a wealth of information that should be digestible for the audience of this book, as does Gelman et al. (2004). Ntzoufras (2009) is a new and comprehensive introduction to WinBUGS focusing on GLMs. It is very useful, but has a higher mathematical level and uses WinBUGS as a standalone application only. Woodworth (2004) is an entry-level introduction to Bayesian inference and also has some WinBUGS code examples.

Link and Barker (2010) is an excellent textbook on Bayesian inference specifically for ecologists and featuring numerous WinBUGS examples.

As an introduction to Bayesianism written mostly in everyday language, Lindley, an influential Bayesian thinker, has written a delightful book, where he argues, among others, that *probability is the extension of logic to all events, both certain (like classical logic) and uncertain* (Lindley, 2006, p. 66). His book is not about practical aspects of Bayesian analysis, but very informative, quite amusing and above all, written in an accessible way.

In this book, we run WinBUGS from within program R; hence, some knowledge of R is required. Your level of knowledge of R only needs to be minimal and any simple introduction to R would probably suffice to enable you to use this book. I like Dalgaard (2001) as a very accessible introduction that focuses mostly on linear models, and at a slightly higher level, featuring mostly GLMs, Crawley (2005) and Aitkin et al. (2009). More comprehensive R books will also contain everything required, e.g., Venables and Ripley (2002), Clark (2007), and Bolker (2008).

This book barely touches some of the statistical models that one would perhaps particularly expect to see in a statistics book for ecologists, namely, Chapters 20 and 21. I say nothing on such core topics in ecological statistics such as the estimation of population density, survival and other vital rates, or community parameters (Buckland et al., 2001; Borchers et al., 2002; Williams et al., 2002). This is intentional. I hope that my book lays the groundwork for a much better understanding of statistical modeling using WinBUGS. This will allow you to better tackle more complex and specialized analyses, including those featured in books like Royle and Dorazio (2008), King et al. (2009), and Link and Barker (2010).

Free documentation for WinBUGS abounds, see *http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml*. The manual comes along with the program; within WinBUGS go `Help > User Manual` or press F1 and then scroll down. Recently, an open-source version of BUGS has been developed under the name of OpenBugs, see *http://mathstat.helsinki.fi/openbugs/*, and the latest release contains a set of ecological example analyses including those featured in Chapters 20 and 21. WinBUGS can be run in combination with other programs such as R, GenStat, Matlab, SAS; see the main WinBUGS Web site. There is even an Excel front-end (see *http://www.axrf86.dsl.pipex.com/*) that allows you to fit a wide range of complex models without even knowing the BUGS language. However, most serious WinBUGS users I know run it from R (see Chapter 5). It turns out that one of the main challenges for the budding WinBUGS programmer is to really understand the linear model (see Chapter 6). One particularly good introduction to the linear model in the context of survival and population estimation is Chapter 6 in Evan Cooch's *Gentle introduction to MARK* (see *http://www.phidot.org/software/mark/docs/book/pdf/chap6.pdf*).

1.7 SUMMARY

This book attempts the following:

1. *demystify Bayesian analyses* by showing their application in the most widely used general-purpose Bayesian software WinBUGS, in a gentle tutorial-like style and in parallel with classical analyses using program R, for a large set of ecological problems that range from very simple to moderately complex;
2. enhance your understanding of the *core of modern applied statistics*: linear, generalized linear, linear mixed, and generalized linear mixed models and features common to all of them, such as statistical distributions and the design matrix;
3. demonstrate the *great value of simulation*; and
4. thereby building a solid grounding of the use of WinBUGS (and R) for relatively simple models, so you can tackle more complex ones, and to help *free the modeler in you*.