CHAPTER

2

Introduction to the Bayesian Analysis of a Statistical Model

OUTLINE

This is a practical book that does not cover the theory of Bayesian analysis or computational methods in any detail. Nevertheless, a very brief introduction is in order. For more detailed expositions of these topics see: Royle and Dorazio (2008), King et al. (2009), or Link and Barker (2010),

Introduction to WinBUGS for Ecologists
DOI: 10.1016/B978-0-12-378605-0.00002-8

which are all specifically aimed at ecologists. In this chapter, I will first motivate a statistical view of ecology and the world at large and define a statistical model, then, contrast classical and Bayesian statistics, briefly touch upon Bayesian computation, sketch the steps of a typical Bayesian analysis, and finally, end with a brief pointer to special topics illustrated in this book.

2.1 PROBABILITY THEORY AND STATISTICS

Both probability theory and statistics are sciences that deal with uncertainty. Their subject is the description of stochastic systems, i.e., systems that are not fully predictable but include random processes that add a degree of chance—and therefore, uncertainty—in their outcome. Stochastic systems are ubiquitous in nature; hence, probability and statistics are important not only in science but also to understand all facets of life.

Indeed, stochastic systems are everywhere! They may be the weather (" will it rain tomorrow?"), politics (" will my party win?"), life (" will she marry me?"), sports ("will my football team win?"), an exam ("will I pass?"), the sex of an offspring ("will I have a daughter?"), body size of an organism, and *many, many more*. Indeed, it is hard to imagine anything observable in the world that is not at least in part affected by chance, i.e., at least partly unpredictable. For such observables or *data*, probability and statistics offer the only adequate framework for rigorous description, analysis, and prediction (Lindley, 2006).

To formally interpret any observation, we *always* need a model, i.e., an abstract description of how we believe our observations are a result of observable and unobservable quantities. The latter are called parameters, and one main aim of analyzing the model is to obtain numerical estimates for them. A model is always an abstraction and thus strictly always wrong. However, according to one of the most famous sayings in statistics, some models are useful and our goal must be to search for them. Useful models provide greater insights into a stochastic system that may otherwise be too complex to understand or to predict.

Both probability theory and statistics deal with the *characteristics of a stochastic system* (described by the parameters of a model) and its outcomes (the observed data), but these two fields represent different perspectives on stochastic systems. Probability theory specifies parameters and a model and then examines a variable outcome, whereas statistics takes the data, assumes a model, and then tries to infer the system properties, given the model. Parameters are key descriptors of the stochastic system about which one wants to learn something. Hence, statistics deals with making inferences (i.e., probabilistic conclusions about system components—parameters) based on a model and the observed outcome of a stochastic system.

2.2 TWO VIEWS OF STATISTICS: CLASSICAL AND BAYESIAN

In statistics, there are two main views about how one should learn about the parameter values in a stochastic system: classical (also called conventional or frequentist) and Bayesian statistics. Although practical applications of Bayesian statistics in ecology have greatly increased only in recent years, Bayesian statistics is, in fact, very old and was the dominating school of statistics for a long time. Indeed, the foundations of Bayesian statistics, the use of conditional probability for inference embodied in Bayes rule, were laid as early as 1763 by Thomas Bayes, an English minister and mathematician. In contrast, the foundations of classical statistics were not really laid until the first half of the twentieth century. So what are the differences?

Both classical and Bayesian statistics view data as the observed realizations of stochastic systems that contain one or several random processes. However, in classical statistics, the quantities used to describe these random processes (parameters) are fixed and unknown constants, whereas in Bayesian statistics, parameters are themselves viewed as unobserved realizations of random processes. In classical statistics, uncertainty is evaluated and described in terms of the frequency of hypothetical replicates, although these inferences are typically only described from knowledge of a single data set. Therefore, classical statistics is also called *frequentist statistics*. In Bayesian statistics, uncertainty is evaluated using the posterior distribution of a parameter, which is the conditional probability distribution of all unknown quantities (e.g., the parameters), given the data, the model, and what we knew about these quantities before conducting the analysis.

In other words, classical and Bayesian statistics differ in their definition of probability. In classical statistics, probability is the relative frequency of a feature of observed data. In contrast, in Bayesian statistics, probability is used to express one's uncertainty about the likely magnitude of a parameter; no hypothetical replication of the data set is required.

Under Bayesian inference, we fundamentally distinguish observable quantities x from unobservable quantities θ. Observables x are the data, whereas unobservables θ can be statistical parameters, missing data, mismeasured data, or future outcomes of the modeled system (predictions); they are all treated as random variables, i.e., quantities that can only be determined probabilistically. Because parameters θ are random variables under the Bayesian paradigm, we can make probabilistic statements about them, e.g., say things like "the probability that this population is in decline is 24%." In contrast, under the classical view of statistics, such statements are impossible in principle because parameters are fixed and only the data are random.

Central to both modes of inference is the sampling distribution $p(x \mid \theta)$ of the data x as a function of a model with its parameters θ. For instance, the sampling distribution for the total number of heads among 10 flips of a fair coin is the Binomial distribution with $p = 0.5$ and trial size $N = 10$. This is the distribution used to describe the effects of chance on the outcome of the random variable (here, sum of heads). In much of classical statistics, the likelihood function $p(x \mid \theta)$ is used as a basis for inference. The likelihood function is the same as the sampling distribution of the observed data x, but "read in the opposite direction": That value $\hat{\theta}$, which yields the maximum of the likelihood function for the observed data x is taken as the best estimate for θ and is called the *maximum likelihood estimate* (MLE) of the parameter θ. That is, much of classical inference is based on the estimation of a single point that corresponds to the maximum of a function. Note that θ can be a scalar or a vector.

The basis for Bayesian inference is Bayes rule, also called Bayes' theorem, which is a simple result of conditional probability. Bayes rule describes the relationship between the two conditional probabilities $p(A \mid B)$ and $p(B \mid A)$, where $\mid$ is read as "given":

$$p(A \mid B) = \frac{p(B \mid A)p(A)}{p(B)}.$$

This equation is an undisputed fact and can be proven from simple axioms of probability. However, what used to be more controversial, and partly still is (e.g., Dennis, 1996; de Valpine, 2009; Lele and Dennis, 2009; Ponciano et al., 2009), is how Bayes *used* his rule. He used it to derive the *probability of the parameters θ, given the data x*, that is, the *posterior distribution $p(\theta \mid x)$*:

$$p(\theta \mid x) = \frac{p(x \mid \theta)p(\theta)}{p(x)}.$$

We see that the posterior distribution $p(\theta \mid x)$ is proportional to the product of the likelihood function $p(x \mid \theta)$ and the prior distribution of the parameter $p(\theta)$. To make this product a genuine probability distribution function, with an integral equal to 1, a normalizing constant $p(x)$ is needed as a denominator; this is the probability of observing one's particular data set x. Ignoring the denominator (which is just a constant and does not involve the unknowns θ), Bayes, rule as applied in Bayesian statistics can be paraphrased as

$$\text{Posterior distribution} \propto \text{Likelihood} \times \text{Prior distribution},$$

where $\propto$ reads as "is proportional to." Thus, Bayesian inference works by using the laws of probability to combine the information about parameter θ contained in the observed data x, as quantified in the likelihood function

$p(x \mid \theta)$, with what is known or assumed about the parameter before the data are collected or analyzed, i.e., the prior distribution $p(\theta)$. This results in a rigorous mathematical statement about the probability of parameter θ, given the data, the posterior *distribution* $p(\theta \mid x)$. Hence, while classical statistics works by estimating a single point for a parameter (which is an unknown constant), Bayesian statistics makes inference about an entire distribution instead, because parameters are random variables described by a statistical distribution.

A prior distribution does not necessarily imply a temporal priority, instead, it simply represents a specific assumption about a model parameter. Bayes rule tells us how to combine such an assumption about a parameter with our current observations into a logical, quantitative conclusion. The latter is represented by the posterior distribution of the parameter.

I find it hard not to be impressed by the application of Bayes rule to statistical inference, because it so *perfectly mimics the way in which we learn* in everyday life! In our guts, we always weigh any observation we make, or new information we get, with what we know to be the case or believe to know. For instance, if someone tells me that he went to the zoo and saw an elephant that stood 5 m tall, I believe this information and find the observation remarkable. However, if someone claimed that he just saw an elephant that stood 10 m tall, I don't believe him. This shows that human psychology works exactly as Bayes rule applied to statistical inference: we always weigh new information by its prior probability in drawing our conclusions (here, "Oh, that's amazing!" or "You must be mad!"). An elephant height of 10 m has a prior probability close to zero *to me*, hence, I am not all too impressed by this claim (except as to find it outrageous). Note, however, that I am not particularly knowledgeable about elephants. Perhaps someone with more specialist knowledge about pachyderms would already have serious doubts about the former claim (I haven't checked). This is the reason for why many Bayesians emphasize that all probability is subjective, or personal: it depends on what we knew before observing a datum (Lindley, 2006). It is easy to find plenty more examples of where we naturally *think* according to Bayes rule.

Inference in Bayesian statistics is a simple probability calculation, and one of the things Bayesians are most proud of is the parsimony and internal logic of their framework for inference. Thus, the entire Bayesian theory for inference can be derived using just three axioms of probability (Lindley, 1983, 2006). Bayes rule can be deduced from them, and the entire framework for Bayesian statistics, such as estimation, prediction, hypothesis testing, is based on just these three premises. In contrast, classical statistics lacks such an internally coherent body of theory.

However, the requirement to determine a prior probability $p(\theta)$ for model parameters ("prior belief") has caused fierce opposition to the

Bayesian paradigm because this was (and partly still is) seen to bring into science an unwanted subjective element. However, as we shall see, it is easy to exaggerate this issue, for several reasons. First, objective science or statistics is an illusion anyway: there are always decisions to be made, e.g., what questions to ask, what factor levels to study, whether to transform a response, and literally myriads more. Each one of these decisions may have an effect on the outcome of a study. Second, it is possible to use the Bayesian machinery for inference (Bayes rule and Markov chain Monte Carlo [MCMC] computing algorithms, see later) with so-called flat priors (also vague, diffuse, uninformative, minimally informative, or low-information priors). Such priors represent our ignorance about a parameter or our wish to let inference, i.e., the posterior distribution, be dominated by the observed data. Actually, this is exactly what we do throughout this book. Third, the prior is seen by some statisticians as a strength rather than a weakness of the Bayesian framework (Link and Barker, 2010): it lets one formally examine the effect on one's conclusions of different assumptions about the parameters. Also, anybody using informative priors must say so and justify this choice. When the choice of priors is suspected to have an undue influence on the posterior distribution, it is good practice to conduct a sensitivity analysis to see how much one's conclusions are changed when a different set of priors is used.

Nevertheless, it is fair to say that there can be challenges involving the priors. First, one possible problem is that priors are not invariant to transformation of parameters. A prior that is uninformative for θ may well be informative for a one-to-one transformation $g(\theta)$ of θ, such as $\log(\theta)$ or $1/\theta$. Hence, it is possible to introduce information into an analysis without intending to do so. Especially in complex models—and these are the ones where a Bayesian treatment and the Bayesian model fitting algorithms offer most rewards—it is quite possible that one unknowingly introduces unwanted information by the choice of ostensibly vague priors. Hence, for more complex models, a sensitivity analysis of priors is even more useful. Still, these challenges are not seen as insurmountable by many statisticians, and Bayesian statistics has now very much entered the mainstream of statistical science. This can be seen immediately when browsing journals such as *Biometrics*, *Biometrika*, or the *Journal of the American Statistical Association*, which contain both frequentist and Bayesian work. Many statisticians now use Bayesian and classical statistics alike, and some believe that in the future, we will see some kind of merging of the paradigms (e.g., Little, 2006).

Finally, in view of the profound philosophical difference between the two paradigms for statistical inference, it is remarkable how little parameter estimates actually differ numerically in practical applications when vague priors are used in a Bayesian analysis. We shall see this in almost every example in this book. Indeed, MLEs are an approximation to the mode of the posterior distribution of a parameter when vague priors are assumed. This is one of the reasons for the ironic claim made by I. J. Good "People who don't know they are Bayesians are called non-Bayesians."

2.3 THE IMPORTANCE OF MODERN ALGORITHMS AND COMPUTERS FOR BAYESIAN STATISTICS

For most modeling applications, the denominator in Bayes rule, $p(x)$, contains high-dimensional integrals which are analytically intractable. Historically, they had to be solved by more or less adequate numerical approximations. Often, they could not be solved at all. Ironically therefore, for a long time Bayesians thought that they had the better solutions in principle than classical statisticians but unfortunately could not practically apply them to any except very simple problems for want of a method to solve their equations.

A dramatic change of this situation came with the advent of simulation-based approaches like MCMC and related techniques that draw *samples from the posterior distribution* (see for instance the article entitled "Bayesian statistics without tears" by Smith and Gelfand, 1992). These techniques circumvent the need for actually computing the normalizing constant in Bayes rule. This, along with the ever-increasing computer power which is required for these highly iterative techniques, made the Bayesian revolution in statistics possible (Brooks, 2003).

It seems fair to say that the ease with which difficult computational problems are solved by MCMC algorithms is one of the main reasons for the recent upsurge of Bayesian statistics in ecology, rather than the ability to conduct an inference without pretending one is completely stupid (i.e., has no prior knowledge about the analyzed system). Indeed, so far there are only few articles in ecological journals that have actually used this asset of Bayesian statistics, i.e., have formally injected prior knowledge into their Bayesian analyses. They include Martin et al. (2005); McCarthy and Masters (2005); Mazzetta et al. (2007); and Swain et al. (2009). Nevertheless, it is likely that analyses with informative priors will become more common in the future.

2.4 MARKOV CHAIN MONTE CARLO (MCMC) AND GIBBS SAMPLING

MCMC is a set of techniques to simulate draws from the posterior distribution $p(\theta \mid x)$ given a model, a likelihood $p(x \mid \theta)$, and data x, using dependent sequences of random variables. That is, MCMC yields a sample from the posterior distribution of a parameter. MCMC was developed in 1953 by the physicists Metropolis et al., and later generalized by Hastings (1970), so one of the main MCMC algorithms is called the Metropolis–Hastings algorithm. Many different flavors of MCMC are available now. One of the most widely used MCMC techniques is Gibbs sampling (Geman and Geman, 1984). It is based on the idea that to solve a large problem, instead of trying to do all at once, it is more efficient to break the problem down into smaller subunits and solve each one in turn. Here is a sketch of how Gibbs sampling works taken from a course taught in 2005 by Nicky Best and Sylvia Richardson at Imperial College in London.

Let our data be x and our vector of unknowns θ consist of k subcomponents $\theta = (\theta_1, \theta_2, \ldots, \theta_k)$, hence we want to estimate k parameters.

1. Choose starting (initial) values $\theta_1^{(0)}, \theta_2^{(0)}, \ldots, \theta_k^{(0)}$
2. Sample $\theta_1^{(1)}$ from $p(\theta_1 \mid \theta_2^{(0)}, \theta_3^{(0)}, \ldots, \theta_k^{(0)}, x)$
Sample $\theta_2^{(1)}$ from $p(\theta_2 \mid \theta_1^{(1)}, \theta_3^{(0)}, \ldots, \theta_k^{(0)}, x)$
....
Sample $\theta_k^{(1)}$ from $p(\theta_k \mid \theta_1^{(1)}, \theta_2^{(1)}, \ldots, \theta_{k-1}^{(1)}, x)$
3. Repeat step 2 many times (e.g. 100s, 1000s)
—eventually obtain a sample from $p(\theta \mid x)$

Step 2 is called an update or iteration of the Gibbs sampler and after convergence is reached, it leads to one draw (=sample) consisting of k values from the joint posterior distribution $p(\theta \mid x)$. The conditional distributions in this step are called "full conditionals" as they condition on all other parameters. The sequence of random draws for each of k parameter resulting from step 3 forms a Markov chain.

So far, a very simplistic summary of a Bayesian statistical analysis as illustrated in this book would go as follows:

1. We use a degree-of-belief definition of probability rather than a definition of probability based on the frequency of events among hypothetical replicates.
2. We use probability distributions to summarize our beliefs or our knowledge (or lack thereof) about each model parameter and apply Bayes rule to update that knowledge with observed data to obtain the posterior distribution of every unknown in our model. The posterior distribution quantifies all our knowledge about these unknowns given the data, our model, and prior assumptions. All statistical inference is based on the posterior distribution.
3. However, posterior distributions are virtually impossible to compute analytically in all but the simplest cases; hence, we use simulation (MCMC) to draw series of dependent samples from the posterior distribution and base our inference on that sample.

How do we construct such a Gibbs sampler or other MCMC algorithm? I am told by statistician colleagues that this is surprisingly easy (but that the art consisted of constructing an *efficient* sampler). However, for most ecologists, this will arguably be prohibitively complicated. And this is where WinBUGS comes in: WinBUGS constructs an MCMC algorithm for us for the model specified and our data set and conducts the iterative simulations for as long as we desire and have time to wait. Essentially, WinBUGS is an MCMC blackbox (J. A. Royle, pers. comm.).

2.5 WHAT COMES AFTER MCMC?

Once we are done with MCMC, we have a series of random numbers from the joint posterior distribution $p(\theta|x)$ that may look like this for a two-parameter model such as the model of the mean in Chapters 4 and 5 (showing only the first six draws):

$$\mu: 4.28, 6.09, 7.37, 6.10, 4.72, 6.67, \ldots$$
$$\sigma^2: 10.98, 11.23, 15.26, 9.17, 14.82, 18.19, \ldots$$

Now what should we do with these numbers?

Essentially, we have to make sure that these numbers come from a stationary distribution, i.e., that the Markov chain that produced them was at an equilibrium. If that is the case, then this is our estimate of the posterior distribution. Also, these numbers should not be influenced by our choice of initial parameter values supplied to start the Markov chains (the initial values); remember that successive values are correlated. This is called convergence monitoring. Once we are satisfied, we can summarize our samples to estimate any desired feature of the posterior distribution we like, for instance, the mean, median, or mode as a measure of central tendency as a Bayesian point estimate or the standard deviation of the posterior distribution as a Bayesian measure of the uncertainty of a parameter estimate. Then, we can compute the posterior distribution for derived variables. For instance, if parameter γ is the ratio of α and β and we are interested in γ, we can simply divide α by β for each iteration in the Markov chain to obtain a sample of γ and then summarize that for inference about γ. We can also compute inferences for very complicated functions of parameters, such as the probability that γ exceeds some threshold value. Next, we briefly expand on each of these topics.

2.5.1 Convergence Monitoring

The first step in making an inference from an MCMC analysis is to ensure that an equilibrium distribution has indeed been reached by the

Markov chain, i.e., that the chain has *converged*. For each parameter, we started the chain at an arbitrary point (the initial value or *init* chosen for each parameter), and because successive draws are dependent on the previous values of each parameter, the actual values chosen for the inits will be noticeable for a while. Therefore, only after a while is the chain independent of the values with which it was started. These first draws ought to be discarded as a *burn-in* as they are unrepresentative of the equilibrium distribution of the Markov chain.

There are several ways to check for convergence. Most methods use at least two parallel chains, but another possibility is to compare successive sections of a single long chain. The simplest method is just to inspect plots of the chains visually: they should look like nice oscillograms around a horizontal line without any trend. Visual checks are routinely used to confirm convergence. For example, Fig. 2.1 shows the time-series plot for five parallel Markov chains for a parameter in a dynamic occupancy model (MacKenzie et al., 2003) fitted to 16 years worth of Swiss wallcreeper data (see Chapter 8). Convergence seems to be achieved after about 60 iterations.

Another, more formal check for convergence is based on the Gelman–Rubin (or Brooks–Gelman–Rubin) statistic (Gelman et al., 2004), called Rhat when using WinBUGS from R via R2WinBUGS (see Chapter 5). It compares between- and within-chain variance in an analysis of variance fashion. Values near 1 indicate likely convergence, and 1.1 is considered by some as an acceptable threshold (Gelman et al., 2004; Gelman and Hill, 2007). With this approach, it is important to start the parallel chains at different selected or at random places.

Convergence monitoring may be a thorny issue and there are horror stories about how difficult it can be to make sure that convergence has actually been achieved. I have repeatedly found cases where Rhat

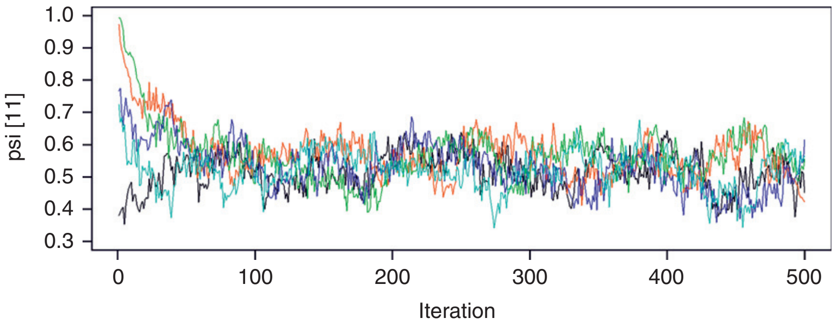

FIGURE 2.1 Time-series plot of five Markov chains for an occupancy parameter in a dynamic occupancy model fitted to Swiss wallcreeper data using WinBUGS (from Kéry et al., 2010a). Typically, the chains of all parameters do not converge equally rapidly.

erroneously indicated convergence; see Chapters 11 and 21 for examples. However, it is also easy to exaggerate this challenge, and with modern computing power, many models can be run for 100 000 iterations or more. Insuring convergence in MCMC analyses is in a sense akin to making sure that the global maximum in a likelihood function has been found in classical statistical analyses. In both, it can be difficult to determine that the desired goals have been achieved.

2.5.2 Summarizing the Posterior for Inference

Again, the aim of a Bayesian analysis is *not the estimate of a single point*, as the maximum of the likelihood function in classical statistics, but the estimate *of an entire distribution*. That means that every unknown (e.g., parameter, function of parameters, prediction, residual) has an entire distribution. This usually appears a bit odd at first. The posterior can be summarized graphically, e.g., using a histogram or a kernel-smoother. Alternatively, we can use mean, median, or mode as a measure of central tendency of a parameter (i.e., as a point estimate) and the standard deviation of the posterior as a measure of the uncertainty in the estimate, i.e., as the standard error of a parameter estimate. (Beware of challenging cases such as estimating a parameter that represents a standard deviation, e.g., the square root of a variance component. We will obtain the posterior distribution of that standard deviation, which will itself have a standard deviation that is used as the standard error of the estimate of the standard deviation … simple, eh?). Finally, the Bayesian analog to a 95% confidence interval is called a (Bayesian) credible interval (CRI) and is any region of the posterior containing 95% of the area under the curve. There is more than one such region, and one particular CRI is the highest-posterior density interval (HPDI). However, in this book, we will only consider 95% CRIs bounded by the 2.5 and the 97.5 percentile points of the posterior sample of a parameter.

2.5.3 Computing Functions of Parameters

As mentioned earlier, one of the greatest features of the Bayesian mode of inference using MCMC is the ease with which *any function* of model parameters can be computed along with their standard errors exactly, while fully *accounting for all the uncertainty involved in computing this function* and without the need for any approximations such as the delta method (Powell, 2007). To get a posterior sample for a population growth rate r from two estimates of population size, N_1 and N_2, we simply compute the ratio of the two at every iteration of the Markov chain and summarize the resulting posterior sample for inference about r.

2.5.4 Forming Predictions

Predictions are expected values of the response for future samples or of hypothetical values of the explanatory variables of a model, or more generally, of any unobserved quantity. Predictions are very important for (a) presentation of results from an analysis and (b) to even understand what a model tells us. For example, the biological meaning of an interaction or a polynomial term can be difficult to determine from a set of parameter estimates. Because predictions are functions of parameters and of data (values of covariate), their posterior distributions can again be used for inference with the mean and the 95% CRIs often used as the predicted values along with a 95% prediction interval.

2.6 SOME SHARED CHALLENGES IN THE BAYESIAN AND THE CLASSICAL ANALYSIS OF A STATISTICAL MODEL

Other, much more general and also more difficult topics in a Bayesian analysis include model criticism (checking whether the chosen model is adequate for a data set), hypothesis testing and model selection, and checking parameter identifiability (e.g., making sure that there are enough data to actually estimate a parameter). These topics are also challenging in classical statistics, although they are frequently neglected.

2.6.1 Checking Model Adequacy

In models with a single error term (e.g., linear model [LM] and generalized linear model [GLM]), the usual residual diagnostics can be applied, e.g., plots of residuals vs. fitted values, histograms of the residuals, and so on, can be produced. We will see some examples of this in this book. In hierarchical models (i.e., models that include random effects other than residuals), checking model adequacy is more difficult and may have to involve (internal) cross-validation, validation against external data, or posterior predictive checks, see Gelman et al. (1996, 2004). We will see several examples for posterior predictive checks, including computation of the Bayesian p-value, which is not about hypothesis testing as in a classical analysis, but a measure of how well a model fits a given data set, i.e., a measure for goodness-of-fit.

2.6.2 Hypothesis Tests and Model Selection

As for classical statistics with a confidence interval, in the Bayesian paradigm, hypothesis tests can be conducted based on a CRI: if it overlaps zero, then the evidence for an effect of that parameter is ambiguous. In addition, we can make direct probability statements about the magnitude of a parameter as mentioned above.

To measure the evidence for a single or a collection of effects in the model, we can use an idea described by Kuo and Mallick (1998) and also Dellaportas et al. (2002): premultiply each regression parameter with a binary indicator w and give w a Bernoulli($p = 0.5$) prior. The posterior of w will then measure the probability that the associated effect belongs in the model; see Royle (2008) for an example of this. Based on the MCMC output from such a model run, model-averaged parameter estimates can also be produced. It must be noted, though, that implementing this feature greatly slows down MCMC samplers.

For model selection in nonhierarchical models, that is, models with a single random component, e.g., LMs and GLMs, there is a Bayesian analog to the Akaike's information criterion (AIC) called the DIC (deviance information criterion; Spiegelhalter et al., 2002). Similar to the AIC, the DIC is computed as the sum of the deviance plus twice the effective number of parameters (called *pD*) and expresses the trade-off between the fit of a model and the variance of (i.e., uncertainty around) its estimates. All else being equal, a more complex model fits better than a simpler one but has less precise parameter estimates, so the best choice will be some intermediate degree of model complexity.

The DIC as computed by WinBUGS seems to work well for nonhierarchical models, but unfortunately, for models more complex than GLMs, especially hierarchical models (e.g., linear mixed models and generalized linear mixed models), the DIC needs to be computed in a different and more complicated manner; see Millar (2009). Thus, in this book, we are not going to pay special attention to the DIC, nor to the estimate of the effective number of parameters (*pD*). However, you are invited to observe, for any of the models analyzed, whether *pD* or the DIC score computed make sense or not, for instance, when you add or drop a covariate or change other parts of a model.

There are other ways to decide on how much complexity is warranted in a model, one of which goes under the name reversible-jump (or RJ-) MCMC (King et al., 2009; Ntzoufras, 2009). Simple versions of RJ-MCMC can be implemented in WinBUGS, see *http://www.winbugs-development.org.uk/*.

No doubt the lack of a semiautomated way of conducting model selection or model averaging (except for RJ-MCMC) will come as a disappointment to many readers. After all, many ecologists have come to think that the problem of model selection has been solved and that this solution has a name, AIC (see extensive review by Burnham and Anderson, 2002). However, this rosy impression is probably too optimistic as argued, for instance, by Link and Barker (2006). It is perhaps instructive for an

TABLE 2.1 Examples for the Illustration of Some Important Special Topics

Topic	Location (Chapters)
Importance of detection probability when analyzing counts	13, 16, 16E, 17, 18, 20, and 21
Random effects/hierarchical models/mixed models	9, 12, 16, 19, 20, and 21
Computing residuals	7, 8, 13, 18, and 20
Posterior predictive checks, including Bayesian *p*-value	8, 13, 18, 20, and 21
Forming predictions	8, 10, 13, 15, 20, and 21
Prior sensitivity	5E, 20, and 21
Nonconvergence or convergence wrongly indicated	11 and 21
Missing values	4E and 8E
Standardization of covariates	11
Use of simulation for assessment of bias (or estimability) in an estimator	10

E denotes the exercises at the end of each chapter.

ecologist to browse through Kadane and Lazar (2004); this article published in one of the premier statistical research journals reviews model selection and clearly shows that in the field of statistics (as opposed to parts of ecology), the challenge of model selection is not yet viewed as having been solved.

2.6.3 Parameter Identifiability

A parameter can be loosely said to be identifiable when there is enough information in the data to determine its value unambiguously. This has nothing to do with the precision of the estimate. For instance, in the equation $a + b = 7$, no parameter is estimable, and we would need an additional equation in a and/or b to be able to determine the values of the two parameters, i.e., to make them estimable.

Strictly speaking, parameter identifiability is not an issue in the Bayesian framework because, in principle, we can always compute a posterior distribution for a parameter. At worst, the posterior will be the same as the prior, but then we haven't learned anything about that parameter. A common check for identifiability is therefore to compare the prior and the posterior distribution of a parameter: if the two coincide approximately, there does not seem to be any information in the data about that parameter. Assessing parameter identifiability is another difficult topic in the Bayesian world, and one where more research is needed, but the same state of affairs exists also for any complex statistical model analyzed by classical methods (see for instance Dennis et al., 2006). In WinBUGS, nonconvergence may not only indicate lack of identifiability of one or more parameters but also other problems. Perhaps one of the simplest ways of finding out whether a parameter is indeed identifiable is by simulation: we simulate a data set and see whether we are able to recover parameter values that resemble those used to generate the data. To distinguish sampling and estimation error from lack of estimability, simulations will have to be repeated many, e.g., 10 or 100, times.

2.7 POINTER TO SPECIAL TOPICS IN THIS BOOK

In this book, we learn extensively by example. Hence, Table 2.1. lists some special topics and shows where examples for them may be found.

2.8 SUMMARY

I have given a very brief introduction to Bayesian statistics and how it is conducted in practice using simulation-based methods (e.g., MCMC, Gibbs sampling). This chapter—and indeed the whole book—is *not* meant to deal with the theory of Bayesian statistics and the associated computational methods in any exhaustive way. Rather, books like King et al. (2009) and Link and Barker (2010) should be consulted for that.

EXERCISE

1. *Bayes rule in classical statistics*: Not every application of Bayes rule makes a probability calculation Bayesian as shown next in a classical example from medical testing. An important issue in medical testing is the probability that one actually has a disease (denoted "*D*"), given that one gets a positive test result, denoted "+" (which, depending on the test, in common language may be very negative, just think about an AIDS test). This probability is $p(D\,|\,+)$. With only three pieces of information that are often known for diagnostic tests and a given population, we can use Bayes rule to compute $p(D\,|\,+)$. We simply need to know the sensitivity of the diagnostic test, denoted $p(+\,|\,D)$, its specificity $p(-\,|\,\text{not } D)$, and the general prevalence, or incidence, of the disease in the study population, $p(D)$. Note that sensitivity and specificity are the two possible kinds of diagnostic error.

Compute the probability of having the disease, given that you got a positive test result. Assume the following values: sensitivity = 0.99, specificity = 0.95, and prevalence = 5%. Start with $p(D\,|+) = \frac{p(+|D)p(D)}{p(+)}$ and note that a positive test result, which has probability $p(+)$, can be obtained in two ways: either one has the disease (with probability $p(D)$) or one does not have it (with probability $p(\text{not } D)$). Does the result surprise you?