CHAPTER

2

Brief Introduction to Bayesian Statistical Modeling

OUTLINE

2.1 INTRODUCTION

In this chapter, we attempt to give a brief overview of the following topics: (1) role of models in science, (2) statistical models, (3) Bayesian and frequentist analysis of statistical models, (4) Bayesian computation, (5) WinBUGS, (6) advantages and disadvantages of Bayesian analysis by posterior sampling, and (7) hierarchical models. This list of topics is vast and it would be impossible to give them extensive coverage even in a whole book. For topics 3–7, unless you understand the theory of

Bayesian Population Analysis using WinBUGS
DOI: 10.1016/B978-0-12-387020-9.00002-X

frequentist and Bayesian inference fairly well, we would expect you to also read some books or parts of books that delve more deeply in that, for instance, Gelman et al. (2004), McCarthy (2007), chapter 2 of Royle and Dorazio (2008), Carlin and Louis (2009), Ntzoufras (2009), or the introductory chapters in Link and Barker (2010). There are also useful introductory articles, such as Ellison (2004) or Clark (2005).

2.2 ROLE OF MODELS IN SCIENCE

Science is about rationally explaining nature by obtaining mechanistic or other explanations for the workings of a system and/or being able to predict the results of the system. However, most observable phenomena in nature are too complex for us to understand directly by simply staring at them, or rather, the system that has generated them is too complex, that is, affected by too many factors, too variable over space or time, and so on. So, explaining always requires simplifying things. Broadly, a model is nothing but a formal simplification of a complex system that we would like to explain or whose behavior we would like to predict. Indeed, at the start of this book, we have claimed that *every* interpretation of *any* observation *always* requires a model, that is, a simplification of the system, so that everybody who offers an explanation of anything has in fact a model, whether he or she knows it or not. It could also be said that explanation is impossible without a model.

So, an explanation or a more formal model is an abstraction of nature, that is, a rendition of nature with much reduced complexity. The crucial point is that we should use a good model, that is, one in which we retain the important features of the system in nature that we want to explain and only ignore the less important features. Then, by looking at this greatly simplified toy version of nature, we hopefully get a better understanding of nature herself and also can use our toy to make predictions of future or unobserved things in nature.

There are many famous sayings about models, some of which follow here. We think that they express nicely some key features of models, statistical or not:

> *Modeling is as much art as it is science* (McCullagh and Nelder, 1989): this statement expresses the fact that there are not, nor can ever be, automatic, brain-free rules for building a model (although in some disciplines much effort is spent in this pursuit).
>
> *All models are wrong, but some are useful* (Box): this is perhaps the most famous saying about models. It emphasizes that one must not look for an exact rendering of nature in a model; it is in this sense that every model is wrong. However, by simplifying, we should get some use out of a model.

Another meaning, which is perhaps not so widely appreciated, is that not all models are useful, so we should try and find the useful ones. Of course, it also begs the question of how wrong a model can be to still be useful.

There has never been a straight line nor a Normal distribution in history, and yet, using assumptions of linearity and normality allows, to a good approximation, to understand and predict a huge number of observations (Youden): This statement again expresses the fact that models are mere approximations, but that they can be hugely successful.

Everything should be a simple as possible, but not simpler (Einstein): this statement is related to the principle of parsimony and is an important guide for creating models. Very similar is the "Occam's razor" attributed to the English logician William of Ockham, which states that the explanation of any phenomenon should make as few assumptions as possible, eliminating (or *shaving off*) those that make no difference in the observable predictions.

Nothing is gained if you replace a world that you don't understand with a model that you don't understand (we heard Maynard Smith quote this one, but he may have had it from somebody else): we like this statement because it reminds us of the importance of the principle of parsimony in modeling—"*Keep it as simple as possible*". It also expresses the notion that we must replace a world that we *do not* understand by a model that we *do* understand, that is, typically something simpler. Of course, we could also understand this statement as a call for becoming a better modeler.

Finally, here is a claim we have made elsewhere (Kéry, 2010): *It is difficult to imagine another method that so effectively fosters clear thinking about a system than the use of a model written in the language of algebra.* There are various ways to express a model; words (language), graphs, and equations are some of them. Unfortunately, the human language is often very inadequate to express the subtle details of the multitude of potential explanations (= "models") for a given system. Trying to put down on paper all the elements of an explanation in the language of algebra has the big advantage that it forces us to think much more clearly about the system we want to understand. One of the things we like most about the WinBUGS software (Section 2.5) is the BUGS language (Gilks et al., 1994). Describing a model in BUGS comes very close to describing it in simple algebra. So to us, describing a model in the BUGS language is one of the most transparent ways of building a model.

Conceptually, and written in algebra, a model looks something like the following:

$$y = f(x, \theta)$$

Here, y is a response, something that our study system has produced and whose genesis we would like to understand. Often, we are interested in predicting future responses produced by the study system or responses for particular values of one or several explanatory variables x. The response is a function f of one or several explanatory variables x and of some system descriptors, or parameters, θ. Here, f would include the particular parametric form of the relationship between y and x. In essence, then, modeling means to replace a complicated reality of very large dimension with a much smaller set of system descriptors called parameters (θ). An explicit simplified system description in algebra is called a mathematical model.

There are broadly two different objectives of modeling, and they may lead to two different modes of building a model: explanation and prediction. Explanation means understanding and will typically require simpler models than prediction (Caswell, 1988). The explanatory mode of modeling focuses on the actual model structure. It is hoped that the kind of parameters and their values have some relevance for how nature generated the observed output. The focus is more on the parameters θ. In contrast, prediction focuses on the system output, the response y, and thus aims at predicting the response as well as possible either within the sample studied or for the entire statistical population that is represented by the sample. Common to both modes of modeling is that we must first build a model and estimate its parameters θ.

There is an important distinction between what might be called implicit and explicit models. We have claimed before that any interpretation of nature requires a model, but we believe that many people are not aware of this. When we talk to somebody in the general public or to not-so-modeler-types of ecologists, we often sense a certain distrust in formal, explicit modeling explanations of nature. Also, folks often have a strong feeling that the observed data are somehow superior to an inference made from these same data under a formal model. We often hear exclamations like the following: "oh yes, but this is only a model and we all know models are wrong; better stick to the data—there at least we *know* where we're at". On the whole, we think that the reasoning behind this feeling is flawed in the sense that *any attempt at explanation requires a model*.

Of course, it is possible to build models that have little to do with reality and are useless to understand or to predict a particular system. However, there cannot ever be a conclusion, deduction, or inference from any observation alone. Data need models, simply some people have explicit models and others have only implicit models. Implicit modelers frequently do not know that they are modelers, too, and that their conclusions are always contingent upon a certain set of assumptions.

These assumptions are usually unstated and may or may not be appropriate for any particular case. But just because you do not describe assumptions explicitly does not mean that these assumptions do not exist. Worse yet, if assumptions are not made explicit, they cannot be scrutinized. Just go and ask an implicit modeler about the goodness-of-fit of his or her explanation, that is, implicit model! So, again, everybody is a modeler, but some recognize this and some do not.

2.3 STATISTICAL MODELS

Almost anything in nature is affected by such a large number of factors that we could never measure or even identify them all. The result is that virtually any system that we encounter in nature will be stochastic, that is, its outcome is to some degree unpredictable. This means that a response is best thought of as the realization of a random variable. In colloquial language, we might say that chance is involved in the generation of our observations. Chance does not mean that something has no reason for happening, in the sense that there is no cause for it: there is always a cause, simply we do not know it and therefore cannot understand it completely or predict an observation perfectly.

In our models for explaining or predicting nature, we then need a description of the combined effects of all unknown and un- or mismeasured factors. A convenient mathematical description is by use of the concept of a random variable with a probability distribution function (pdf). For a random variable, this function assigns a probability of occurring to each element of a set of outcomes that are possible. To account for the unpredictable element in our observations, a model must incorporate a stochastic component and then a mathematical model becomes a statistical model. Our sketch of a model might then become:

$$y = f(x, \theta) + \varepsilon \quad \text{with} \quad \varepsilon \sim g(\phi)$$

Here, ε is the part of the response that is not explained by the functional form of the model f, the explanatory variable(s) x, and the parameter θ, and g is a function describing that unexplained part using parameter ϕ. A statistical model is often paraphrased as

$$\text{response} = \text{systematic} + \text{stochastic}$$

That is, we imagine that our response consists of a systematic and a stochastic part. Other pairs of terms for the same idea are deterministic + random and signal + noise. We will see later (Chapter 3) that this concept must be extended in so-called hierarchical models, where it applies separately to each component model, that is, level in a hierarchy.

2.4 FREQUENTIST AND BAYESIAN ANALYSIS OF STATISTICAL MODELS

One often hears the phrase "*we analyzed the data*". We believe that data analysis should be seen as consisting of two fairly distinct activities: first, to construct a plausible model of the processes that could have produced the data we observe and second, to analyze that model, for example, to find values for its parameters or to predict what the observed data might be under specific circumstances. Of course, the two activities are intertwined, but nevertheless we think that it is useful to distinguish them conceptually. One example of where this helps is by recognizing that in a sense, there is no "Bayesian so-and-so model". Rather, we first build a model and then, we may decide to analyze it in a Bayesian or in a classical framework. So what is the difference between a Bayesian and a classical (or frequentist) analysis?

The difference between classical and Bayesian statistics really starts with the analysis of a model, if one forgets for a moment the obvious difference that any model analyzed in a Bayesian mode of inference must contain prior distributions (see below). Frequentists and Bayesians differ in the way they treat the uncertainty about what is unknown in a model, especially the uncertainty about a parameter θ.

For a frequentist, parameters are fixed and unknown quantities and uncertainty about them is expressed in terms of the variability of hypothetical replicate data sets produced by them. Uncertainty is evaluated over these hypothetical replicates, even if the only thing we ever have is a single data set. Probability is defined as the long-run frequency of events in such hypothetical replicates; therefore, classical statistics is often called frequentist statistics. Frequentists only make probability statements about the data, given fixed parameter values, but never about the parameters themselves, as one might want to. In other words, frequentists do not assign a probability to a parameter; rather, they ask about the probability of observing certain kinds of data given certain values of the unknown parameters. Probability statements such as standard errors refer to hypothetical replicate data that would be expected if certain parameter values hold; they are never directly about these parameters. In the frequentist world, it is impossible in principle to make a statement such as "I am 95% certain that this population is declining".

Bayesians define probability in a fundamentally different way. Their probability is the individual belief that an event happens or that a parameter takes a specific value. No hypothetical replicates are required in Bayesian inference (though they are useful for instance in model checking; see posterior predictive checks; Gelman et al., 1996). For a Bayesian, probability is the sole measure of uncertainty about all unknown quantities: parameters, unobservables, missing or mismeasured values, or future or unobserved

responses (predictions). Bayesians use probability as their unified measure of uncertainty. This allows them to apply the mathematical laws of probability for parameter estimation and all their statistical inference. Therefore, it is possible to make probability statements about the unknown quantities, given the data, by simple use of conditional probability.

One often reads that in frequentist statistics, a parameter is a fixed and unknown quantity, while in Bayesian statistics it is a random variable. This is misleading. Rather, also for Bayesians, parameters may represent fixed and unknown quantities, but because Bayesians describe their uncertainty, or their imperfect knowledge, about the unknown parameter in terms of probability, they are *treating* parameters as random variables (Link and Barker, 2010).

So how do frequentists and Bayesians go about parameter estimation and inference? Both usually start with the sampling distribution of the data, also called the data distribution. This is the statistical description of the mechanism that could have produced the observed data, that is, the statistical model. The sampling distribution of the data y is a function of a possibly vector-valued parameter θ. It is denoted $p(y \mid \theta)$ and read "the probability of y, conditional on (given) θ". An example might be that conditional on θ, a set of counts y has a Poisson distribution: $p(y \mid \theta) \sim \text{Pois}(\theta)$. This is often abbreviated to $y \mid \theta \sim \text{Pois}(\theta)$ or even $y \sim \text{Pois}(\theta)$.

In frequentist statistics, the likelihood function plays a central role for inference about parameters. The likelihood function is the same as the sampling distribution, but "read in reverse": we interpret the sampling distribution of the observed data as a function of the unknown parameters θ, with the data y fixed . This is denoted $L(\theta \mid y)$ and read as "the likelihood of parameter θ, given the data y". We choose as our best guess of θ that parameter value which leads to the maximum function value when plugged into the sampling distribution function for the observed data y. The likelihood is not a probability because it does not integrate to 1, and the maximum function value may be greater than 1. Frequentists estimate a single point of the likelihood function and call the value which maximizes that function the *maximum likelihood estimate* or *MLE*. In other words, the MLE represents parameter value(s) which maximize the probability of getting the data actually observed. Any other value gives a lower probability of getting one's data.

Here is an example for a simple maximum likelihood analysis. Let us assume we wanted to empirically determine the detection probability of tadpoles by releasing some in a small artificial pond and counting them later. Say, we released $n = 50$ tadpoles and then count $y = 20$ of them and we want to estimate the probability that a tadpole is seen (θ). The typical sampling distribution assumed for this scenario is a binomial, that is,

$$p(y \mid \theta) = \frac{n!}{y!(n-y)!}\theta^y(1-\theta)^{n-y}.$$

The method of maximum likelihood takes as the best estimate that value of θ, which, when plugged into this sampling function along with the data (y), yields the highest function value, that is, the highest likelihood $L(\theta \mid y)$, where $L(\theta \mid y) \propto p(y \mid \theta)$. Thus, we can plug in different possible values of θ (we know that the value must be in the range from 0 to 1), compute the value of the likelihood function, and take that values of θ for which the likelihood is maximal (Fig. 2.1). We see that the likelihood function reaches a maximum for $\theta = 0.4$, so this is the maximum likelihood estimate (MLE) of θ, often denoted $\hat{\theta}$, and written as $\hat{\theta} = 0.4$.

In this simple case, it is possible to obtain the MLE analytically. This requires that we calculate the first derivative of the likelihood function, set it to zero, and solve the equation with respect to θ. Since the binomial coefficient (the ratio of factorials just after the equal sign above) is a constant, we do not need to include it in this calculation. Thus, we have

$$L(\theta \mid y) \propto \theta^{y}(1-\theta)^{n-y}$$

$$L(\theta \mid y)\partial\theta = \theta^{y}(1-\theta)^{n-y}\left(\frac{y}{\theta} - \frac{n-y}{1-\theta}\right)$$

$$\theta^{y}(1-\theta)^{n-y}\left(\frac{y}{\theta} - \frac{n-y}{1-\theta}\right) = 0$$

$$\hat{\theta} = \frac{y}{n}$$

We see that the ratio 20/50 is the MLE of θ, which is what we have expected.

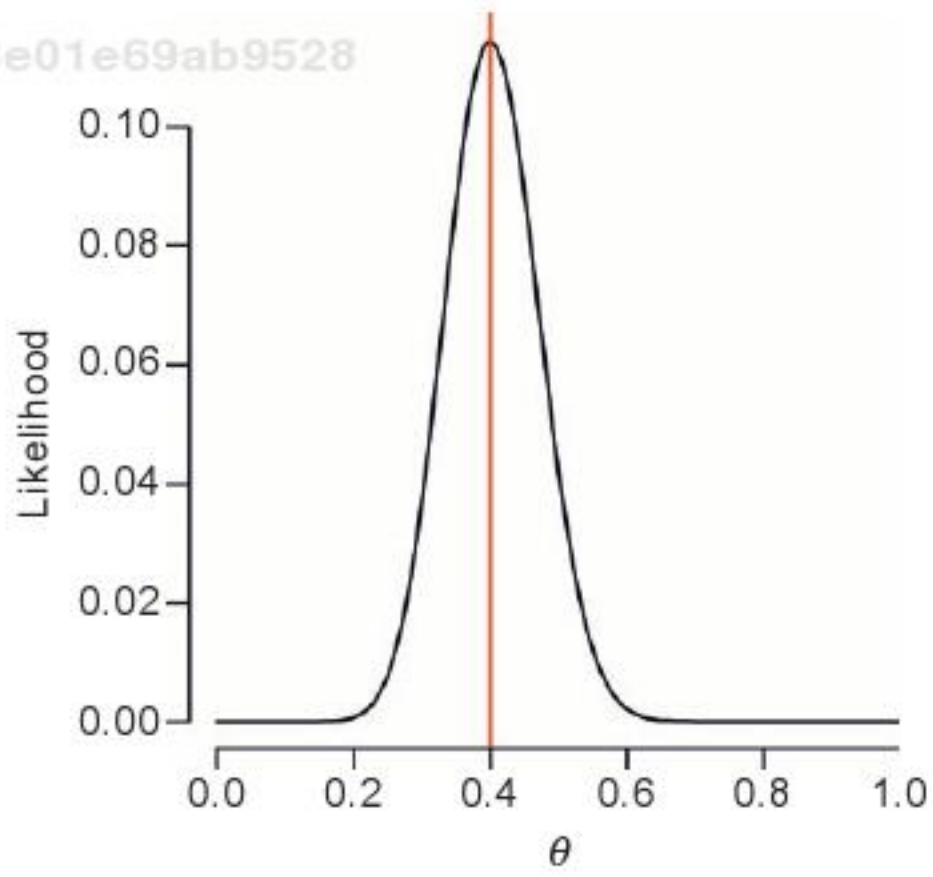

FIGURE 2.1 Binomial likelihood function for detection probability (θ) in the tadpole example, where 20 of 50 released tadpoles were seen. The MLE of θ is 0.4.

MLEs have several desirable features, such as asymptotic unbiasedness, consistency, and invariance to transformation (see, e.g., chapter 2 in Royle and Dorazio, 2008, or any book about mathematical statistics). However, the method of maximum likelihood is based on asymptotic approximations; for instance, MLEs are only unbiased and associated standard error estimates valid when sample size goes to infinity. Every ecologist knows that we rather rarely have infinite samples in ecology. How well MLEs and their standard errors perform in the typical small sample size situations in ecology is an open question in any actual application (Le Cam, 1990).

In contrast, the basis for Bayesian inference is the so-called Bayes rule. Bayes rule is attributed to the seventeenth century English minister and mathematician Thomas Bayes (Bayes, 1763). It is an undisputed mathematical fact which is easily proven from the rules of probability. Consequently, not every application of Bayes rule makes an analysis Bayesian: the application of Bayes rule to observables is undisputed. However, what Bayesians do is to apply Bayes rule also to unobservable quantities, such as, most importantly, the parameters in a statistical model. As Lindley (1983, p. 2) put it so succinctly, the recipe for every Bayesian analysis about any uncertain quantity is quite simple and mechanical:

- *What is uncertain and of interest to you? Call it* θ.
- *What do you do know? Call it* D *[…].*
- *Then calculate* $p(\theta \mid D)$.
- *How? Using the rules of probability, nothing more, nothing less.*

To describe how Bayesians learn from data using the rules of probability, we will introduce Bayes rule in the context of two sets of mutually excluding, *observable* events, A and B. Remember that a vertical bar ($|$) means "conditional on" and is read as "given".

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

This says that the conditional probability of observing A, given that B has happened or is true, $p(A \mid B)$, is equal to the conditional probability of observing B given A, $p(B \mid A)$, times the marginal probability of A, $p(A)$, divided by the marginal probability of B, $p(B)$. To better see how Bayes rule works for statistical learning from data, consider the following example which is inspired by a similar example in Pigliucci (2002). Assume that our activity after work consists of bird watching (B) or watching football on TV (F) and that this depends on whether the weather is good (g) or bad (b) on a particular night. Let us assume that you knew the following: the joint probability of good weather and us watching birds is 0.5, the marginal probability of good weather is 0.6 and the marginal probability of us watching birds is 0.7. Now if you are told that we were watching football on a particular night, what is your best guess about the weather that night?

For illustration, we present all involved probabilities in a two-by-two table with margins added (Table 2.1). In this example, we deal with mutually exclusive events (e.g., the weather cannot simultaneously be good and bad); hence, probabilities must add up within the same rows and columns, respectively. We also know that the four main cells must sum to 1, so we can fill in all cells in the table from the information just given.

Note that $p(A, B) = p(A \mid B)p(B) = p(B \mid A)p(A)$: the joint probability that A and B both occur is equal to the product of the probabilities that A occurs given that B has occurred and that B occurs, and vice versa. We had asked for your best guess about the weather on a night we were watching football. As a response, we can compute $p(b \mid F)$, the conditional probability of bad weather, given that we were watching football:

$$p(b \mid F) = \frac{p(b, F)}{p(F)} = \frac{0.2}{0.3} \approx 0.66$$

So at the outset, without any additional information, your best guess of the probability of bad weather that night would simply have been the marginal probability $p(b) = 0.4$. However, given that we watch football more frequently when the weather is bad, the knowledge that on that particular night we were watching football has increased your best guess at the probability of bad weather from 0.4 to 0.66.

This example illustrates how the information about our postwork activity (D in Lindley's recipe) influences our knowledge about the weather on a given night. In other words, it shows how our prior knowledge about the weather, $p(\theta)$, was updated by the observed data D to become $p(\theta \mid D)$. This example deals with observable events of a binary nature and nicely illustrates the use of conditional probability for learning from data, which is the basis for using Bayes rule for statistical inference about unknown quantities. In Bayesian inference, parameters take the place of the weather in our example and the data correspond to the knowledge about our

TABLE 2.1 Joint and Marginal Probabilities of Events for Two Sets of Mutually Exclusive Events, Bird Watching/Watching Football and Good/Bad Weather

	Good Weather (g)	**Bad Weather (b)**	
Go Bird Watching (B)	**0.5**	0.2	**0.7**
Watch Football (F)	0.1	0.2	0.3
	0.6	0.4	1.0

Note: The probabilities given in the text are printed in bold face and the remainder can be obtained by simple addition and subtraction.

postwork activity. In addition, we use Bayes rule to assess the uncertainty about both discrete events and continuous quantities.

When Bayes rule is applied to statistical inference about parameter θ based on the information in the data D, the probability $p(\theta \mid D)$ is called the posterior distribution of θ: it is the conditional probability of parameter θ, given the (known) data D, the prior and the model:

$$p(\theta \mid D) = \frac{p(D \mid \theta)p(\theta)}{p(D)}$$

There are three further quantities in Bayes rule, apart from the posterior distribution $p(\theta \mid D)$. In some ways, $p(D \mid \theta)$ is the opposite of the posterior: it is the probability of the data, given the parameters, or, as used here, the likelihood function. It may appear confusing that the likelihood in Bayes rule is traditionally written as $p(D \mid \theta)$, that is, in the same way as the sampling distribution of the data. Quantity $p(\theta)$ is the probability of the parameters, that is, the prior distribution. Finally, $p(D)$ is the marginal probability of the data and is defined as the integral of the numerator over θ. This is a constant used to normalize the right-hand side of Bayes rule so that the result integrates to one and becomes interpretable as a probability. As an aside, we note that it would be wrong to say that Bayesian inference is not likelihood based. Obviously, the likelihood function is a central part of Bayesian inference.

Following up the tadpole example from above, how would we estimate the unknown parameter θ in the Bayesian framework? Well, we have already defined the likelihood function. We now need to define a prior distribution of θ, that is, specify what we know *a priori* about θ and express this knowledge in a probability distribution. We know that θ must lie between 0 and 1. A useful probability distribution defined on the interval (0, 1) is the beta distribution with parameters α and β. By specifying values of α and β, we can express our knowledge about θ. Generally, we have

$$p(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}.$$

Note that we have excluded a constant of the beta distribution as it has no relevance for estimating θ. Having chosen likelihood and prior, we can analytically obtain the posterior distribution:

$$p(\theta \mid y) \propto p(y \mid \theta)p(\theta)$$

$$p(\theta \mid y) \propto \theta^{y}(1-\theta)^{n-y}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

$$p(\theta \mid y) \propto \theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1}$$

We see that this posterior distribution is also a beta distribution, with mean $\frac{y+\alpha}{n+\alpha+\beta}$ and mode $\frac{y+\alpha-1}{n+\alpha+\beta-2}$.

Absent any prior knowledge about θ, we would specify $\alpha = 1$ and $\beta = 1$ because this would result in a uniform prior distribution for θ, representing a belief that any value of θ between 0 and 1 is equally likely. A prior which says that we do not know anything about the parameter (or do not care about what might be likely values) is called noninformative, vague, flat, or diffuse prior. In the tadpole example, the resulting posterior mean is 0.404 and the mode is 0.400. The posterior mean is very close to the MLE and the mode is exactly the MLE, as we would expect—with a noninformative prior, the posterior mode of a parameter in a Bayesian analysis corresponds to the MLE of that parameter in a frequentist analysis. We notice that the impact of the prior (the values of α and β) on the posterior distribution diminishes with larger sample size (n). The posterior distribution can also be plotted and functionals (e.g., probability that $\theta > 0.5$) may be computed (see later).

So Bayesian statistical analysis is conceptually very simple: probability (via Bayes rule) is the basis for all inference about parameters and any other unknown quantities in a system analyzed. Bayesian statistics has great philosophical appeal (Link and Barker, 2010) since it is conceptually so simple (all inference is based on Bayes rule), exact (e.g., standard errors are those for your actual data set and not for some infinite version of it), and coherent (logically consistent).

Bayes rule is often paraphrased like the following:

$$\text{posterior} \propto \text{likelihood} \times \text{prior}$$

that is, the posterior distribution is proportional to the product of the likelihood and the prior. This makes it clear that in a Bayesian analysis, one's conclusion, that is, the posterior distribution, is very openly *always* a result of both the information contained in the data (as embodied in the likelihood function) *and* of our prior knowledge (our assumptions) about the unknowns in the model. We cannot conduct a Bayesian analysis without formally expressing our *a priori* uncertainty/knowledge about the parameters in the form of a probability distribution.

Several other points are noteworthy about Bayes rule as a basis for inference. First, Bayes rule formalizes the way in which humans learn. Learning always consists of updating what we knew before with what we see now. Bayes rule is thus a mathematical formalization of how we deal with new information: we always weigh the information of any new observation with the knowledge, or prior experience base, that we possessed before making that observation. The result, our conclusion, is then affected by both, and the relative importance of one or the other can vary. For instance, if we know something for almost certain, we would require large quantities of data to overthrow that prior belief. In contrast, if we do not know anything at all about a system, we might be happy to draw a conclusion based on very little data. This conclusion would then be the result almost entirely of the new

data. In Bayes rule, this weighting of information happens in a formal and mathematically rigorous way.

As another illustration of this point, note that in virtually every analysis in ecology we know something about the system analyzed and we always use that information, even in a frequentist framework. For example, if we get parameter estimates that seem to make no sense when compared with what we think we know about the system ("results do not make sense biologically"), many of us are prepared to dismiss these results in an act of *ad hoc* Bayesianism. In contrast, in a proper Bayesian analysis, prior expectations about the results could be formally introduced into an analysis. Such expectations really amount to available knowledge that is not formally used otherwise in a frequentist analysis, and this does not seem a very sensible thing to do, if we think about it.

Second, Bayes rule shows us how we can combine several pieces of information in a mathematically rigorous manner. Simply treat one piece of information as prior information and the other piece of information as the data, form a likelihood for the latter, apply Bayes rule and out comes your combination of the information in the form of the posterior distribution.

Third, priors can simply be seen as assumptions. Hence, Bayes rule represents an instrument by which we can compare formally the effect of different assumptions about model parameters by repeating the calculations with different priors.

Finally, and fourth, Bayes statistics is normative in the sense that it prescribes a mathematically rigorous way of arriving at a logical conclusion from data, a model, and *a priori* assumptions (Lindley, 2006). A Bayesian will not argue about what prior assumption one should make; this is really in the realm of the subject-matter scientist. However, once people have decided on their priors, then Bayes rule *prescribes* a mathematically rigorous way in which our statistical conclusions ought to be drawn from some observed data, using a model and these prior assumptions.

Hence, one might think that the ability to specify prior distributions was widely regarded as an advantage of the Bayesian approach. However, interestingly, priors are more often viewed as a liability of Bayesian analyses, for several reasons (Dennis, 1996). The first reason is that priors need to be decided upon. This choice is somewhat subjective even if based on past data because whether these data are relevant for the current analysis may be debatable. Hence, Bayesian analysis is intrinsically subjective (but at least explicitly so, one might want to add). Many people feel uneasy at making an explicit decision about what might be plausible values for a parameter and find it difficult to make a choice for the prior distributions.

Second, if two persons use different priors for their analysis of the same data set, they may clearly get different answers because the posterior is always the result of combining the information in the data with that in

the prior. Worse yet, even if both agreed to specify vague priors, which does not contain information, they might still end up with different answers under two different such vague priors. So, even vague priors can be challenging because different forms of specifying absence of knowledge about a parameter might not be equal. For instance, for a parameter representing a probability, we might use a uniform distribution on the interval 0–1 to say that any value is equally likely. When we specify that same parameter on the logit scale as we often do (see Chapter 3), then something analogous would be to use a uniform distribution with a large range, for example, from −1000 to 1000. Although the two prior distributions are both vague on their scales, the posterior distributions will not be exactly the same.

In spite of all this, it must be said that it is easy to exaggerate prior-related difficulties with the Bayesian approach. For once, and perhaps to console some doubters, typically parameter estimates from the Bayesian analysis of a model with vague priors numerically match pretty closely the MLEs from a frequentist analysis of the model. Second, with reasonable sample sizes, the data overwhelm the prior in their influence on the posterior distribution because the effect of the prior diminishes as sample size increases. Data cloning, a method to use Markov chain Monte Carlo simulation (see Section 2.5) to obtain MLEs without effects of priors, is based on this fact (Lele et al., 2007). Third, in any Bayesian analysis, it is customary to report the priors used. If an analyst disagrees with the choice, he or she could—at least in principle—repeat the analysis with his or her favorite prior. Finally, it is a good practice (though far from always done) to try out several priors and see what their effect on the inference is, that is, do a prior sensitivity analysis and report its results.

In this book, we follow Royle and Dorazio (2008), and indeed most applied Bayesian analysts, and specify vague priors for a natural parameterization of a model. For a parameter representing a probability, we often use a uniform(0, 1) or beta(1, 1) distribution or adopt a uniform distribution with a suitably wide range (e.g., −10, 10) for the same parameter on the logit scale. Alternatively, we often specify a flat normal distribution, that is, a normal distribution with suitably large standard deviation. What represents a suitably wide range or large standard deviation depends on the support of the likelihood function. If the likelihood of a parameter is essentially zero outside of, say, 0.4–0.6, then a uniform prior with a range between 0 and 1 may be sufficient to not affect the posterior distribution. On the other hand, if the likelihood function has nonnegligible support over a larger range, a uniform prior intended to be vague must also have a wider range. Whether a prior is sufficiently vague may be easily ascertained by repeating an analysis with narrower or wider priors and seeing whether the posterior is affected or not. In the case of a uniform prior, a posterior distribution that is truncated by either or both limits

shows that the analyst has not succeeded in choosing a vague prior. Consequently, a wider range must be chosen.

For variance parameters, the typical prior chosen used to be an inverse gamma distribution for a long time. However, currently, the preferred choice of many Bayesian analysts seems to be a suitably wide uniform for the variance parameter on the scale of the standard deviation (Gelman, 2006). This is our typical choice. We do not usually adopt explicitly informative priors. However, when the Markov chains (see Section 2.5) for a parameter fail to converge, we may narrow the range of a uniform prior or reduce the standard deviation of a flat normal prior to achieve convergence.

We saw that the basis of all Bayesian inference was the posterior probability distribution of the parameters. So once we have that, what should we do with it?

In special cases, we might choose to plot the posterior distribution for an important parameter. However, in most cases, it will be enough to summarize the posterior distribution for some or all model parameters by reporting its central tendency and its spread. Typically, for a point estimate, the posterior mean, median, or mode is used. With vague priors, the posterior distribution reflects the likelihood function directly; hence, the posterior mode is equivalent to the MLE of a corresponding frequentist analysis. The standard deviation of the posterior distribution is analogous to the standard error of a parameter estimate in a frequentist analysis. Any interval which contains 95% of the posterior mass is a Bayesian analogue to the frequentist confidence interval (CI) and is usually called a credible interval (CRI), or sometimes also a Bayesian confidence interval. Often the 2.5th and 97.5th percentiles of the posterior samples are taken as a 95% CRI; this is what we do in this book.

At this stage, and especially because posterior-based Bayesian parameter estimates often very closely match their MLE analogs numerically, it is important not to forget the exact meaning of these quantities. This has to do with the different definitions of probability. For instance, a 95% frequentist CI does *not* contain the target parameter with probability 0.95. In frequentist statistics, probability statements are about the data, or in this case, about the method, and never about the parameters. Hence, the 95% refer to the reliability of the method of constructing a 95% CI. If we sampled data from the same population 100 times and for each formed a 95% CI for a certain parameter, then about 95 intervals would indeed contain the population value and another 5 would not.

In contrast, a Bayesian 95% CRI *does* contain the parameter with probability 0.95. Also, we can make other probability statements about parameters, for instance, of the kind "I am 92% sure that this population is declining", by looking at the proportion of the mass $r < 1$ of the posterior distribution for a population growth rate r. Or else, "I am 50% sure that the growth rate lies between 0.5 and 0.8". This is a great asset of a Bayesian analysis,

especially when describing the results to the public or resource managers. The Bayesian definition of probability (and especially of uncertainty intervals) conforms much more closely to the human concept of probability than the repeated-sample definition in frequentist statistics.

2.5 BAYESIAN COMPUTATION

To analytically evaluate the posterior distribution, solving Bayes rule for all but the simplest models involves high-dimensional integrations, which can be very difficult or actually impossible to solve in most cases. The tadpole example above (Fig. 2.1) is a fairly simple example which is not difficult analytically. Most posterior distributions are, however, much more complicated, and no closed-form formulas exist. Hence, up to about 20 years ago, Bayesian analysis of a more complex model was typically not really an option. However, at the beginning of the 1990s, some statisticians rediscovered pioneering work done by physicists back in the 1950s (Smith and Gelfand, 1993). This work showed that simulation techniques could be used to draw samples from the posterior distribution instead of solving the equations. Specifically, Metropolis et al. (1953) and Hastings (1970) developed so-called Markov chain Monte Carlo (MCMC) algorithms. MCMC yields samples of arbitrary size of dependent (i.e., autocorrelated) draws from a distribution and can be constructed so that this distribution approximates the desired posterior distribution. Hence, these samples can be summarized for inference about the posterior distribution; for instance, mean and standard deviation of the samples can be interpreted as the posterior mean and posterior standard deviation, that is, as a Bayesian point and interval estimate of a parameter. Samples can also be plotted, in a raw or a smoothed histogram, for a picture of the posterior distribution.

The rediscovery and successive refinement of MCMC algorithms, along with the ever-increasing power of personal computers, sparked a revolution in statistics and also in the empirical sciences such as ecology (McCarthy, 2007). This revolution is still going on and has catapulted Bayesian methods to the center of the ecological data analysis scene (Brooks, 2003). However, for most ecologists, constructing their own MCMC algorithms would be prohibitively difficult. Hence, the Bayesian revolution has only recently reached ecology.

2.6 WinBUGS

What has brought the Bayesian revolution to ecology has a name: WinBUGS (Gilks et al., 1994; Lunn et al., 2000, 2009). WinBUGS is the Windows version of a free computer program developed as part of the

BUGS project, which means *Bayesian inference using Gibbs sampling*; see www.mrc-bsu.cam.ac.uk/bugs. WinBUGS originally used a particular variant of MCMC called Gibbs sampling (Geman and Geman, 1984), but now uses a variety of other MCMC sampling techniques. For a history of the BUGS project, an appreciation and outlook, see Lunn et al. (2009). The active development in the BUGS project now takes place with OpenBUGS; see www.openbugs.info. At the time of writing, the two BUGS sisters are nearly identical in practice and will run most code from one another fine. The same goes for a BUGS clone called JAGS (*Just another Gibbs sampler*, see www-fis.iarc.fr/~martyn/software/jags). JAGS is another MCMC engine that uses the BUGS language; hence, most BUGS code should run also in JAGS.

For most ecologists, WinBUGS is simply an ingenious MCMC blackbox. The analyst communicates with the MCMC engine by providing a data set and describing a statistical model for it using a simple and effective model definition language, the BUGS language (Gilks et al., 1994). In our opinion, the BUGS language can claim a large part of the value for ecologists of the WinBUGS software. All statistical models that we have had to do with are specified more simply and—to us—in a *much* more transparent way in the BUGS language than when using custom code for maximum likelihood estimation. For the latter, one needs to define the likelihood for a model explicitly and then use some function optimizer (for instance, `nlm` or `optim` in R) to find the MLEs (Bolker, 2008). Alternatively, one uses software that shields us from most of the complexity but makes it easy to fit models that one does not understand or that may not make sense. In contrast, BUGS code often looks trivially simple and concise. In the BUGS language, all stochastic models are described by specifying local stochastic or deterministic relationships between quantities such as parameters and data. By breaking apart an entire model into its smaller component parts, understanding is greatly enhanced for an ecologist. Moreover, the construction of even very complex models becomes relatively feasible and transparent. BUGS model descriptions are naturally hierarchical, and indeed, WinBUGS is ideal for fitting hierarchical models (see Section 2.9). We will see many examples for this later in the book. Indeed, we have found that WinBUGS frees the creative modeler in many ecologists.

Once the model is specified, WinBUGS constructs an MCMC algorithm in perfect blackbox manner and runs that for the required length. Its primary product is a long stream of numbers, one for each model parameter that we choose to estimate. If the MCMC algorithm has been constructed adequately and the chains have converged to the desired posterior distribution, then these numbers represent a random sample from these posterior distributions. There is an autocorrelation built into these numbers, since they form a Markov chain. This means

that the first part of the chains, where the effect of the arbitrarily chosen starting values will still be felt, must be discarded as a so-called burnin. Whether the burnin period is over or not can be judged by visual means, that is, by inspecting a time-series plot of the sampled values for each parameter. The plot should now randomly jump up and down around a constant mean. There are formal criteria to decide whether convergence has been reached. For instance, the Brooks–Gelman–Rubin statistic (Brooks and Gelman, 1998) is often used. It requires two or more chains for each parameter and compares the between-chain with the within-chain variance in an ANOVA fashion. At convergence, the value of this test statistic, sometimes called Rhat, is 1. After a chain has converged onto the desired target distribution, to save computer space and reduce autocorrelation, one may thin it by k, that is, keep every kth value only. Thus, one gets a smaller, but more information-dense (because less autocorrelated) sample from the posterior distribution.

WinBUGS can be used as standalone software, see McCarthy (2007), Ntzoufras (2009) or chapter 4 in Kéry (2010). However, we find it more efficient to harness it to R via the communicator package R2WinBUGS (Sturtz et al., 2005). This is how we use WinBUGS throughout this book.

There are many Bayesian statistics books that explain MCMC algorithms and give examples (e.g., McCarthy, 2007; Ntzoufras, 2009; Link and Barker, 2010; King et al., 2010); therefore, we skip this here. We feel that the importance of being able to code one's own MCMC algorithm may easily be overstated. After all, hardly any ecologist would nowadays be able to code a Newton–Raphson algorithm for fitting a GLM or a Laplace approximation for the integrals that need to be solved for obtaining mixed-model estimates. And yet, many of us routinely use these methods for our research.

Admittedly, MCMC may be somewhat more difficult than these techniques and may fail in perhaps more ways than other computing algorithms commonly used for a frequentist analysis. We do not doubt that it can be a great advantage to actually *know* how to code MCMC algorithms, not least because custom-written MCMC code often runs much faster than WinBUGS. Nevertheless, a simple intuitive understanding of the nature of MCMC techniques will often be enough for ecologists. Such an understanding may be obtained by simply using an MCMC blackbox, such as WinBUGS and experiencing the behavior of the chains in many situations for many different models. This is what we do in this book.

WinBUGS is a fantastic program, but may exhibit a fair dose of pretty idiosyncratic behavior. There are many things that one just must know in order to succeed. A collection of survival tips can be found in Appendix 1.

2.7 ADVANTAGES AND DISADVANTAGES OF BAYESIAN ANALYSES BY POSTERIOR SAMPLING

There are many advantages of the Bayesian analysis of a model by posterior simulation via MCMC techniques (Kéry, 2010). Some of them which are particularly relevant for an ecologist include the following:

- Even difficult models can be fit, including some which cannot be fitted in the classical framework.
- Derived quantities may be computed trivially easily, with full propagation of the uncertainty in the components that make up the derived quantity. This can be a very hard problem in the classical framework.
- All results are exact; there are no asymptotics involved in the estimates as for MLEs, which may be of questionable value in small-data situations so typical of ecological studies.
- The BUGS language allows the typical quantitative ecologist to actually understand the construction of even complex models so that the code can be modified to fit one's own purposes.

On the other hand, there are also challenges with the Bayesian approach. At first sight, like any new theory or method, Bayesian statistics may appear difficult. Then, the choice of priors and the sensitivity of the estimates to that choice need some thinking. MCMC engines such as WinBUGS are blackboxes and are hard to understand, leaving a certain uneasiness with people who like to understand most of what they do, and convergence of the Markov chains may be difficult to assess.

MCMC-based analyses in general can be slow compared to other ways of model fitting (see, e.g., the comparisons in Kéry, 2010). Just because WinBUGS is an extremely flexible, generic MCMC engine, it is a rather slow software when compared with custom-written MCMC algorithms. For complex models applied to large data sets, WinBUGS may become too slow to be of practical value. Novel algorithmic techniques that provide approximate analytical solutions to the integrations involved in Bayesian analysis (e.g., AD model builder; see http://admb-foundation.org/ or R-INLA, Rue et al., 2009) may then be exciting new avenues for the fitting of some classes of hierarchical models.

Some other difficult topics include the detection of parameter identifiability and model selection. Lunn et al. (2009) say that the flexibility of WinBUGS to specify even very complex models may let the user fit models that do not make sense, for instance, models with parameters that are not identifiable, that is, for which the data do not contain any information. Nonidentifiability of a parameter is often difficult to diagnose in complex models, but perhaps harder still in a Bayesian than in a frequentist analysis.

This is due in part because in a sense, the problem does not really exist in the Bayesian mode of analysis: if there is no information about a parameter in the data (the likelihood), then there is always information coming from the prior, and we still technically get a posterior distribution for that parameter. Hence, one way of checking whether a parameter is indeed identified is by comparing the prior with the posterior and seeing whether changing the prior induces large changes in the posterior (Gimenez et al., 2009b; see Section 7.9). Simulating a data set and seeing whether the analysis is able to recover estimates that resemble the known input values is perhaps one of the best ways for an ecologist to check for nonidentifiability of a parameter (e.g., Schaub, 2009).

Another big topic is model and variable selection. In the frequentist world, many ecologists use model selection criteria such as Akaike's information criterion (AIC; see review by Burnham and Anderson, 2002). Yet, it appears sometimes as if a bunch of models is thrown up into the air in the hope that AIC will do all the work of sorting through them. That such a view is overly simplistic becomes clear when reading through a review paper on model selection in the primary statistical literature (e.g., Kadane and Lazar, 2004). Model and variable selection are deep waters and even among statisticians there is no consensus view on what is the best—and practically feasible—approach. Furthermore, model selection using the AIC is an unsolved problem for mixed models due to the challenge of counting the effective number of parameters (Link and Barker, 2010). Hence, for mixed models even in the classical arena, there does not seem to be a simple approach available.

These challenges appear, if anything, even more acute when one moves to a Bayesian analysis. There is an AIC-analogue called deviance information criterion or DIC (Spiegelhalter et al., 2002), but again, its standard version computed by WinBUGS appears to be problematic for hierarchical models—and most models that ecologists nowadays want to fit have more than one random component and therefore are mixed, or hierarchical, models (see Chapter 4). The DIC can be computed for such hierarchical models (see Millar, 2009, which includes R code), but the required computations are involved and computationally very demanding. Hence, in spite of long-standing criticisms of stepwise model selection and model selection by significance tests, one may effectively be back at one of those. We can look at the significance of a parameter by checking whether its 95% CRI covers zero and based on that decide whether it is warranted in a model or not. This is what we sometimes do. Other times, we simply fit one model that is biologically plausible to us and stick to that. There are yet other approaches, for instance, Bayes factors and reversible jump Markov chain Monte Carlo (RJMCMC); see for instance, King et al. (2010) and Link and Barker (2010). But be warned, Link and Barker's chapter 7 on Bayesian multimodel selection is not easy reading. Also see

the overview by O'Hara and Sillanpää (2009). It is likely that we will see more work in the future on Bayesian model selection and hypothesis testing.

2.8 HIERARCHICAL MODELS

In hierarchical models, complex stochastic systems are decomposed into a dependent sequence of simpler submodels. This partitioning is beneficial for a better understanding of a system, for an honest accounting for all levels of uncertainty or for computational ease. A model can be hierarchical in two ways, statistically and conceptually.

In a purely statistical sense, a hierarchical model is composed of a sequence of random variables, with the realization of the random variable at one level being a parameter of the random variable at the next level down. For instance, a hierarchical model with two levels is (dropping indices):

1. $x \sim f(\omega)$
2. $y \sim g(x, \theta)$

That is, at level 1 of the hierarchy, x is a realization of a random variable described by probability distribution f with parameter ω. At level 2, y is a realization of another random variable described by probability distribution g, which depends on the realization of the first random variable, x, and on another parameter, θ. Of course, there may be more than two levels in a hierarchical model. Hierarchical models abound in ecology. For instance, a nested ANOVA model (Kéry, 2010) is an example of a hierarchical model with two levels, with f and g being a normal distribution, such that $x \sim N(\mu, \sigma_x^2)$ and $y \sim N(x, \sigma^2)$, where μ is the grand mean, σ_x^2 the variance among group means x, and σ^2 the residual variance of measurements y around the group means (note indices have been omitted for clarity).

In the context of hierarchical models for population analysis, Royle and Dorazio (2008) make the important distinction between *implicit* and *explicit hierarchical models*. Explicit hierarchical models have random variables or parameters with an explicit ecological interpretation, while implicit hierarchical models do not. As an example of an implicit hierarchical model, the Poisson GLMMs in Chapter 4 have a quantity called the expected count (λ). This is not a real ecological parameter because it is the product of population size and detection probability. In contrast, in the hierarchical models in Chapter 6, N is the sum of the latent indicator variables z and corresponds exactly to the local population size. In explicit hierarchical models, the lowest level in the hierarchy typically represents an explicit description of the binomial observation process. As a result, the ecological parameters in the model become directly interpretable and do have an ecological meaning. In addition, inference from explicit hierarchical models

is protected against possible misinterpretations due to a confounding of the ecological and the observation processes in implicit hierarchical models. All else equal, we prefer explicit over implicit hierarchical models.

There is another, conceptual sense in which a model can be hierarchical, and this has to do with exactly this accounting for the observation process. For example, the CJS model fitted via the m-array (Section 7.10) is not a hierarchical model in the statistical sense of the term, but the state-space version of the model (Section 7.2) is. With the m-array, the hierarchical genesis of the observed data is lost by aggregation when creating the m-array. Nevertheless, this model is still an explicit hierarchical model in a conceptual sense because its parameters have an explicit ecological meaning owing to the explicit modeling of the observation process. Similarly, the N-mixture model (Chapter 12) is intrinsically hierarchical, even when fit in a frequentist framework, where the latent abundance states are integrated out from the likelihood and thus the hierarchy is collapsed (Royle, 2004c).

2.9 SUMMARY AND OUTLOOK

In this chapter, we have briefly reviewed statistical models and their analysis in WinBUGS. We have claimed that any interpretation of data requires a model, either an implicit or an explicit one. Then, we briefly reviewed two philosophies for formal learning from data, with their associated methods for fitting models to data and making inferences about their parameters, that is, of obtaining estimates of the parameters and of the uncertainty around these estimates. One is maximum likelihood and the other is Bayesian inference. Bayesian inference is based on the posterior distribution, which is a product of the likelihood (representing the information contained in the data) and the prior distribution (representing what is known about the parameters beforehand). Bayesian inference uses a fact of conditional probability, Bayes rule, to let the data update our prior state of knowledge to the posterior state of knowledge. In this way, what we learn from data, the posterior distribution, is a weighted average of the prior distribution and the information of the data at hand.

We have seen that priors can be regarded both as an asset and as a liability in Bayesian inference (of course, we believe that the former outweigh the latter). We have also seen that the results of a Bayesian analysis based on the posterior distribution are much more easily explained to the public owing to the more intuitive Bayesian definition of probability. Bayesian analysis in practice nowadays means obtaining samples from the posterior distribution by simulation techniques such as Markov chain Monte Carlo (MCMC). The free WinBUGS software (along with its "sisters", OpenBUGS and JAGS) is the most widely used MCMC engine currently available.

It allows us to specify almost arbitrarily complex models using an ingenious and simple model definition language. It then constructs an MCMC algorithm, runs it for the requested length, and produces a stream of numbers which, if all went well, represents a sample from the posterior distribution of interest. These samples can be summarized for inference, for instance, posterior means and standard deviations are customarily treated as Bayesian point and interval estimates. Important advantages of the Bayesian model fitting by posterior sampling include the numerical tractability of even very complex models, exact rather than asymptotic inference, and the ease with which derived parameters can be estimated with full propagation of all uncertainty. Some disadvantages are that it may be difficult at first, it may be slow, and parameter nonidentifiability and model selection may be even harder challenges than in the frequentist framework.

We are now armed with the motivation for population analysis and have a basic understanding for how estimation and inference about model parameters is achieved in the Bayesian framework of statistics. Hence, we are ready to move on to see our first population models. In the next two chapters, we will deal with simple models for time series of counts. Importantly, these chapters will also provide an introduction to what may be the three most essential topics of applied statistical modeling: linear and generalized linear models in Chapter 3 and random effects in Chapter 4. You will meet these concepts over and over again in your statistical modeling. If you understand them in a simple model, your understanding for more complex models will be greatly enhanced.

This page intentionally left blank