PART I

GETTING STARTED

[T]he theory of probabilities is basically just common sense reduced to calculus; it makes one appreciate with exactness that which accurate minds feel with a sort of instinct, often without being able to account for it.

Laplace, 1829

1 The basics of Bayesian analysis

1.1 General principles

The general principles of Bayesian analysis are easy to understand. First, uncertainty or "degree of belief" is quantified by probability. Second, the observed data are used to update the *prior* information or beliefs to become *posterior* information or beliefs. That's it!

To see how this works in practice, consider the following example. Assume you are given a test that consists of 10 factual questions of equal difficulty. What we want to estimate is your ability, which we define as the rate θ with which you answer questions correctly. We cannot directly observe your ability θ. All that we can observe is your score on the test.

Before we do anything else (for example, before we start to look at your data) we need to specify our prior uncertainty with respect to your ability θ. This uncertainty needs to be expressed as a probability distribution, called the *prior distribution*. In this case, keep in mind that θ can range from 0 to 1, and that we do not know anything about your familiarity with the topic or about the difficulty level of the questions. Then, a reasonable "prior distribution," denoted by $p(\theta)$, is one that assigns equal probability to every value of θ. This uniform distribution is shown by the dotted horizontal line in Figure 1.1.

Now we consider your performance, and find that you answered 9 out of 10 questions correctly. After having seen these data, the updated knowledge about θ is described by the *posterior distribution*, denoted $p(\theta \mid D)$, where D indicates the observed data. This distribution expresses the uncertainty about the value of θ, quantifying the relative probability that each possible value is the true value. Bayes' rule specifies how we can combine the information from the data—that is, the likelihood $p(D \mid \theta)$—with the information from the prior distribution $p(\theta)$, to arrive at the posterior distribution $p(\theta \mid D)$:

$$p(\theta \mid D) = \frac{p(D \mid \theta)\, p(\theta)}{p(D)}. \tag{1.1}$$

This equation is often verbalized as

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{marginal likelihood}}. \tag{1.2}$$

Note that the marginal likelihood (i.e., the probability of the observed data) does not involve the parameter θ, and is given by a single number that ensures that

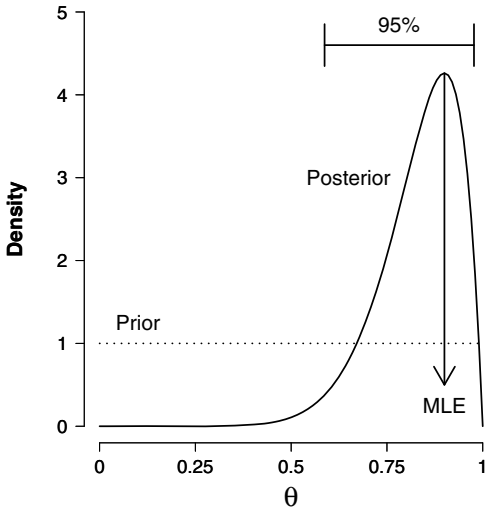

Fig. 1.1 Bayesian parameter estimation for rate parameter θ, after observing 9 correct responses and 1 incorrect response. The mode of the posterior distribution for θ is 0.9, equal to the maximum likelihood estimate (MLE), and the 95% credible interval extends from 0.59 to 0.98.

the area under the posterior distribution equals 1. Therefore, Equation 1.1 is often written as

$$p(\theta \mid D) \propto p(D \mid \theta)\, p(\theta), \tag{1.3}$$

which says that the posterior is proportional to the likelihood times the prior. Note that the posterior distribution is a combination of what we knew before we saw the data (i.e., the information in the prior distribution), and what we have learned from the data. In particular, note that the new information provided by the data has reduced our uncertainty about the value of θ, as shown by the posterior distribution being narrower than the prior distribution.

The solid line in Figure 1.1 shows the posterior distribution for θ, obtained when the uniform prior is updated with the data. The central tendency of a posterior distribution is often summarized by its mean, median, or mode. Note that with a uniform prior, the mode of a posterior distribution coincides with the classical maximum likelihood estimate or MLE , $\hat{\theta} = k/n = 0.9$ (Myung, 2003). The spread of a posterior distribution is most easily captured by a Bayesian x% credible interval that extends from the $(100 - x)/2^{\text{th}}$ to the $(100 + x)/2^{\text{th}}$ percentile of the posterior distribution. For the posterior distribution in Figure 1.1, a 95% Bayesian credible interval for θ extends from 0.59 to 0.98. In contrast to the orthodox confidence interval, this means that one can be 95% confident that the true value of θ lies in between 0.59 and 0.98.

Exercises

Exercise 1.1.1 The famous Bayesian statistician Bruno de Finetti published two big volumes entitled *Theory of Probability* (de Finetti, 1974). Perhaps surprisingly, the first volume starts with the words "probability does not exist." To understand why de Finetti wrote this, consider the following situation: someone tosses a fair coin, and the outcome will be either heads or tails. What do you think the probability is that the coin lands heads up? Now suppose you are a physicist with advanced measurement tools, and you can establish relatively precisely both the position of the coin and the tension in the muscles immediately before the coin is tossed in the air—does this change your probability? Now suppose you can briefly look into the future (Bem, 2011), albeit hazily. Is your probability still the same?

Exercise 1.1.2 On his blog, prominent Bayesian Andrew Gelman wrote (March 18, 2010): "Some probabilities are more objective than others. The probability that the die sitting in front of me now will come up '6' if I roll it ... that's about 1/6. But not exactly, because it's not a perfectly symmetric die. The probability that I'll be stopped by exactly three traffic lights on the way to school tomorrow morning: that's well, I don't know exactly, but it is what it is." Was de Finetti wrong, and is there only one clearly defined probability of Andrew Gelman encountering three traffic lights on the way to school tomorrow morning?

Exercise 1.1.3 Figure 1.1 shows that the 95% Bayesian credible interval for θ extends from 0.59 to 0.98. This means that one can be 95% confident that the true value of θ lies between 0.59 and 0.98. Suppose you did an orthodox analysis and found the same confidence interval. What is the orthodox interpretation of this interval?

Exercise 1.1.4 Suppose you learn that the questions are all true or false questions. Does this knowledge affect your prior distribution? And, if so, how would this prior in turn affect your posterior distribution?

1.2 Prediction

The posterior distribution θ contains all that we know about the rate with which you answer questions correctly. One way to use the knowledge is *prediction*.

For example, suppose you are confronted with a new set of 5 questions, all of the same difficulty as before. How can we formalize our expectations about your performance on this new set? In other words, how can we use the posterior distribution $p(\theta \mid n = 10, k = 9)$—which, after all, represents everything that we know about θ from the old set—to *predict* the number of correct responses out of the new set of $n^{\text{rep}} = 5$ questions? The mathematical solution is to integrate over the posterior,

$\int p\left(k^{\text{rep}} \mid \theta, n^{\text{rep}} = 5\right) p\left(\theta \mid n = 10, k = 9\right)\ \mathrm{d}\theta$, where k^{rep} is the predicted number of correct responses out of the additional set of 5 questions.

Computationally, you can think of this procedure as repeatedly drawing a random value θ_i from the posterior, and using that value to every time determine a single k^{rep}. The end result is $p\left(k^{\text{rep}}\right)$, the posterior predictive distribution of the possible number of correct responses in the additional set of 5 questions. The important point is that by integrating over the posterior, all predictive uncertainty is taken into account.

Exercise

Exercise 1.2.1 Instead of "integrating over the posterior," orthodox methods often use the "plug-in principle." In this case, the plug-in principle suggests that we predict $p(k^{\text{rep}})$ solely based on $\hat{\theta}$, the maximum likelihood estimate. Why is this generally a bad idea? Can you think of a specific situation in which this may not be so much of a problem?

1.3 Sequential updating

Bayesian analysis is particularly appropriate when you want to combine different sources of information. For example, assume that you are presented with a new set of 5 questions of equal difficulty. You answer 3 out of 5 correctly. How can we combine this new information with the old? Or, in other words, how do we update our knowledge of θ? Consistent with intuition, Bayes' rule entails that the prior that should be updated based on your performance for the new set is the posterior that was obtained based on your performance for the old set. Or, as Lindley put it, "today's posterior is tomorrow's prior" (Lindley, 1972, p. 2).

When all the data have been collected, however, the order in which this was done is irrelevant. The results from the 15 questions could have been analyzed as a single batch; they could have been analyzed sequentially, one-by-one; they could have been analyzed by first considering the set of 10 questions and next the set of 5, or vice versa. For all these cases, the end result, the final posterior distribution for θ, is identical. Given the same available information, Bayesian inference reaches the same conclusion, independent of the order in which the information was obtained. This again contrasts with orthodox inference, in which inference for sequential designs is radically different from that for non-sequential designs (for a discussion, see, for example, Anscombe, 1963).

Thus, a posterior distribution describes our uncertainty with respect to a parameter of interest, and the posterior is useful—or, as a Bayesian would have it, necessary—for probabilistic prediction and for sequential updating. To illustrate, in the case of our binomial example the uniform prior is a beta distribution with parameters $\alpha = 1$ and $\beta = 1$, and when combined with the binomial likelihood

this yields a posterior that is also a beta distribution, with parameters $\alpha + k$ and $\beta + n - k$. In simple *conjugate* cases such as these, where the prior and the posterior belong to the same distributional family, it is possible to obtain analytical solutions for the posterior distribution, but in many interesting cases it is not.

1.4 Markov chain Monte Carlo

In general, the posterior distribution, or any of its summary measures, can only be obtained analytically for a restricted set of relatively simple models. Thus, for a long time, researchers could only proceed easily with Bayesian inference when the posterior was available in closed-form or as a (possibly approximate) analytic expression. As a result, practitioners interested in models of realistic complexity did not much use Bayesian inference. This situation changed dramatically with the advent of computer-driven sampling methodology, generally known as Markov chain Monte Carlo (MCMC: e.g., Gamerman & Lopes, 2006; Gilks, Richardson, & Spiegelhalter, 1996). Using MCMC techniques such as Gibbs sampling or the Metropolis–Hastings algorithm, researchers can directly sample sequences of values from the posterior distribution of interest, forgoing the need for closed-form analytic solutions. The current adage is that *Bayesian models are limited only by the user's imagination.*

In order to visualize the increased popularity of Bayesian inference, Figure 1.2 plots the proportion of articles that feature the words "Bayes" or "Bayesian," according to Google Scholar (for a similar analysis for specific journals in statistics and economics see Poirier, 2006). The time line in Figure 1.2 also indicates the introduction of WinBUGS, a general-purpose program that greatly facilitates Bayesian analysis for a wide range of statistical models (Lunn, Thomas, Best, & Spiegelhalter, 2000; Lunn, Spiegelhalter, Thomas, & Best, 2009; Sheu & O'Curry, 1998). MCMC methods have transformed Bayesian inference to a vibrant and practical area of modern statistics.

For a concrete and simple illustration of Bayesian inference using MCMC, consider again the binomial example of 9 correct responses out of 10 questions, and the associated inference problem for θ, the rate of answering questions correctly. Throughout this book, we use WinBUGS to do Bayesian inference, saving us the effort of coding the MCMC algorithms ourselves.[1] Although WinBUGS does not work for every research problem application, it will work for many in cognitive sci-

[1] At this point, some readers want to know how exactly MCMC algorithms work. Other readers feel the urge to implement MCMC algorithms themselves. The details of MCMC sampling are covered in many other sources and we do not repeat that material here. We recommend the relevant chapters from the following books, listed in order of increasing complexity: Kruschke (2010a), MacKay (2003), Gilks et al. (1996), Ntzoufras (2009), and Gamerman and Lopes (2006). An introductory overview is given in Andrieu, De Freitas, Doucet, and Jordan (2003). You can also browse the internet, and find resources such as `http://www.youtube.com/watch?v=4gNpgSPal_8` and `http://www.learnbayes.org/`.

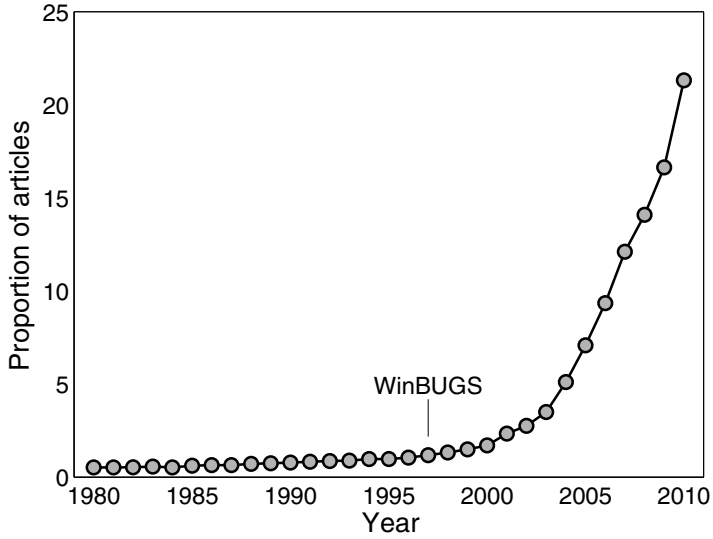

Fig. 1.2 A Google Scholar perspective on the increasing popularity of Bayesian inference, showing the proportion of articles matching the search "bayes OR bayesian -author: bayes" for the years 1980 to 2010.

ence. WinBUGS is easy to learn and is supported by a large community of active researchers.

The WinBUGS program requires you to construct a file that contains the model specification, a file that contains initial values for the model parameters, and a file that contains the data. The model specification file is most important. For our binomial example, we set out to obtain samples from the posterior of θ. The associated WinBUGS model specification code is two lines long:

```
model{
  theta ~ dunif(0,1) # the uniform prior for updating by the data
  k ~ dbin(theta,n)  # the data; in our example, k = 9 and n = 10
}
```

In this code, the "∼" or twiddle symbol denotes "is distributed as", `dunif(a,b)` indicates the uniform distribution with parameters a and b, and `dbin(theta,n)` indicates the binomial distribution with rate θ and n observations. These and many other distributions are built in to the WinBUGS program. The "#" or hash sign is used for comments. As WinBUGS is a declarative language, the order of the two lines is inconsequential. Finally, note that the values for k and n are not provided in the model specification file. These values constitute the data and they are stored in a separate file.

When this code is executed, you obtain a sequence of MCMC samples from the posterior $p(\theta \mid D)$. Each individual sample depends only on the one that immediately preceded it, and this is why the entire sequence of samples is called a *chain*.

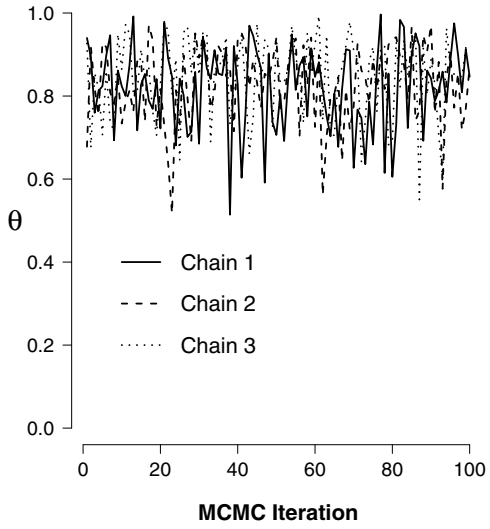

Fig. 1.3 Three MCMC chains for rate parameter θ, after observing 9 correct responses and 1 incorrect response.

In more complex models, it may take some time before a chain converges from its starting value to what is called its stationary distribution. To make sure that we only use those samples that come from the stationary distribution, and hence are unaffected by the starting values, it is good practice to diagnose convergence. This is an active area of research, and there is an extensive set of practical recommendations regarding achieving and measuring convergence (e.g., Gelman, 1996; Gelman & Hill, 2007).

A number of worked examples in this book deal with convergence issues in detail, but we mention three important concepts now. One approach is to run multiple chains, checking that their different initial starting values do not affect the distributions they sample from. Another is to discard the first samples from each chain, when those early samples are sensitive to the initial values. These discarded samples are called *burn-in* samples. Finally, it can also be helpful not to record every sample taken in a chain, but every second, or third, or tenth, or some other subset of samples. This is known as *thinning*, a procedure that is helpful when the chain moves slowly through the parameter space and, consequently, the current sample in the MCMC chain depends highly on the previous one. In such cases, the sampling process is said to be autocorrelated.

For example, Figure 1.3 shows the first 100 iterations for three chains that were set up to draw values from the posterior for θ. It is evident that the three chains are "mixing" well, suggesting early convergence. After assuring ourselves that the chains have converged, we can use the sampled values to plot a histogram, construct a density estimate, and compute values of interest. To illustrate, the three chains from Figure 1.3 were run for 3000 iterations each, for a total of 9000 samples from the posterior of θ. Figure 1.4 plots a histogram for the posterior. To visualize how the

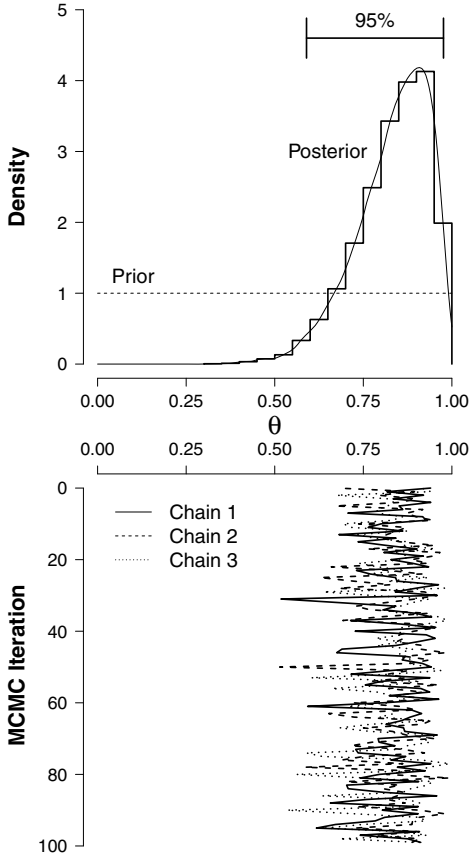

Fig. 1.4 MCMC-based Bayesian parameter estimation for rate parameter θ, after observing 9 correct responses and 1 incorrect response. The thin solid line indicates the fit of a density estimator. Based on this density estimator, the mode of the posterior distribution for θ is approximately 0.89, and the 95% credible interval extends from 0.59 to 0.98, closely matching the analytical results from Figure 1.1.

histogram is constructed from the MCMC chains, the bottom panel of Figure 1.4 plots the MCMC chains sideways; the histograms are created by collapsing the values along the "MCMC iteration" axis and onto the "θ" axis.

In the top panel of Figure 1.4, the thin solid line represent a density estimate. The mode of the density estimate for the posterior of θ is 0.89, whereas the 95% credible interval is $(0.59, 0.98)$, matching the analytical result shown in Figure 1.1.

The key point is that the analytical intractabilities that limited the scope of Bayesian parameter estimation have now been overcome. Using MCMC sampling, posterior distributions can be approximated to any desired degree of accuracy. This