unobserved quantities as random variables that truly distinguishes Bayes. We go into detail about the unique features of Bayes in the next chapter, but here show that using prior information does *not* set it apart from the other prevailing approach to parameter estimation, maximum likelihood.

The first step in estimating parameters using maximum likelihood is to choose a likelihood function, a choice that itself requires some prior knowledge of the parameters we seek to estimate. However, it is also possible to use information from previous studies in a more direct way by including this information in the likelihood function.

We showed earlier how multiple, independent observations from a data set can be used to calculate the likelihood (eq. 4.1.5) or the log likelihood (eq. 4.1.6) of a parameter. It is possible to include prior information on the parameter in the same way—the total likelihood is the product of the individual likelihoods of the data and the prior likelihood (Edwards, 1992; Pawitan, 2001), as follows. We define α and β as the parameters describing the probability distribution of θ based on prior research. We can include prior information on θ using

$$\mathrm{L}(\theta|\mathbf{y}) = [\mathbf{y}|\theta]\,[\theta] \tag{4.4.1}$$

$$= \underbrace{\prod_i [y_i|\theta]}_{a}\ \underbrace{[\theta|\alpha, \beta]}_{b}\ .$$

We find the maximum likelihood value of θ, including current data and prior information, by finding the value of θ that maximizes the total likelihood, that is, the product of the probability or probability density of the data conditional θ (term a in eq. 4.4.1) with the probability density of θ conditional on parameters obtained in earlier studies (term b in equation 4.4.1). Edwards (1992, pg. 36) calls the log of term a the experimental support and the log of term b, the prior support. We show later (sec. 9.1.3.1) that equation 4.4.1 is a specific example of a more general statistical procedure called *regularization.*

5 Simple Bayesian Models

In this chapter we lay out the basic principles of Bayesian inference, building on the concepts of probability developed earlier (chapter 3). Our purpose is to use the rules of probability to show how Bayes' theorem works. We will make use of the conditional rule of probability and the law of total probability, so it might be useful to review these first principles (sec. 3.2) before proceeding with this chapter.

We begin with the central, underpinning tenet of the Bayesian view: the world can be divided into quantities that are observed and quantities that are unobserved. Unobserved quantities include parameters in models, latent states predicted by models, missing data, effect sizes, future states, and data before we observe them. We wish to learn about these quantities using observations. The Bayesian framework for achieving that understanding is applied in exactly the same way regardless of the specifics of the research problem at hand or the nature of the unobserved quanties.

The feature of Bayesian analysis that most clearly sets it apart from all other types of statistical analysis is that Bayesians treat all unobserved quantities as random variables.[1] Because the behavior of random variables is governed by probability distributions, it follows that unobserved

[1]There is some argument among statisticians about whether states of ecological systems and parameters governing their behavior are truly random. Ecologists with traditional statistical training may object to viewing states and parameters as random variables. These objections might proceed as follows. Consider the state "the average biomass of trees in a hectare of Amazon rainforest." It could be argued that there is nothing random about it, that at any instant in time there *is* an average biomass that is fixed and knowable at that instant—it is determined, not random. This is true, perhaps, but the practical fact is that if we were to attempt to know that biomass, which is changing by the minute, we would obtain different values depending on when and how we measured it. These values would follow a probability distribution. So, thinking of unknowns

quantities can be characterized by probability distributions like those we learned about in section 3.4. Bayesian analysis uses the rules of probability (sec. 3.2) to discover the characteristics of the probability distributions of unobserved quanties. Understanding those distributions enables the ecological researcher to make statements about processes tempered by honest specifications of uncertainty.

It is fundamental to Bayesian analysis to understand the distinctions among things that are known versus unknown, observed versus unobserved, and random variables versus fixed quantities. The first distinction is this: things that are *known* are not random variables but, rather, are treated as fixed. This might seem obvious, but it can be slippery. Numerical constants, for example π, are known. Things that are not observed, for example, parameters in a model, latent states, predictions, and missing data are unknown and are always modeled as random variables. But what about things we observe?

Observations of responses (i.e., the y) are always modeled as random variables. How can this be? How can something that we observe be random? The key idea here is that the y are random variables *before they are observed*. After we observe them, they are quantities in hand that represent one instance of a stochastic process. So, this one instance of observation is fixed, but if we repeated our observations of the response, we would not expect always to get identical values. The sources of stochasticity in responses will be treated in greater detail as we proceed.

What about observed predictor variables (i.e., covariates, the x)? Are they random or fixed? Rightly or wrongly (usually wrongly), ecologists often treat predictor variables as being observed perfectly—they are observations but they are treated as if they were known, fixed quantities. They are not random variables if we assume they are measured without error, but they *are* random variables if we assume they have measurement or sampling errors that we seek to include in our model.

as random variables is a scientifically useful abstraction with enormous practical benefits, benefits we demonstrate in later chapters. We leave arguments about whether states and parameters are "truly random" to metaphysics. As an aside, Ben Bolker (personal communication) points out that "The same traditionally trained ecologists who object to treating states as random variables don't mind using hypothesis tests that are grounded in the idea of a long-term frequency of observation in repeated observations, which don't sensibly exist in many cases . . . ."

5.1 Bayes' Theorem

The basic problem in ecological research is to understand processes that we cannot observe based on quantities that we can observe. We represent unobserved processes as models made up of parameters and latent states, which we notate here as θ. We make observations y to learn about θ. Before the data are observed, we treat them as random variables. The chance of observing the data conditional on θ is given by a probability distribution, $[y|\theta]$. Because θ is also a random variable, it is governed by the probability distribution $[\theta]$. We wish to discover the probability distribution of the unobserved θ conditional on the observed data, that is, $[\theta|y]$. Using the basic rules of conditional probability for two random variables, we have

$$[\theta|y] = \frac{[\theta, y]}{[y]}, \tag{5.1.1}$$

$$[y|\theta] = \frac{[\theta, y]}{[\theta]}. \tag{5.1.2}$$

Solving equation 5.1.2 for $[\theta, y]$ we have

$$[\theta, y] = [y|\theta]\,[\theta]. \tag{5.1.3}$$

Substituting the right-hand side of equation 5.1.3 for $[\theta, y]$ in equation 5.1.1 we obtain

$$[\theta|y] = \frac{[y|\theta]\,[\theta]}{[y]}. \tag{5.1.4}$$

Because y is conditional on θ, the law of total probability (eqs. 3.2.13 and 3.2.14) for discrete-valued parameters shows that

$$[y] = \sum_{\theta} [y|\theta][\theta], \tag{5.1.5}$$

where the summation is over all possible values of θ. For parameters that are continuous,

$$[y] = \int [y|\theta]\,[\theta]\,d\theta. \tag{5.1.6}$$

Substituting the right-hand side of equation 5.1.5 for $[y]$ in 5.1.4, we obtain Bayes' theorem for discrete-valued parameters,

$$[\theta|y] = \frac{[y|\theta]\,[\theta]}{\sum_\theta [y|\theta]\,[\theta]}, \tag{5.1.7}$$

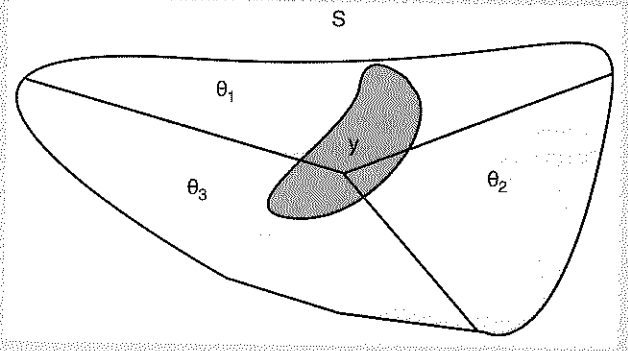
and similarly substituting equation 5.1.6 for $[y]$ in 5.1.4, we find Bayes' theorem for parameters that are continuous,

$$[\theta|y] = \frac{[y|\theta]\,[\theta]}{\int [y|\theta]\,[\theta]\,d\theta}. \tag{5.1.8}$$

Bayes' theorem provides the basis for estimating the probability distribution of the unobserved quantities θ informed by the data y. A simple example illustrates these ideas graphically (box 5.1).

Box 5.1 Illustration of Bayes' Theorem

Imagine support for the parameter θ shown as the light-colored polygon labeled S. Assume that θ can take on three values, θ_1, θ_2, and θ_3. We assume for simplicity that these are the *only* possible values—they are mutually exclusive and exhaustive; that is, $\sum_i$ area of wedge$_i$ $= S$. The area of each θ_i wedge divided by the area of S reflects our prior knowledge of the parameter: area of wedge θ_i/area of S $=$ Pr(θ_i). If we have no reason to favor one value of θ_i over another, $\Pr(\theta_1) = \Pr(\theta_2) = \Pr(\theta_3) = \frac{1}{3}$



(continued)

(Box 5.1 *continued*)

We now collect some data, shown by the dark polygon y. The parameter θ controls how the data arise. So, for example, the data might be the number of survivors observed in a sample of n individuals during time Δt, where θ is the probability that an individual survives over the time interval. We want to use the data to update our knowledge of θ.

Given that we have data in hand, we can limit attention to the wedge of the θ_i contained within the data polygon. The probability of θ_i is $\Pr(\theta_i|y) = \frac{\text{area of } \theta_i \text{ within } y}{\text{area of } y} = \frac{\text{area of } \theta_i \text{ within } y/\text{area of } S}{\text{area of } y/\text{area of } S} = \frac{\Pr(\theta_i \cap y)}{\Pr(y)} = \frac{\Pr(\theta_i, y)}{\Pr(y)}$. Using the conditional rule of probability to substitute for $\Pr(\theta_i, y)$, we have $\Pr(\theta_i|y) = \Pr(y|\theta_i)\Pr(\theta_i)/\Pr(y)$. Using $\Pr(y) = \text{area of } y/\text{area of } S = \sum_j \Pr(y|\theta_j)\Pr(\theta_j)$, we find Bayes' theorem for discrete parameters:

$$\Pr(\theta_i|y) = \frac{\Pr(y|\theta_i)\Pr(\theta_i)}{\sum_j \Pr(y|\theta_j)\Pr(\theta_j)}. \tag{5.1.9}$$

The denominator is a normalizing constant assuring that $\sum_i \Pr(\theta_i|y) = 1$. As the number of wedges in S increases to infinity and their area decreases to 0, we have Bayes' theorem for continuous parameters:

$$[\theta|y] = \frac{[y|\theta][\theta]}{\int [y|\theta][\theta]d\theta}. \tag{5.1.10}$$

Understanding Bayesian inference and why it works requires that we understand each of its components, which we now explain for continuous parameters. The *likelihood* $[y|\theta]$ (fig. 5.1.1) plays a key role in Bayesian analysis by linking the unobserved θ to the observed y. It allows us to answer a central question of science: what is the probability that we will observe the data if our deterministic model ($g(\theta)$) accurately portrays the process that gives rise to the data? We have seen the likelihood before (eq. 4.1.4, fig. 4.1).

The *prior distribution* of the unobserved quantities, $[\theta]$, represents our knowledge about θ before we collect the data (fig. 5.1.1). The prior distribution can be informative, reflecting knowledge gained in previous research, or it can be vague, reflecting a lack of information about θ before we collected the data that are now in hand. We treat priors in greater detail in the next section; for now, we highlight prior distributions as one of the components of Bayes' theorem.

$$[\theta|y] = \frac{[y|\theta]\,[\theta]}{\int [y|\theta]\,[\theta]\,d\theta}$$

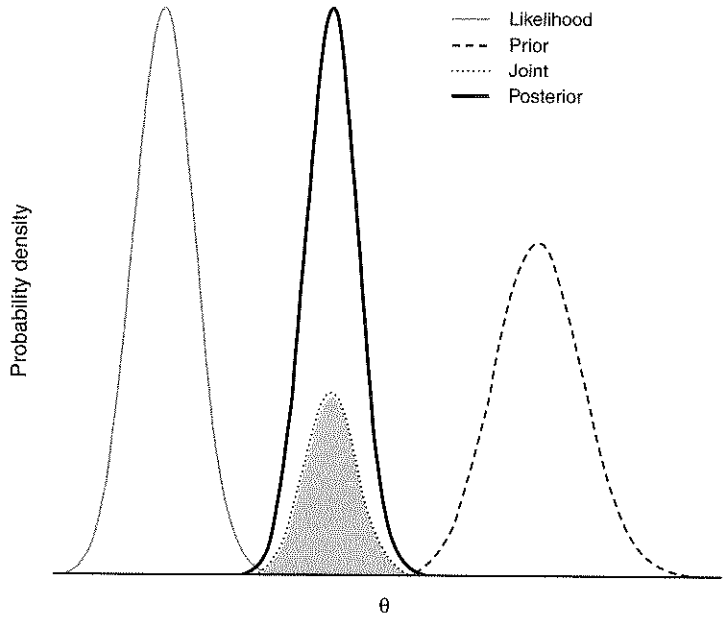


Figure 5.1.1. Illustration of Bayes' theorem for data (y) and unobserved quantities (θ). The likelihood ($[y|\theta]$, gray solid line) gives the probability that we will observe the data conditional on the value of the parameter. The prior ($[\theta]$, dashed line) specifies the probability of θ based on our knowledge of θ before the data are collected. The joint distribution (dotted line) is the product of the prior and the likelihood. The marginal distribution of the data $\left(\int [y|\theta][\theta]d\theta\right)$ is the integral of the joint distribution, shown here as the shaded area. (See sect. 3.4.2 for a review of the concept of marginal distributions.) The posterior is the distribution (black solid line) that results when we divide every point on the joint distribution curve by the area under the curve, effectively normalizing the joint distribution so that the area under the posterior distribution equals 1.

The product of the likelihood and the prior is the joint distribution[2] (fig. 5.1.1). We have seen this product $\left([y|\theta]\,[\theta]\right)$ before (eq. 4.4.1), and we learned that it does not define a probability distribution for θ because the area under the curve $[y|\theta]\,[\theta]$ with respect to θ is not certain to equal 1.

[2]Recall that the joint distribution $[\theta, y] = [y|\theta][\theta]$.

The marginal distribution of the data,

$$[y] = \int [y|\theta]\,[\theta]\,d\theta, \tag{5.1.11}$$

is the area under the joint distribution curve (fig. 5.1.1). Dividing each point on the joint distribution $[y|\theta]\,[\theta]$ by $\int [y|\theta]\,[\theta]\,d\theta$ normalizes the curve with respect to θ, yielding the posterior distribution $[\theta|y]$. The posterior distribution is a true probability density function that meets all the requirements for these functions (sec. 3.4.1), including that $\int [\theta|y]d\theta = 1$. Dividing the joint distribution by $\int [y|\theta]\,[\theta]\,d\theta$ assures that the posterior distribution integrates to 1, which is why $[y]$ is often referred to as a *normalizing constant*.

Before the data are collected, y is a random variable, and the quantity $\int [y|\theta]\,[\theta]\,d\theta$ is a marginal distribution, a concept we will use frequently in later chapters (for review, see sect. 3.4.2). It is also called the *prior predictive distribution*—it tells us what we know about the *data* before they are collected. However, after the data are collected, $\int [y|\theta]\,[\theta]\,d\theta$ is a known, fixed quantity (a scalar). This means that

$$[\theta|y] \propto [\theta, y] \tag{5.1.12}$$

$$\propto [y|\theta]\,[\theta]\,. \tag{5.1.13}$$

We will make extensive use of this proportionality.[3] We can use equation 5.1.13 to learn about the posterior distribution from the joint distribution even when we cannot directly calculate $[y]$, as will often be the case. We call equation 5.1.13 a simple Bayesian model because it represents the joint distribution of the observed and unobserved quantities as the product of the likelihood and the prior distributions.

We could have developed the same ideas about discrete-valued parameters using sums rather than integrals.

5.2 The Relationship between Likelihood and Bayes'

The fundamental difference between inference based on maximum likelihood and inference based on Bayes' theorem is that Bayes' treats all

[3]The constant of proportionality is the reciprocal of the marginal distribution of the data, which is a constant after the data are observed.

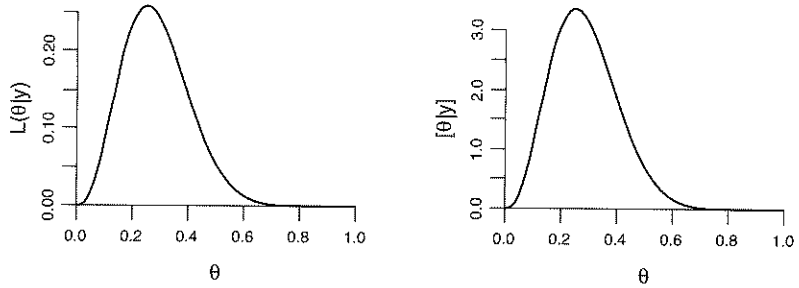


Figure 5.2.1. Likelihood profile (left panel) and posterior distribution (right panel) for the parameter probability of a success, θ, given the observation three successes on 12 trials with uninformative priors on θ. The shapes of the two curves are identical. The area under the likelihood profile does not equal 1. The area under the posterior distribution equals 1.

unobserved quantities as random variables governed by probability distributions. This treatment is possible because dividing the joint distribution by the marginal distribution of the data assures that the posterior distribution is a true probability distribution (fig. 5.2.1). This is a nontrivial result, because it allows Bayesian inference to make probabilistic statements about unobserved quantities of interest. In contrast, the likelihood profile is not a probability distribution—there is nothing that assures that the area under the curve equals 1 (fig. 5.2.1). Unknowns cannot be treated as random variables in the likelihood framework. Instead, likelihood depends on comparing the relative strength of evidence in data for one value of a parameter over another value. Prior information can be used in Bayesian and likelihood analysis with $[y|\theta][\theta]$. In likelihood, we find the values of θ that maximize $[y|\theta]$ $[\theta]$. The normalization of this product by the marginal distribution of the data is what sets Bayesian inference apart from inference based on likelihood—it allows unobserved quantities to be treated as random variables.

5.3 Finding the Posterior Distribution in Closed Form

A simple Bayesian model contains a joint distribution expressed as a likelihood multiplied by a prior (or priors), $[y|\theta][\theta]$. There are special cases of this product where the posterior distribution $[\theta|y]$ has the same form as the prior, $[\theta]$. In these cases, the prior and the posterior are called *conjugate distributions* (or simply conjugates), and the prior is called a

conjugate of the likelihood. Conjugate distributions are important for two reasons. For simple problems, they allow us to calculate the parameters of posterior distributions on the back of a cocktail napkin.[4] Moreover, the ease of calculation of parameters of the posterior for simple problems becomes important for complicated problems if we can break them down into parts that can be attacked one at time. We will learn about the role of conjugates in this process in the chapter on Markov chain Monte Carlo (chapter 7).

It is perfectly possible to make use of conjugate priors effectively without knowing how each one is derived. Seeing a single derivation (box 5.3) is adequate background for most ecologists who seek to use Bayesian methods. However, we offer a couple of examples here to provide intuition for conjugate relationships. More detailed treatment as well as tables showing the known conjugate distributions can be found in Bayesian textbooks (e.g., Gelman, 2006). The ones we use most frequently are shown in appendix table A.3.

Box 5.3 Derivation of the Posterior Distribution for a Beta Prior and Binomial Likelihood

We seek the posterior distribution of the parameter ϕ, the probability of a success conditional on n trials and y observed successes. The beta distribution is a conjugate prior for the binomial likelihood. We use Bayes' theorem to obtain

$$[\phi|y,n] \propto \underbrace{\binom{n}{y}\phi^{y}(1-\phi)^{n-y}}_{\text{binomial likelihood}} \underbrace{\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\phi^{\alpha-1}(1-\phi)^{\beta-1}}_{\text{beta prior}}, \tag{5.3.1}$$

where α and β are the parameters of the beta prior distribution. By dropping the normalizing constants $\left(\binom{n}{y}, \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right)$ we obtain

$$[\phi|y,n] \propto \underbrace{\phi^{y}(1-\phi)^{n-y}}_{\text{binomial likelihood}} \underbrace{\phi^{\alpha-1}(1-\phi)^{\beta-1}}_{\text{beta prior}}. \tag{5.3.2}$$

Simplifying, we get

$$[\phi|y,n] \propto \phi^{y+\alpha-1}(1-\phi)^{\beta+n-y-1}. \tag{5.3.3}$$

(continued)

[4] It is embarrassing to do an elaborate numerical procedure to obtain results that can be obtained on a napkin.

5.4 More about Prior Distributions

We devote an entire section in this chapter to prior distributions because ecologists who have not received formal training in Bayesian methods will be especially unfamiliar with the use of priors, a concept that, in contrast with likelihood, has no parallel in traditional statistical training. We also include this section because ecologists often seek to minimize the influence of the prior on inference. This is a place where it is easy to make errors. Finally, we want to advocate the thoughtful use of informed priors in Bayesian modeling.

Some view the choice of a prior in Bayesian models as a contentious topic because it is a decision that can influence inference. However we will attempt to convince you of the following:

1. There is no such thing as a noninformative prior, but certain priors influence the posterior distribution more than do others.
2. Informative priors, when properly justified, can be tremendously useful in Bayesian modeling (and science, in general).

It is important to remember that one of the objectives of Bayesian analysis is to provide information that can inform subsequent analyses; the posterior distribution obtained in one investigation becomes the prior in subsequent investigation. Thus, we agree with the view of Gelman (2006) that "non-informative" priors are provisional. They are a starting point for analysis. As scientists, we should always prefer to use appropriate, well-constructed, informative priors on θ.

5.4.1 "Noninformative" Priors

We use quotation marks in this section title because there is no such thing as a noninformative prior. By that we mean that all priors will have some influence on the posterior distribution of some transformation of the parameter you may be interested in learning about. Let's begin, then, by studying potential priors for a very simple Bayesian model, a model for binary data. Consider the set of binary data (i.e., zeros and ones) denoted by y_i, for $i = 1, \ldots, n$. If we are interested in inference concerning the probability that a given observation will be one, $p = \Pr(y = 1)$, then we could formulate the parametric model

$$y_i \sim \text{Bernoulli}(p), \tag{5.4.1}$$

where $i = 1, \ldots, n$. In this case, a Bernoulli distribution is the "model" that we assume stochastically generated the data. The Bernoulli distribution contains the parameter p; thus, a complete Bayesian model requires a prior distribution for p. Let's examine a few priors for p as well as their influence on the posterior distribution for the following data set: $\mathbf{y} = (0, 0, 1, 0, 1, 0, 0, 0, 1, 0)'$.

Perhaps the most commonly chosen prior for p is the uniform distribution, such that $0 < p < 1$. The uniform is a specific case of the more flexible beta distribution; thus, it is common to select the prior

$$p \sim \text{beta}(\alpha, \beta), \tag{5.4.2}$$

where if $\alpha = \beta = 1$, this distribution becomes a uniform. The uniform distribution is commonly thought to be "noninformative" in this setting because all possible values of p are equiprobable. The uniform can be contrasted with a prior where larger values of p are more probable, such as when $\alpha = 4, \beta = 1$. We compare the posterior distributions arising from these two choices for a prior in figure 5.4.1. Notice how the prior in figure 5.4.1 B "pulls" the posterior toward the larger values, thus influencing it.

An alternative to the visual approach for assessing the influence of the prior on the posterior is to inspect the closed-form mathematical expression for the posterior (i.e., the result of conjugate relationships, sect. 5.3). For the Bernoulli-beta model[5] we are using in this example, the posterior distribution for p is

$$[p|\mathbf{y}] = \text{beta}\left(\sum_{i=1}^{n} y_i + \alpha, \sum_{i=1}^{n}(1 - y_i) + \beta\right). \tag{5.4.3}$$

In our simple example, for the data $\mathbf{y}$, the resulting beta posterior distribution has parameters $3 + \alpha$ and $7 + \beta$. Notice that larger values for the prior parameters α and β have more of an effect on these parameters in the posterior. Similarly, as both α and β get small, the posterior distribution appears to become less influenced by the prior (leaving only statistics related to the data in the posterior). Thus, a beta prior with $\alpha = 1, \beta = 1$ is less influential on the posterior than a beta prior with $\alpha = 4, \beta = 1$. This is a seemingly sensible result and one that is very commonly used to justify the specification of priors, especially for probabilities (i.e., p), regression

[5] We showed the derivation of the expression for the posterior distribution when the prior is beta and the likelihood is binomial in section 5.3. Recall that the Bernoulli is a special case of the binomial where n, the number of trials equals 1.

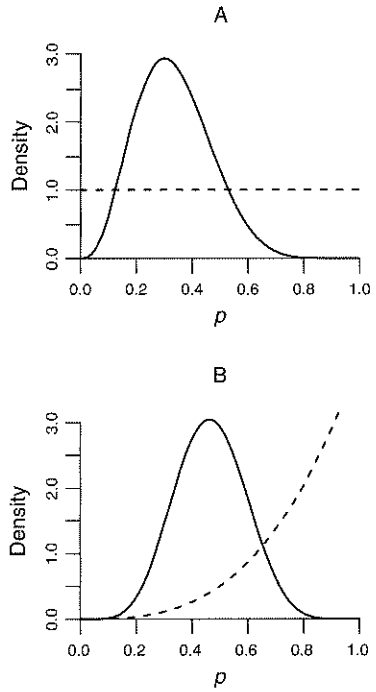


Figure 5.4.1. Prior (dashed line) and resulting posterior distributions (solid line) for a model with a Bernoulli likelihood and a beta prior with two prior specifications: (**A**) $\alpha = 1, \beta = 1$ and (**B**) $\alpha = 4, \beta = 1$.

coefficients (i.e., $\boldsymbol{\beta}$), and variance components (i.e., σ^2). Perfect flatness can be achieved only in bounded priors like the beta, but priors that *approach* flatness are often referred to as "flat" nonetheless. You will also see them called "diffuse," "weak," or "vague."

It is important to recognize that even the uniform prior for p technically has some influence on the posterior distribution, because prior parameters $\alpha = 1, \beta = 1$ yield the posterior parameters $3 + 1, 7 + 1$, which are not the same as $3, 7$, as would be the case if only statistics related to the data appeared in the posterior. Using this argument, one might be tempted to use $\alpha = 0, \beta = 0$ as prior parameters, but recall from the definition of the beta distribution that both parameters must be greater than zero to ensure a valid probability distribution. Furthermore, the sensibility of using very small values for α and β in the beta prior breaks down because, as we see in figure 5.4.2, a beta prior with $\alpha = 0.001, \beta = 0.001$ actually pulls the posterior distribution toward zero. Most of the mass of a U-shaped prior

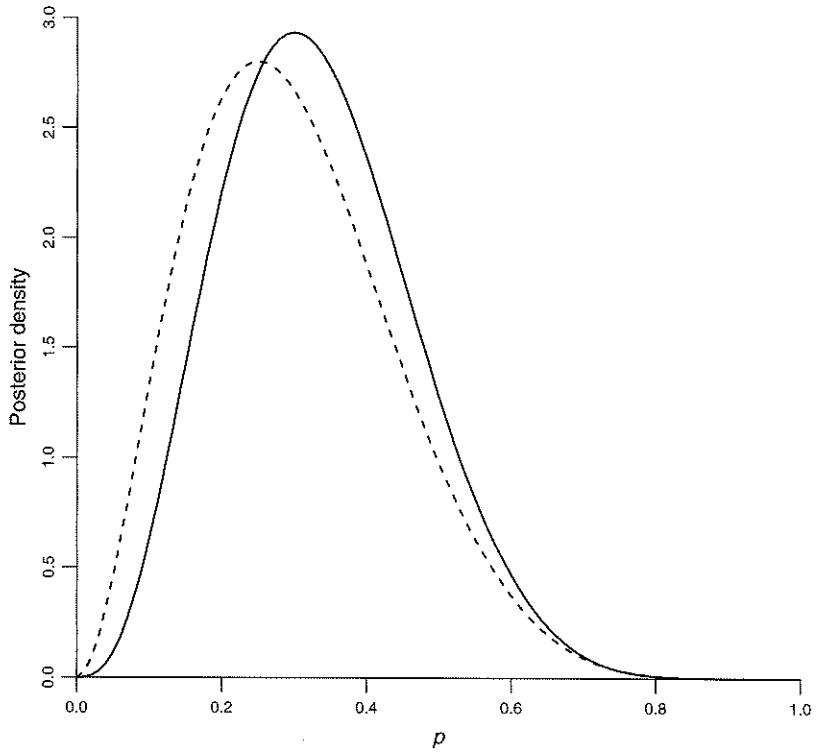


Figure 5.4.2. Resulting posterior distributions for the Bernoulli-beta model with prior specifications $\alpha = 1, \beta = 1$ (solid line) and $\alpha = 0.001, \beta = 0.001$ (dashed line).

distribution implied by the beta$(0.001, 0.001)$ is near 0 and 1, suggesting that p is more likely to be large or small but not moderate (i.e., close to 0.5).

The take-home message is that all priors have an influence on the posterior distribution, and what might seem like a good trick to minimize the prior influence may not always do what you think it should. You can always overwhelm any amount of prior influence with enough data. In our example, if n gets large, then any prior values for α and β become inconsequential in the posterior; they will be very minimal compared with the large values for $\sum_{i=1}^{n} y_i$ and $\sum_{i=1}^{n}(1 - y_i)$. Thus, to some extent, the simplest way to minimize prior influence is to collect a larger data set!

Another caution in specifying priors that appear to minimize the influence on the posterior distribution pertains to "propriety." A proper probability distribution is a positive function that integrates to 1 over the support of its

random variable (sec. 3.4). If the function does not integrate to 1, then it is termed "improper" and is not technically a valid probability distribution. That means we can't use it for statistical inference, because all statistical theory depends on the basic axioms about probability distributions. For example, continuing the previous discussion about how to make the beta distribution less influential, we would be tempted to use $\alpha = 0, \beta = 0$. However, because both parameters must be positive to guarantee a proper prior distribution, the beta(0, 0) is not a valid probability density function and thus its use is not advised. Interestingly, the resulting posterior, which we can still work out analytically, ends up being a beta(3, 7), which is proper in this specific case. Therefore, an improper prior can *sometimes* lead to a proper posterior, but that result has to be shown for the particular model being fit and almost always depends on the data. If you cannot mathematically show that your posterior is proper, then it's best to avoid improper priors.

Let's consider another situation. Suppose you have the same data and Bayesian model but are interested in obtaining inference related to the quantity p^2, rather than p. The seemingly benign uniform prior (i.e., beta(1, 1)) for p then becomes quite informative for p^2. To illustrate this point, we can find the implied prior distribution for p^2 using a Jacobian transformation technique.[6] In this case, if we use a uniform prior for p, the implied prior for p^2 (the quantity about which we desire inference) is proportional to $1/p$. Therefore, the values of p^2 under its implied prior are not equiprobable, as they are for p. Specifically, the uniform prior for p says that smaller values for p^2 are more probable than larger values. That result may not be what we had in mind when we chose the beta(1, 1) prior for p. A prior whose information about a parameter does not change when we transform the parameter is called "invariant to transformation." The *Jeffreys prior* was developed for this exact purpose, to help specify priors that are invariant to transformation.

The Jeffreys prior depends on the form of the likelihood (also called the *data model*). More specifically, the Jeffreys prior is proportional to Fisher information raised to the half power.[7] That is, if we can calculate the negative expectation of the second derivative of the log likelihood,

$$-E_y\left(\frac{d^2 \log[\mathbf{y}|p]}{dp^2}\right), \tag{5.4.4}$$

[6]The details of this technique are beyond the scope of this book but can be found in any graduate-level mathematical statistics book.

[7]This is the same Fisher information used to find asymptotic variance of an MLE.

then we have something proportional to the Jeffreys prior. The Jeffreys prior for our ongoing binary data example (eq. 5.4.1) is, perhaps surprisingly, a beta(0.5, 0.5) distribution. This Jeffreys prior will contain the same information for p as it will for p^2, or any other transformation of p for that matter. Unfortunately, the Jeffreys prior is often called "noninformative," but for the same reasons cited earlier, it is not noninformative. We might use a Jeffreys prior when we don't know what else to use, in this case, because it happens to be invariant to transformation. For our example, the Jeffreys prior is U-shaped; not quite as extremely U-shaped as the beta(0.001, 0.001) prior for p, but it will still give more prior preference to those values close to 0 and 1 than to 1/2. The Jeffreys prior for this particular example turns out to be proper, but it is not guaranteed to be proper for all models.

You will commonly see a normal prior with large variance used as a prior distribution for a variety of parameters. A normal distribution with large variance (i.e., normal (0, 1000)) is often justified as an attempt to find a vague prior that is conjugate.[8] Given that the normal distribution is not bounded, it will be impossible to make it perfectly flat, so the large variance serves as a mechanism to at least spread it out.[9] A normal with infinite variance would be flat, but then it would also not be proper (i.e., would not integrate to 1). The use of a normal prior with large, but finite, variance seems to work well without complications for parameters that are means and where the data contain plenty of information. However, for other types of parameters, say transformations of probabilities such as logit(p), the normal prior with large variance can have a dubious influence on the posterior.

To illustrate our point, suppose we have the same Bernoulli model for the binary data we've been discussing in this section, and we use the prior for logit(p) such that

$$\text{logit}(p) \sim \text{normal}(0, \sigma_p^2)\,, \tag{5.4.5}$$

where, σ_p^2 is set to be a large number. The question is, what prior does this imply for p (rather than logit(p))? Simulating 10, 000 random draws from a normal distribution and taking the inverse logit transformation, we can see in figure 5.4.3 that a normal with $\sigma_p^2 = 100$ is much more informative than a normal with $\sigma_p^2 = 2$.

[8]Recall that conjugacy occurs when a prior and posterior have the same form. There can be many analytical and computational advantages to using conjugate priors, but they are not always the best choice.

[9]Keep in mind that the prior variance is always relative to the scale of the parameter. For example, if the data indicate that a parameter should be 1000, then a N(0, 100) prior for that parameter will probably be informative unless the sample size is huge, because a variance of 100 is small relative to 1000, as is the prior mean of 0.

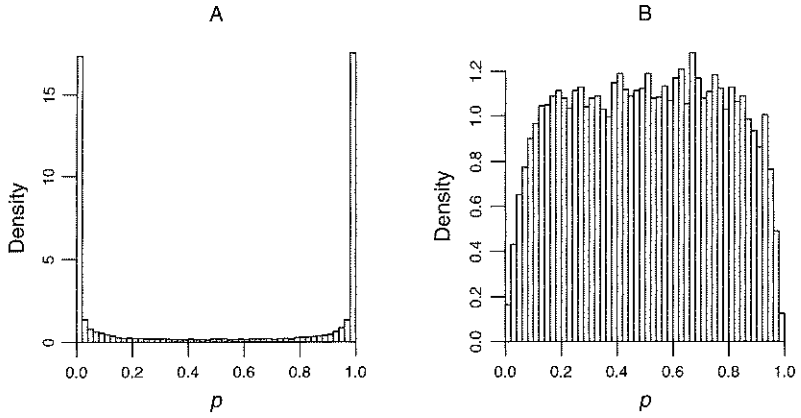


Figure 5.4.3. Histograms of p based on samples drawn from prior distributions for (**A**) logit(p) $\sim$ normal(0, 100) and (**B**) logit(p) $\sim$ normal(0, 2).

Priors with large variance might seem vague or less informative, but they are not always, thus it is a good idea to check the implied prior distribution in the transformation of the parameter for which you desire inference. You can do this by varying the values of the parameters for the prior and examining how that variation effects the posterior.

It's worth mentioning that the same methods are commonly used for choosing priors for variance components. In fact, we present models that contain such priors throughout this book. It is important to realize that such priors are not truly noninformative, for the same reasons we described earlier. For example, suppose we have data that can be sufficiently modeled with a normal distribution,

$$y_i \sim \text{normal}(\mu, \sigma^2), \tag{5.4.6}$$

for $i = 1, \ldots, n$, and where the mean μ is assumed to be known (for now). Our interest lies in obtaining inference about the variance, σ^2. A conjugate prior for the variance parameter is the inverse gamma distribution,

$$\sigma^2 \sim \text{inverse gamma}(\alpha, \beta), \tag{5.4.7}$$

which yields the posterior for σ^2:

$$\left[\sigma^2|\mathbf{y}\right] = \text{inverse gamma}\left(\frac{n}{2} + \alpha, \frac{\sum_{i=1}^{n}(y_i - \mu)^2}{2} + \beta\right). \tag{5.4.8}$$

Notice that, much like the beta posterior discussed previously, here in the inverse gamma posterior for σ^2, if α and β get small, then the influence of

the prior on the posterior is minimized. Thus, it is common to see priors for variance components specified as inverse gamma (0.001, 0.001) in an attempt to minimize prior influence (but see Gelman, 2006). However, these priors are not "noninformative" and are not invariant to transformation. This sort of prior could be misleading for example, if, one was interested in obtaining inference about the standard deviation, σ, rather than the variance, σ^2.[10] As an alternative, the Jeffreys prior could be used for σ^2. For this model, the Jeffreys prior turns out to be proportional to $1/\sigma^2$, which has the form of an inverse gamma with $\alpha = 0$ and $\beta = 0$. This formulation for the inverse gamma does not yield a proper prior because both parameters (α and β) must be positive. However, the Jeffreys prior, as always, is invariant to transformation. As in the case with the Bernoulli model previously discussed, the Jeffreys prior for σ^2 yields a proper posterior as long as at least one observation is available (i.e., $n \geq 1$).

Finally, there is another approach to finding priors whose influence is minimal on the posterior; these priors are called *reference priors*. A reference prior is found by maximizing the Kullback-Leibler (K-L) divergence between the posterior and prior distributions.[11] The heuristic concept behind reference priors is that a prior which is as different as possible from the posterior may be desirable if you have no prior information or expertise and need a default prior to use just to obtain Bayesian inference.[12] Interestingly, for univariate parameters, the reference prior approach yields the Jeffreys prior! However, in multivariate situations, the reference prior must be found for each individual model where it is being used. However, calculating the correct reference prior can be quite challenging analytically and numerically. The field of objective Bayesian inference focuses on this task for various models.

5.4.2 Informative Priors

We have learned that all priors influence the posterior in some way but that we can often assess the amount of influence and sometimes even control it. But when formulating statistical models, we might ask

[10]This is more common than you might think, as it is easier to interpret the standard deviation, σ, than the variance, σ^2.

[11]The development of this concept is beyond the scope of this book, but in short, the K-L divergence provides a way to measure discrepancy between two distributions; it involves explicit integration and can be difficult to compute in practice, making this approach quite technical.

[12]Some argue that this very concept seems contrary to the Bayesian spirit by trying to avoid its biggest utility, the ability to properly account for previous research efforts in making scientific conclusions.

ourselves why we're trying to limit the influence of the prior on the posterior in the first place. The illusion of objectivity has been put on a pedestal in science, almost to the extent that we are to believe that only new data can be used to reach scientific conclusions. Extrapolating this concept to Bayesian statistics would then imply that we *should* be looking for priors that have no influence on the posterior (hence the previously mentioned subfield of objective Bayesian inference). However, a point not often recognized is that all parametric statistical modeling approaches are subjective, including maximum likelihood. The very fact that we have to choose a likelihood function implies that we have made a strong assumption about the data-generating mechanism. Nonparametric statistical approaches seek to minimize such assumptions, but they make their own set of strong assumptions based on their associated computational algorithms for providing inference. Any constraint we put on data or parameters to obtain inference imparts subjectivity. As we discuss in chapter 9, the various forms of regularization, including penalization methods and model selection, put extreme constraints on parameters, yet they are used throughout statistics and across all applied fields without much fanfare concerning their inherent subjectivity. More important, these approaches are recognized as being helpful in many ways!

Our view is that we would be remiss if we were to ignore decades of important scientific learning in the field of ecology and that there should be a way to rigorously incorporate this learning into our statistical approaches. Fortunately, the Bayesian framework provides such a mechanism. The posterior distribution itself is a formal, mathematically valid way to combine information from current as well as previous scientific studies. In that light, it is not difficult to see that the posterior distribution and Bayesian framework are a mathematical representation of the scientific method itself.

In the scientific method, we use existing data and expertise to formulate hypotheses about how the world works, then we make conclusions and update hypotheses using new data. In Bayesian statistics, we summarize our understanding of how the world works in a prior distribution and then "update" (i.e., compute the posterior distribution) our understanding using new data. Science would be completely haphazard if we threw out everything we knew about the world every time we began a new study. Haphazard is not even a strong enough word to describe science performed in a manner where we pretend to be completely ignorant about our study system; perhaps, lazy or irresponsible would be a better descriptor! In all seriousness, we challenge readers (and ourselves) to provide an example of a parameter in a statistical model they wish to fit knowing absolutely nothing about it—nothing at all. At a minimum, we should all know

at least the support (i.e., the values the parameter can assume) for any parameter, but we often know quite a bit more than that. Ignoring prior information you have is like selectively throwing away data before an analysis.

Instead, we argue that science would be better off if we all took the time to carefully collect and represent our prior understanding of parameters in Bayesian models. Doing so can be hard work, as it sometimes requires a mathematical transformation of moments into natural parameters in the distribution we, as experts, value as best representing the data and parameters. It also could include being more responsible in our knowledge of preexisting scientific findings, for example, by more carefully reading the literature and translating those findings into quantitative information we can use in our prior. Formulating honest and responsible priors may also involve communicating with other experts on the topic under study, probing them for details that can be represented in probability distributions to serve as priors. Yes, this is beyond what we normally do in statistical analyses, but Bayesian methods provide the tools for incorporating such information, and we should be obligated to use them responsibly.

In addition to being helpful in accounting for the body of accumulated scientific knowledge when we make new inferences, strong so-called informative priors can be beneficial in the following ways:

- They allow us to benefit from several sources of information, including different data sources and expert knowledge. Given that priors are most influential when paired with small data sets, it can be incredibly helpful to obtain meaningful inference by having a formal mechanism for combining several smaller but independently collected data sets into a single modeling framework. An additional likelihood involving a separate data source can often be written as a prior in the original model containing the primary data source. We cover use of multiple likelihoods in the joint distribution in section 6.2.5.
- Informative priors stabilize computational algorithms. This benefit is not an inferential one but definitely a practical one. When statistical models accumulate parameters in such a way that the ratio of unknowns to knowns grows, the probability surfaces we need to explore during the fitting process can acquire pathological problems such as lack of identifiable[13] parameters, multicollinearity, and flat likelihood or posterior surfaces. Such issues can cause numerical approaches to become unstable (e.g., fail to converge). Stronger priors add definition

[13]Parameters are identifiable if they can be estimated given a large amount of data. They are unidentifiable if they cannot. See section 6.3 for a more complete definition of identifiability.

to the surfaces being explored by the statistical fitting algorithms and thus improve computational stability.

- Stronger priors offer a formal way to place constraints on the unknowns in statistical models. A seldom recognized fact is that such constraints are the basis for important inferential tools such as model selection. We cover this topic in great detail in chapter 9, but as a preview we note here that important concepts such as information criteria, penalized likelihood methods, ridge regression, Lasso, and cross-validation are used regularly in many fields and can all be considered as different ways to improve inference through the use of stronger priors. Most statisticians now recognize that imposing a constraint on an optimization problem (e.g., maximizing a likelihood) is the same concept as specifying a prior in a Bayesian model, and both can be helpful for the same reasons.

Excellent examples of the benefits of using informative priors can be found in Crome et al. (1996); McCarthy and Masters (2005); Elderd et al. (2006) and McCarthy et al. (2008).

Until now, we have considered informative priors as single distributions as if they were obtained from a single, previously conducted investigation. What do we do if we have multiple sources of prior knowledge informing a parameter θ? Recall the idea of a mixture distribution (sec. 3.4.5). We can compose a prior from multiple previous studies by mixing their estimates of θ. A prior on θ using information from L different studies can be written as

$$[\theta] = \sum_{l=1}^{L} w_l \, [\theta]_l \,, \tag{5.4.9}$$

$$\sum_{l=1}^{L} w_l = 1, \tag{5.4.10}$$

where the w_l are weights, and $w_l \geq 0$. If we believe that all studies were conducted equally well, then we would choose the w_l to be equal. As an example, assume we had three studies of the intercept β_0 in a regression with an associated variance that we wished to combine in a prior. We might reasonably use

$$[\beta_0] = \tfrac{1}{3}\text{normal}\left(\beta_{0,1}, \sigma_1^2\right) + \tfrac{1}{3}\text{normal}\left(\beta_{0,2}, \sigma_2^2\right) + \tfrac{1}{3}\text{normal}\left(\beta_{0,3}, \sigma_3^2\right). \tag{5.4.11}$$

Now that we can see the potential value of priors informed by single or multiple studies, we need to know how to represent existing scientific knowledge in the form of a probability distribution. There are different approaches for manifesting expert knowledge about a parameter into a prior distribution, but rather than cover each one generically, we present the following example.[14]

5.4.3 Example: Priors for Moth Predation

A particular species of nocturnal moth rests during the day on tree trunks, and its coloration acts as camouflage to protect it against predatory birds. A study was conducted to evaluate predation of a common moth species. Suppose that n sites (for $i = 1, \ldots, n$) were selected, and a varying number of dead moths, N_i, were glued to tree trunks at each site. After 24 hours, the number of moths that had been removed, y_i, presumably by predators, was recorded. A reasonable data model for the moth counts would be a binomial with N_i "trials" per site such that

$$y_i \sim \text{binomial}(N_i, p), \tag{5.4.12}$$

where the parameter p corresponds to the probability of predation and is the unknown about which we desire inference. Consider the following three scenarios in formulating an appropriate prior distribution for this model:

1. We desire a relatively vague prior that contributes information equivalent to two additional "placed" moths and an expected prior probability of predation of .5.
2. We desire an informative prior based on a previous observational study that reported an average of 10% (standard deviation of 2.5%) of the moths in a population were eaten by predators in a 24-hour period.
3. We desire an informative prior based on a pilot study that suggests the proportion removed in any given 24-hour period is unlikely to exceed 0.5 or be less than 0.1.

In scenario 1, the fact that we do not feel we have much prior information pertaining to p means that we want to spread out the probability mass in the prior between 0 and 1 such that our prior has a mean of 0.5 but no strong preference for any range of values. A beta distribution could work well here such that $p \sim \text{beta}(\alpha, \beta)$. But how do we assess the information content of

[14]This example is gratefully modified from the excellent text of Ramsey and Schafer (2012) using ideas from Kiona Ogle.

the prior in terms of an effective increase in sample size? The answer comes from looking at the form of the posterior distribution for this model:

$$[p|\mathbf{y}] = \text{beta}\left(\sum_{i=1}^{n} y_i + \alpha, \sum_{i=1}^{n}(N_i - y_i) + \beta\right). \tag{5.4.13}$$

In equation 5.4.13 we can see a form for the posterior very similar to the one in the Bernoulli model discussed previously, where each of the updated posterior parameters contains a sum of two components, one from the data and one from the prior. The first parameter, $\sum_{i=1}^{n} y_i + \alpha$, is the sum of y_i over all sites—that is, the total number of moths placed that were preyed on—plus the prior parameter α. The second posterior parameter, $\sum_{i=1}^{n}(N_i - y_i) + \beta$, is the total number of moths not removed by predation, plus prior parameter β. Thus, if we set $\alpha = 1$ and $\beta = 1$, it is equivalent to adding two moths to the sample size in such a way that the information does not impose any preference for predation. In this case, the implied prior is a beta(1, 1) or a uniform distribution. Of course, we could have started with a uniform, but it is instructive to see that the prior parameters α and β can be thought of as augmenting the sample size if that helps specify prior information. Given the preceding, what prior would be induced if we had the equivalent of 10 extra moths, worth of prior information such that 60% was in favor of predation and the other 40% was against predation? The answer could easily be visualized by plotting a beta probability density function with parameters $\alpha = 4$ and $\beta = 6$.

In scenario 2, we have information from a former study on moth predation. That study provides inference pertaining to the mean and standard deviation of the proportion of moths preyed on. In translating this information into our beta prior, we can consider the mean and variance equations associated with a beta random variable:

$$\text{E}(p) = \frac{\alpha}{\alpha + \beta}, \tag{5.4.14}$$

$$\text{Var}(p) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}. \tag{5.4.15}$$

Setting $\text{E}(p) = 0.1$ and $\sqrt{\text{Var}(p)} = 0.025$, we can backsolve for α and β to find the appropriate prior (as discussed in sect. 3.4.4). We then arrive at $\alpha = 14.3$ and $\beta = 128.7$ as parameters in our prior.

Scenario 3 is slightly more involved, but entirely realistic, in that it is common for prior information to arise as bounds on likely values for a parameter. In this scenario, if we assume that the term "unlikely" implies that p should fall between a lower bound and an upper bound with high probability (e.g., 95%), then we need to take an approach similar to the moment matching technique, but instead of relating moments to the results of a pilot study, we relate quantiles of the distribution to the results of a pilot study. That is, we must solve the system of equations

$$\int_0^{0.1} \text{beta}(p|\alpha, \beta)dp = .025 \tag{5.4.16}$$

$$\int_{0.5}^{1} \text{beta}(p|\alpha, \beta)dp = .025 \tag{5.4.17}$$

for α and β, where equations 5.4.16 and 5.4.17 integrate the beta probablity density function. This calculation would be quite difficult to perform analytically (i.e., with pencil and paper) but could be approximated numerically using an optimization algorithm in a mathematical or statistical software package. We used the function `optim()` in R (R Core Team, 2013) to find the appropriate prior parameter values $\alpha = 4.8$ and $\beta = 12.7$ by minimizing the difference between the output of the beta cumulative distribution function (i.e., `pbeta()` in R) and .025.

Thus, there are a variety of ways to convert preexisting scientific information and expertise into probability distributions for use as priors in Bayesian models. These informative priors can be very useful in many ways, but only when care is taken to appropriately specify them. It is a common concern that if Bayesian models fell into the wrong hands, they could be misused by those seeking to mislead science or policy. However, even under such dubious circumstances, the priors would have to be clearly spelled out in any scientific communication and would be scrutinized just as any other scientific finding is scrutinized during peer review. Furthermore, those with villainous intentions have much easier ways to mislead science or the general public, for example, by outright fabrication of scientific studies. We feel that carelessness by well-intentioned scientists (in the field, in the lab, or in specifying inappropriate likelihoods or priors) is probably a much more common cause of erroneous inference than is mischief.

5.4.4 Guidance

We admit that the cautionary statements in this section could make the choice of priors seem complicated and difficult; however, that is not our aim. We feel that priors can be an important component of science and can be helpful in obtaining useful models for inference. Our goal in this discussion of priors is to instill a sense of awareness about the decisions

being made in the model-building process. If you are more thoughtful about specifying priors and the associated consequences after reading this section, then we have done our job.

The fact is, few of these details are made clear in other texts on applied Bayesian statistics, and we wrote this section, at least in part, as a reminder to ourselves to think deeply about how we can incorporate prior scientific knowledge in the form of a probability distribution for use as a prior. You'll notice that we commonly use default priors in examples throughout this book. It would seem that by doing so, we encourage this practice, but in reality we don't claim to be experts in all the applied subjects in the diverse examples we offer. Thus, it is with a touch of "do as we say, not as we do" that we suggest that our model specifications throughout are only placeholders for a model that might actually be used by an expert in the relevant field. This section also serves as a prelude to chapter 8, where we give a concrete example of the value of prior information and to chapter 9, where we describe ways that priors are an example of regularization, an approach widely used in statistics to improve model fit.

Although we have provided several approaches for specifying priors for specific models in this section, a list of all possible options for all possible models would be too lengthy. Thus, we echo the guidance provided by Seaman et al. (2012) and leave you with a few further general diagnostics and remedies to consider when specifying priors in Bayesian models:

- Bear in mind that one of the objectives of Bayesian analysis is to provide knowledge that can inform subsequent analyses; the posterior distribution obtained in one investigation becomes the prior in subsequent investigation. Thus, we agree with the view of Gelman (2006) that vague priors are provisional—they are a starting point for analysis. As scientists, we should always prefer to use appropriate, well-constructed informative priors.
- Visualize the prior you choose in terms of the parameters for which you desire inference. We did this earlier for the logit(p) (i.e., fig. 5.4.3). Sometimes you can do this analytically (i.e., with pencil and paper, using calculus), but it's often easier just to simulate values from your prior, then transform them to represent the desired quantity and plot a histogram.
- Perform a prior sensitivity analysis. Try several different priors, maybe by simply choosing different prior variances, and see how much the posterior distribution moves around as a result. Often, you'll see little posterior sensitivity to priors when there is a high ratio of data to parameters. However, if the posterior is sensitive to the prior and you truly desire a prior that is only weakly informative, you will need to rethink your prior by changing its form or parameters. Alternatively, you must carefully justify your choice of prior in relation to the inference you seek.

- An influential prior might be indicated if the posterior inference differs greatly from maximum likelihood inference. Of course, this can be confirmed only in models where both approaches can be implemented easily, so it may not be practical for more complicated Bayesian models. Still, in some Bayesian models, inference will approach what would be obtained with inference based on maximum likelihood if certain priors are used.
- Dependent parameters should probably have priors that acknowledge the dependence. We are often lured into thinking that we can simply specify independent priors for parameters, but a prior that is a joint multivariate distribution or conditional distribution for one parameter given another is often more appropriate. An example in regression is $y_i \sim \text{normal}(\beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i}, \sigma^2)$ for $i = 1, \ldots, n$. In this case, it is common to use the independent normal priors $\beta_1 \sim \text{normal}(0, \sigma_\beta^2)$ and $\beta_2 \sim \text{normal}(0, \sigma_\beta^2)$, but it can be helpful to use a multivariate normal prior for both regression coefficients simultaneously, $\boldsymbol{\beta} \sim \text{multivariate normal}(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\beta}$ is the vector containing β_1 and β_2, $\mathbf{0}$ is a vector of zeros, and $\boldsymbol{\Sigma}$ is a covariance matrix.
- Keep in mind that even with large sample sizes there may be not be enough information in the data to tease apart different parameters, regardless of their priors. This is more of an identifiability problem rather than a problem with the prior, and the form of the model itself should be reconsidered. For example, with binomial data $y_i \sim \text{binomial}(N, p)$ for $i = 1, \ldots, n$, where N and P are unobserved, there are not enough data in the world to learn about both N and p individually, but a strong prior on one of the two parameters (if warranted) can help focus the inference on the other. However, without sufficient prior information this is not a useful model in an inverse (i.e., statistical) setting.