4 Likelihood

It is possible to learn Bayesian analysis with a bare-bones treatment of likelihood, but we include a full chapter on likelihood for two reasons. Likelihood forms the fundamental link between models and data in the Bayesian framework. Understanding this linkage is central to the aims of this book. In addition, *maximum likelihood* is a widely used alternative to Bayesian methods for estimating parameters in ecological models (Hilborn and Mangel, 1997; Bolker, 2008). It will be useful to understand the similarities and differences between Bayesian analysis and analysis based on maximum likelihood. They are more similar than you might think. Thus, although this is a book about Bayesian modeling, we believe it is important to provide an overview of likelihood and maximum likelihood before we turn to Bayesian inference.

4.1 Likelihood Functions

Until now, we have defined probability distributions in terms of a random variable, z, and parameters. This broad definition makes sense because the Bayesian approach views all things that are not observed[1] as random variables and thus requires a broad framework for treating many kinds of quantities. To explain likelihood, however, we begin by narrowing the arguments in probability distributions to include a data point y and parameters $\boldsymbol{\theta}$ in an ecological model. If we have a *known and fixed* value

[1]Data, before we observe them, are treated as random variables. After we observe them, they are fixed.

of θ, then we are able to calculate the probability (or probability density) of a variable observation y conditional on our model being a true representation of the process that gives rise to y. We use one of the probability distributions described in section 3.4 to make that calculation. A couple of simple examples illustrate how we do that.

EXAMPLE 1 You collect data on the number of tadpoles per volume of water in a pond. You observe 14 tadpoles in a 1 L sample. You *know* that the true average number of tadpoles per liter of water in the pond is 23. The probability of your observation is $[y_1|\lambda] = \text{Poisson}(y_1 = 14|\lambda = 23) = .0136$. A second sample contains 34 tadpoles. Conditional on the known, fixed average number of tadpoles in the pond ($\lambda = 23$), the probability of this additional observation is $[y_2|\lambda] = \text{Poisson}(y_2 = 34|\lambda = 23) = .0069$. The joint probability of the data conditional on λ is the product of the individual probabilities, $.0136 \times .0069 = 9.38 \times 10^{-5}$, assuming these observations are independent, which means that knowledge of one observation tells us nothing about the other. We could extend this calculation of joint probability by taking the product of the probabilities of any number of independent observations.

EXAMPLE 2 You are investigating decomposition of leaf litter over time using a simple model of exponential decay, $\mu_t = e^{-kt}$, where μ_t is the mean proportion of the initial mass of leaf litter remaining at time t (units: day), assuming that the proportion at $t = 0$ is 1, and k is the mass specific rate of decay (units: day^{-1}). Because the data are observed proportions, y_t, that can take on values continuously from 0 to 1, you choose a beta distribution to calculate the probability density of an observation given that the parameter k and the variance of the estimates of your model (σ^2) are known and fixed. Using moment matching, you obtain

$$\alpha_t = \frac{\mu_t^2 - \mu_t^3 - \mu_t\sigma^2}{\sigma^2}, \tag{4.1.1}$$

$$\beta_t = \frac{\mu_t - 2\mu_t^2 + \mu_t^3 - \sigma^2 + \mu_t\sigma^2}{\sigma^2}, \tag{4.1.2}$$

$$[y_t|\mu_t, \sigma^2] = \text{beta}(y_t|\alpha_t, \beta_t). \tag{4.1.3}$$

Because $\mu_t = e^{-kt}$, you can also write $[y_t|e^{-kt}, \sigma^2] = \text{beta}(y_t|\alpha_t, \beta_t)$. Conditional on a *known*, fixed decay rate $(k = 0.01\,\text{day}^{-1})$ and *known*, fixed $\sigma^2 = 6 \times 10^{-4}$, you calculate parameters for the beta

distribution of the mass remaining on day 30: $\alpha_{30} = 236.33$ and $\beta_{30} = 82.68$. The probability density that an observation of $y_t = 0.7$ of the mass remains at time $t = 30$ is 4.040.

These examples were deliberately chosen to make things feel entirely backward. In both cases we know the value of the parameters—they are fixed—and we don't know the data, which are random variables. When the parameter is fixed and the data are random variables, we can calculate the probability of an observation (for discrete data) or the probability density of an observation (for continuous data) using the distributions described in section 3.4.

Alternatively, the usual case for ecological researchers is that the parameters are unknown and the data are fixed. That is, we have a set of observations in hand and we want to know what the observations tell us about parameters. We need a way to evaluate the evidence in the fixed data for variable parameter values, and we do this by using a *likelihood function* $\mathrm{L}(\theta|y)$, defined as

$$\mathrm{L}(\theta|y) = [y|\theta]\,. \tag{4.1.4}$$

Equation 4.1.4 simply says that the likelihood of the parameter given the data is equal to the probability (or probability density) of the data conditional on the parameter.[2] For n independent observations,

$$\mathrm{L}(\theta|\mathbf{y}) = \prod_{i=1}^{n} [y_i|\theta]\,, \tag{4.1.5}$$

$$\log(\mathrm{L}(\theta|\mathbf{y})) = \sum_{i=1}^{n} \log[y_i|\theta]\,. \tag{4.1.6}$$

It is important to be clear on terminology. The left-hand side of equation 4.1.4 ($\mathrm{L}(\theta|y)$) is called a *likelihood function*. Bayesians refer to the distribution on the right-hand side ($[y|\theta]$) as a likelihood or a data model to differentiate it from other types of distributions used in Bayesian analysis, as will be discussed in the next chapter.

[2]Older sources on likelihood (Edwards, 1992; Azzalini, 1996; Royall, 1997) express equation 4.1.4 as a proportionality, $\mathrm{L}(\theta|y) = c[y|\theta]$. However, contemporary texts (Pawitan, 2001; Casella and Berger, 2002; Clark, 2007) drop the constant of proportionality c by assuming that $c = 1$. This is permissible because we use likelihood to make comparisons between models or parameter values, one relative to another. For these types of comparisons, the value of c is irrelevant and we simplify notation by letting $c = 1$.

4.2 Likelihood Profiles

The key difference between a probability distribution and a likelihood function is that the parameter is fixed and the data are random variables in a probability mass or density function, whereas in the likelihood function, the data are fixed and the parameters are variable. The relationship between likelihood functions and probability distributions can be most easily seen by plotting them (fig. 4.2.1). If we hold parameters constant and plot a probability density function $[y|\theta]$ as a function of different values of continuously valued y, the area under the curve equals 1 (fig. 4.2.1 A). However, if we hold y constant and plot the same probability density function as a function of θ (fig. 4.2.1 B), we obtain a *likelihood profile*[3] (Hilborn and Mangel, 1997; Bolker, 2008), where the area under the curve does *not* equal 1. A single point is shared between the two curves, showing that probability and likelihood are the same only when the parameter is treated as a fixed quantity.

We see the same type of relationship when the data are discrete. When we hold θ constant and plot a probability mass function $[y|\theta]$ for different values of y, then $\sum_y [y|\theta] = 1$. (fig. 4.2.1 C). However, when we plot the same probability mass function with y fixed as a function of a varying θ, we get a likelihood profile (fig. 4.2.1 D), and, again, the area under the curve does not equal 1.

The units of the y-axis of a likelihood profile are arbitrary and can be scaled to any quantity (because likelihood is strictly defined in terms of a multiplicative constant; see footnote 2). Often, the likelihood profile is scaled such that the peak of the curve equals 1, which is accomplished by dividing all the likelihoods by the likelihood at the peak of the profile, that is, the maximum likelihood. This provides a convenient scaling, but it does not change the relationship between likelihood and probability.

These plots (fig. 4.2.1) illustrate an important but somewhat subtle distinction between likelihood functions and probability distributions. Saying that the parameter θ is not fixed allows us to calculate $L(\theta|y)$ by allowing it to vary, but this does *not* mean that θ is treated as a random variable in the likelihood framework. It is not a random variable, because random variables are defined as quantities governed by probability distributions, and likelihood functions do not define the probability or probability density of θ. This distinction causes some authors to use notation aimed at preventing any confusion with conditional probability; that is, they use $L(\theta; y)$ instead of $L(\theta|y)$.

[3] Also called a "likelihood curve" (Edwards, 1992).

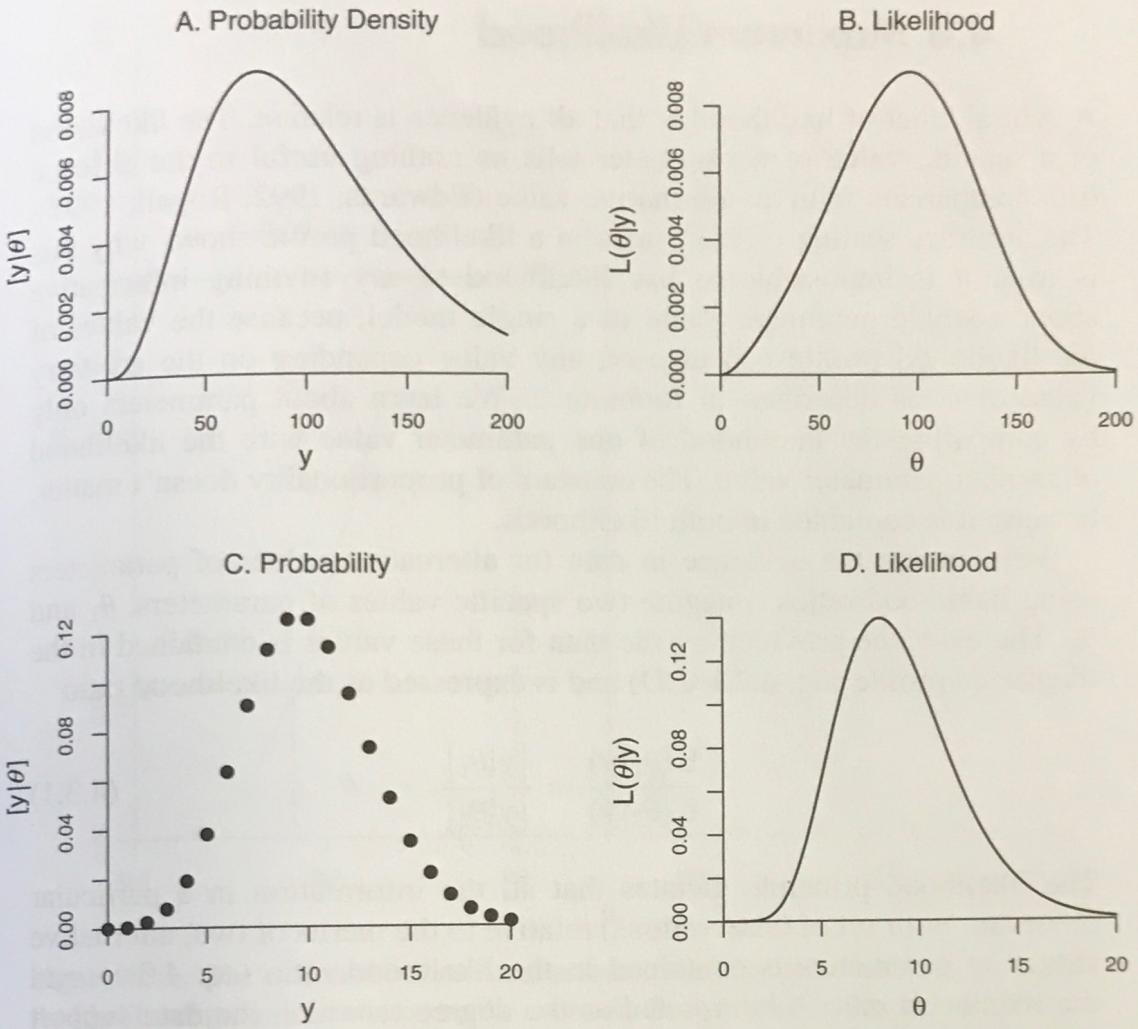

Figure 4.2.1. Illustration of relationships between probability distributions and likelihood profiles for continuous (**A, B**) and discrete (**C, D**) data. Assume we have continuous, strictly positive data with a fixed mean $\theta = 100$ and variance $\sigma^2 = 2500$. Panel **A** shows a gamma probability density function plotted over $y = 0, \ldots, 200$ conditional on the fixed parameters $\alpha = \theta^2/\sigma^2$, $\beta = \theta/\sigma^2$. The integral of this function taken from 0 to ∞ is 1. Assuming a fixed value of $y = 75$ and a fixed value of $\sigma^2 = 2500$, panel **B** shows a gamma probability density function (with moments matched to parameters) over $\theta = 0, \ldots, 200$. This curve is a likelihood profile, and it does not integrate to 1. A single point is shared between the two curves, $[y|\theta, \sigma^2]$, where $y = 75, \theta = 100$, and $\sigma^2 = 2500$, illustrating that probability and likelihood are the same only when the parameters are treated as fixed. In panel **C**, we assumed a fixed mean, $\theta = 10$, for a Poisson probability mass function of the data $y = 1, 2, 3, \ldots, 20$. The sum of the probabilities from 0 to ∞ equals 1. In panel **D**, a likelihood profile is plotted as the Poisson probability mass function, where its mean θ varies from 0 to 20 assuming a fixed $y = 8$. The integral of this curve over 0 to 1 does not equal 1.

4.3 Maximum Likelihood

A central tenet of likelihood is that all evidence is relative. The likelihood of a specific value of a parameter tells us nothing useful in the absence of a comparison with an alternative value (Edwards, 1992; Royall, 1997). The arbitrary scaling of the y-axis in a likelihood profile shows why this is true; it is impossible to use likelihood to say anything informative about a single parameter value or a single model, because the values of the likelihood profile can take on any value depending on the arbitrary value of c, as described in footnote 2. We learn about parameters only by comparing the likelihood of one parameter value with the likelihood of another parameter value. The constant of proportionality doesn't matter, because it is contained in both likelihoods.

We compare the evidence in data for alternative values of parameters using likelihood ratios. Imagine two specific values of parameters, θ_1 and θ_2. The evidence provided by the data for these values is contained in the likelihood profile (fig. 4.2.1 C,D) and is expressed as the likelihood ratio

$$\frac{\mathrm{L}(\theta_1|y)}{\mathrm{L}(\theta_2|y)} = \frac{[y|\theta_1]}{[y|\theta_2]}. \tag{4.3.1}$$

The likelihood principle dictates that all the information in a particular observation (or set of observations) relative to the merits of two, alternative values of parameters is contained in the likelihood ratio (eq. 4.3.1), and the likelihood ratio is interpreted as the degree to which the data support one value of a parameter over another (Edwards, 1992; Azzalini, 1996; Royall, 1997). The evidence in data for alternative values of parameters is often obtained using the natural logarithm of the likelihood ratio, which is formally defined as the *support* for one value of a parameter over another, conditional on the data.[4]

Thus, the ratio of the likelihoods or the difference between the log likelihoods provides the basis for evaluating the evidence in data for alternative values of parameters (fig. 4.3.1). In most practical problems in ecology, we are particularly interested in the value of the parameter θ that has the maximum support in data, which is found at the peak of the

[4]There is potential for confusion here. Edwards (1992) defined support as the log of the likelihood ratio. Statisticians also use support to mean the domain of a probability function or a probability density function, where the values of the probability or probability density of a random variable exceed 0. So, used this way, the support for the Poisson distribution is the set of all nonnegative integers; the support for the normal distribution is all real numbers. For the remainder of the book, we use the latter meaning.

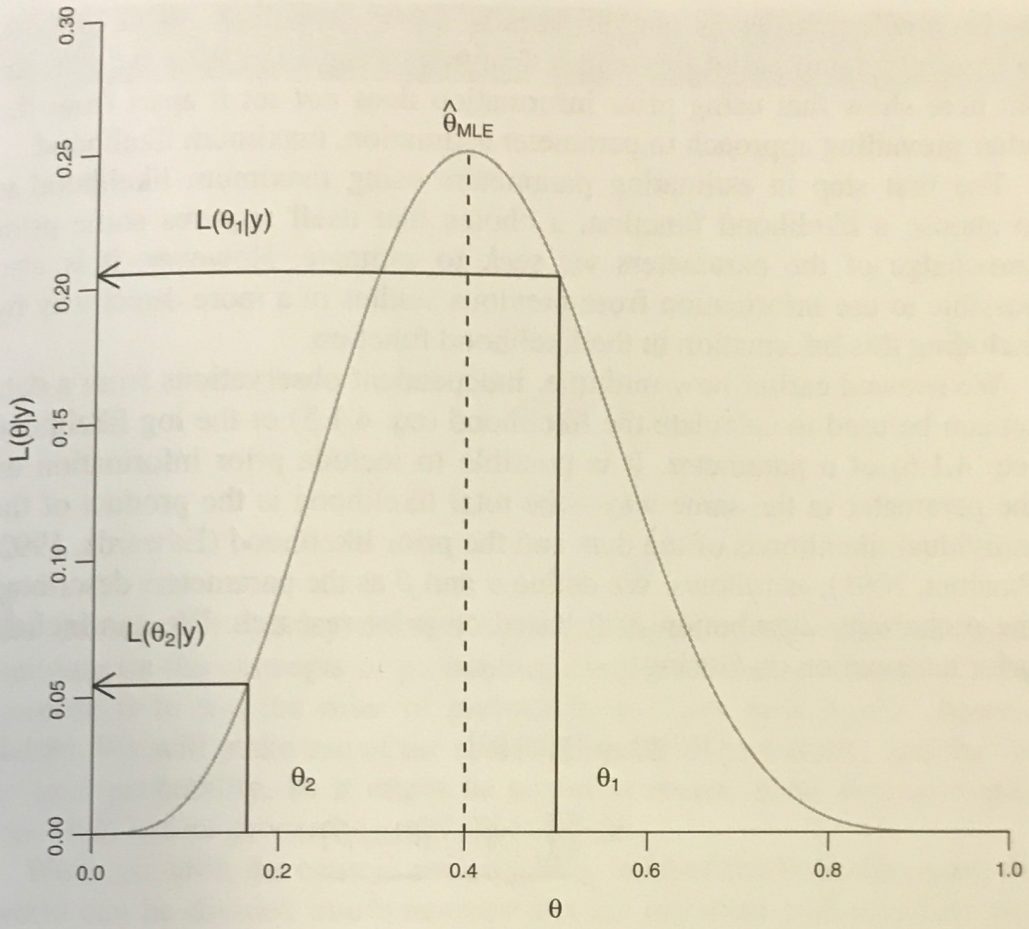

Figure 4.3.1. The evidence in data for alternative values of parameters is contained in the likelihood profile, illustrated by the curve. In the example here, there is greater evidence for $\theta = \theta_1$ relative to $\theta = \theta_2$, because $\mathrm{L}(\theta_1|y) > \mathrm{L}(\theta_2|y)$ (arrows). The maximum likelihood estimate for $\theta = \hat{\theta}_{MLE}$ is the value of θ for which $\mathrm{L}(\theta|y)$ is greater than for any other value (dashed line).

likelihood profile (fig. 4.3.1). This is the value of θ that maximizes the likelihood function (eq. 4.1.5) or the log likelihood function (eq. 4.1.6). We can find this maximum likelihood value of θ analytically for simple models or by numerical methods for more complex ones (Hilborn and Mangel, 1997; Pawitan, 2001; Bolker, 2008).

4.4 The Use of Prior Information in Maximum Likelihood

It is often heard that the difference between Bayesian analysis and likelihood is that Bayes uses prior information, although it is the treatment of

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

$$\begin{aligned} L(\theta|\mathbf{y}) &= [\mathbf{y}|\theta]\,[\theta] \qquad (4.4.1) \\ &= \underbrace{\prod_i [y_i|\theta]}_{a}\,\underbrace{[\theta|\alpha, \beta]}_{b}\,. \end{aligned}$$

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.*