1. **Clark, James S.; Lavine, Michael. 2001**. Bayesian statistics estimating plant demographic parameters. In: Scheiner SM, Gurevitch J, eds. Design and Analysis of Ecological Experiments. Oxford, UK:Oxford University Press. p 327–346. Chapter 17.

Bayesian analysis differs from other topics in this book, so we approach it in a different way. Our gentle introduction is intended for the ecologist who might find either Bayesian or classical approaches useful, depending on the application at hand. So our chapter includes some comparisons, but they are not the insidious examples that rely on strange or unrealistic distributions to generate discord. Although most of this book is designed for the practitioner, providing the bridge from concept to software, Bayesian analysis still requires programming. Thus, although we cannot direct the reader to a broad range of software options, we adopt the general philosophy of this volume by providing a simple and practical introduction to a topic that is generally treated at a more advanced level in graduate statistic courses and beyond.

17

Bayesian Statistics

Estimating Plant Demographic Parameters

JAMES S. CLARK
MICHAEL LAVINE

17.1 Introduction

There are times when external information should be brought to bear on an ecological analysis. Experiments are never conducted in a knowledge-free context. The inference we draw from an observation may depend on everything else we know about the process. Bayesian analysis is a method that brings outside evidence into the analysis of experimental and observational data.

With the increasing use of Bayesian methods in ecology, our science has co-opted the philosophical controversy that attended the twentieth-century rise of "classical" statistics (Stigler 1986). Limitations of classical hypothesis testing and *P*-values (Berger and Berry 1988) on the one hand, or of Bayesian priors and subjective probability (Dennis 1996) on the other, allow smart people to come down on either side of a polarized debate (Edwards 1996). The debate will undoubtedly continue.

This chapter is not one of the battlegrounds over Thomas Bayes' thinking when he described his famous "billiard" example or its application since (Bayes 1763; Fisher 1959; Stigler 1986). Although we are not always enamored with classical hypothesis testing in general, we often use it. And, although priors can sometimes sound like a bad idea in theory, it is usually harder to abuse them than some people think. Regardless of whether it is always sensible to regard an unknown parameter as having a distribution of values (in a Bayesian sense), this can be the best way to model many ecological problems. For those of us most interested in interval estimation, the fact that both methods usually give similar answers tends to be lost in the fray. The large divergences that can occur with

small sample sizes or strong prior opinions are less common than the impression left by some authors. We leave these arguments for others, but refer readers to the lively treatments in the Special Feature of *Ecological Applications* (Dixon and Ellison 1996). Hilborn and Mangel (1998) lay out the utility of Bayesian methods as part of a set of tools for analysis of ecological data. An overview of Bayesian statistics is given in Berger (in press).

Bayesian analysis differs from other topics in this book, so we approach it in a different way. Our gentle introduction is intended for the ecologist who might find either Bayesian or classical approaches useful, depending on the application at hand. So our chapter includes some comparisons, but they are not the insidious examples that rely on strange or unrealistic distributions to generate discord. Although most of this book is designed for the practitioner, providing the bridge from concept to software, Bayesian analysis still requires programming. Thus, although we cannot direct the reader to a broad range of software options, we adopt the general philosophy of this volume by providing a simple and practical introduction to a topic that is generally treated at a more advanced level in graduate statistics courses and beyond.

We cannot go far using Bayesian methods without the routine application of calculus (including numerical methods that require an understanding thereof). Models with multiple parameters get complicated fast, but the conceptual background laid by simpler models generally applies. Rather than attempt a broad survey that would risk losing the reader in technique, we limit this chapter to one sampling distribution (the binomial). This limited scope allows us to introduce a number of concepts (the basic elements of Bayesian methods, conjugacy, comparison with classical methods) that apply generally. Useful introductory texts include Berry (1996), Lee (1989), and Box and Tiao (1973).

17.2 The Basic Elements

Bayesian statistics has two distinguishing characteristics:

1. It combines, in a formal way, data from the experiment at hand with data from any other experiment or information deemed relevant.
2. It summarizes the analysis with a probability distribution that shows how well the various values of the parameter are supported by all of this information.

For the purpose of illustrating concepts, we begin with a simple example. To understand the dynamics of plant populations, ecologists estimate survival from census data. Because annual rates tend to be high (often >95% per year for trees), it can be difficult to obtain data sufficient to make confident estimates (i.e., enough deaths). Information that is external to the study at hand can help to sharpen estimates. This Bayesian example combines census data from a typical field study with external information to evaluate survival of *Acer rubrum* trees in the southern Appalachian Mountains (Wyckoff and Clark 2000).

The probability density in figure 17.1A summarizes the analysis of tree survival. The data include annual censuses of trees; the parameter of interest is the

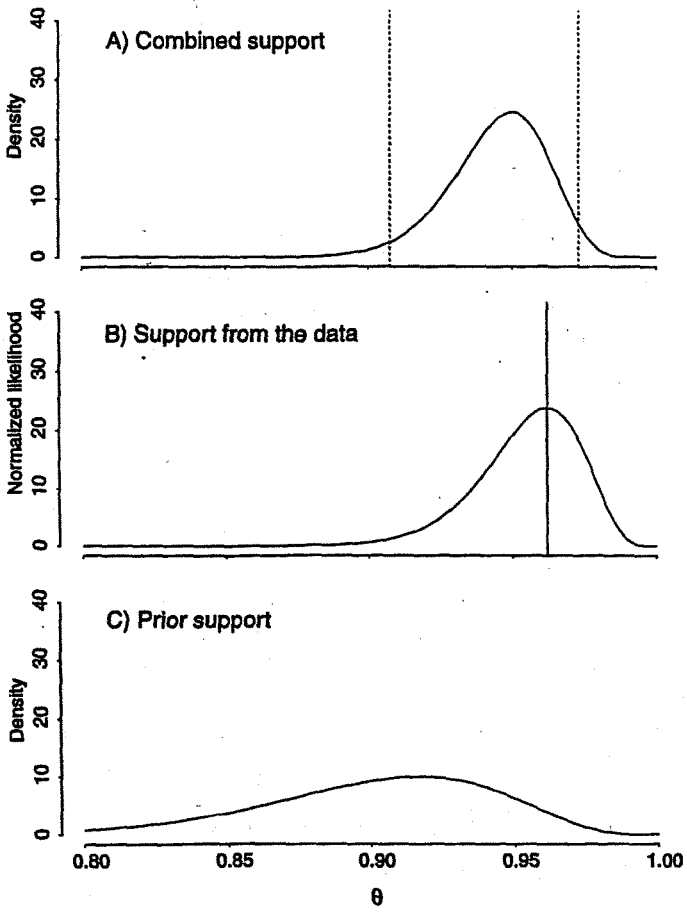

Figure 17.1 The elements of a Bayesian analysis. The posterior density (A) represents the support of values of the survival rate θ in light of data at hand (B) and prior inputs (C).

probability θ that a randomly selected tree survives from one year to the next. Figure 17.1A shows that values of θ around 0.95 are most likely, but all values from about 0.91 to about 0.98 are plausible. The data support the value $\theta = 0.95$ about twice as well as 0.93 or 0.97 and about ten times as well as 0.91 or 0.98.

Where did figure 17.1A come from? The repeated censuses of Wyckoff and Clark (2000) included 127 survivors from a total of 132 individuals. The chance of 127 survivals from 132 total trees is calculated from a binomial distribution as

$$f(\text{data}\,|\,\theta) = \Pr[127 \text{ survivals in } 132 \text{ trees}] = \binom{132}{127} \theta^{127} (1-\theta)^5 \qquad (17.1)$$

This function of θ is called a likelihood function and is plotted in figure 17.1 The likelihood function indicates how well the data support each value of θ. this example, the data best support the value θ = 127/132, and they lend decre ing support to values on either side. The value best supported by the data has maximum likelihood and is termed the maximum-likelihood (ML) estimate. T likelihood function in figure 17.1B shows only the support from the data, so i not yet a Bayesian analysis.

A Bayesian analysis combines the likelihood function of figure 17.1B w other information, which could be data from other experiments or scientific sight. The other information is summarized by a function $f(\theta)$ called the pr density. For example, if we had reason to believe, say, from previous exp ments, that θ was most likely to be about 0.9 and very likely to be somewh between 0.85 and 0.95, then we might use a prior density that looks like fig 17.1C. The name *prior* is used because the other information usually comes to prior to the information from the experiment we are analyzing. But priority time is not necessary. Perhaps a better term would be external information (R tery and Zeh 1993).

The likelihood function and prior density are combined according to Bay theorem,

$$f(\theta \mid \text{data}) = \frac{\text{prior} \times \text{likelihood}}{\int \text{prior} \times \text{likelihood}\, d\theta} = \frac{f(\theta) f(\text{Data} \mid \theta)}{\int f(\theta) f(\text{Data} \mid \theta) d\theta} \qquad (17$$

The left-hand side is called the posterior density and is the combination of lik hood and prior. The theorem says that the posterior density is calculated by mu plying the prior density by the likelihood. (The integral in the denominator is a normalizing constant.) The posterior represents, at least approximately, h well various values of θ are supported by all the information, including data hand and the prior information. The prior and likelihood pictured in figures 17 and 17.1C, respectively, combine to give the posterior in figure 17.1A.

In summary, a Bayesian analysis takes as inputs both data, by way of a lik hood, and additional information, summarized by the prior, to produce a poste density. The posterior expresses how the combination of data and prior toge support values for the unknown parameter.

17.3 A Few Details

To introduce some of the techniques necessary to arrive at a posterior, we pu a bit further the example from Wyckoff and Clark (2000), who compared m mum-likelihood and Bayesian approaches to estimate how survival rates cha over time.

17.3.1 Arriving at a Posterior

Let n be the number of trees counted at a first census, and k be those that sur to be counted in the second census. For the *Acer rubrum* example, with $n =$

and $k = 127$, intuition tells us that the best estimate of θ is simply $k/n = 0.96$, the fraction that survive. The strong intuitive sense we have about this simple problem helps us grasp the basics of a Bayesian approach. Recall the two ingredients, a likelihood and a prior density for the parameter of interest θ. Rewriting equation 17.1 in terms of generic parameters gives the binomial likelihood,

$$f(k \mid \theta) = \binom{n}{k} \theta^k (1 - \theta)^{n-k} \tag{17.3}$$

where θ is the probability of survival, and the combinatorial $\binom{n}{k}$ is the binomial coefficient. The prior $f(\theta)$ depends on insights about θ other than those obtained from the data. The flat prior, representing the view that all values of θ are equally probable, is a uniform density on the interval (0, 1):

$$f(\theta) = 1 \qquad 0 < \theta < 1 \tag{17.4}$$

(dashed line in figure 17.2A). We might use a flat prior if we desire an outcome that is influenced by the data alone and not by external information. To write

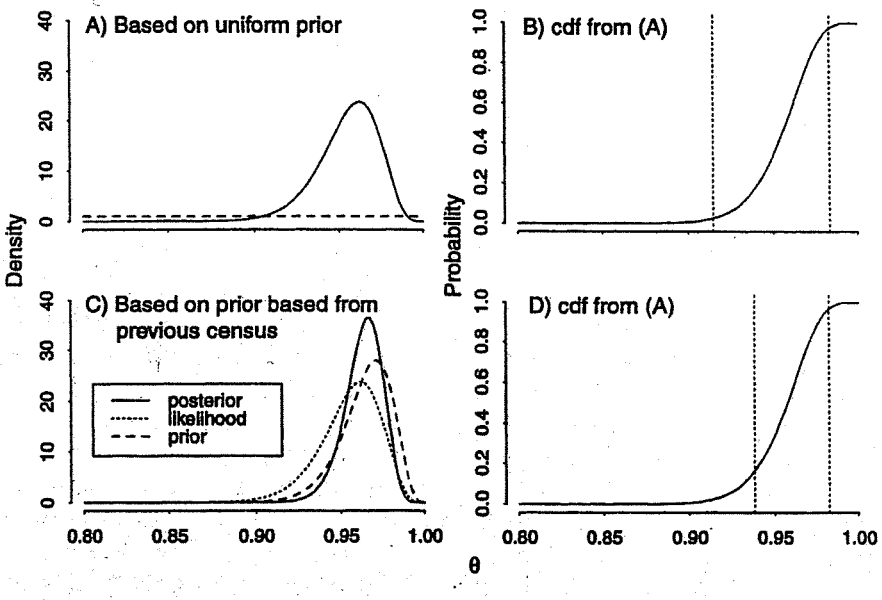

Figure 17.2 Bayesian analyses for uniform (A) and nonuniform (B) priors. In part (A) the posterior coincides with the likelihood. The cumulative plots at right are the cumulative posterior distributions with vertical dashed lines enclosing 95% of the posterior densities.

Bayes' rule (equation 17.2), we note that the combinatorials in the numerator and denominator cancel, leaving

$$f(\theta \mid k) = \frac{\theta^k(1-\theta)^{n-k}}{\int_0^1 \theta^k(1-\theta)^{n-k}\, d\theta} = \frac{\theta^k(1-\theta)^{n-k}}{B(k+1, n-k+1)} = (n+1)\binom{n}{k}\theta^k(1-\theta)^{n-k} \qquad (17.5$$

$B(\bullet, \bullet)$ is the beta function. The final step in equation 17.5 makes use of some well-known relationships among beta functions, gamma functions, and factorials that can be found in standard probability texts.

The posterior in equation 17.5 is a beta density, $f(\theta|k) = B(k+1, n-k+1)$ and expresses the level of certainty assigned to values of θ. The mode of this density is the most probable value of θ and occurs at the critical point, where $df(\theta|k)/d\theta = 0$. Differentiation is simplified if we first take logs, $d\ln f(\theta|k)/d\theta = \frac{k}{n} - \frac{n-k}{1-\theta}$. Setting this derivative equal to zero shows that the mode of the posterior agrees with our intuitive estimate k/n (figure 17.2A). We can summarize our degree of confidence in θ with quantiles that contain the central $100(1-\alpha)\%$ of the posterior. The right-hand side (figure 17.2B) is the cumulative distribution for the posterior showing 95% quantiles (dashed lines).

Now consider how our noninformative (flat) prior affects the result. The uniform prior density means that the posterior beta density (equation 17.5) has the same shape as the likelihood function (equation 17.3); the two differ only by a constant and, thus, contain the same information about the parameter θ. The normalized likelihood (divide the likelihood function by the denominator of equation 17.1) coincides with the posterior in figure 17.2A. Because we had no prior insight, the census data governed the result. Before considering how the posterior is influenced by the particular choice of the prior, we compare the Bayesian method with a classical approach.

17.3.2 Comparison with a Classical Approach

How does this Bayesian approach differ from a classical view? A classical (frequentist) approach might involve fitting the parameter θ to data and then deriving a probability statement (a *P*-value) based on a comparison of that result with some alternative null model. The maximum-likelihood (ML) estimate of θ is that which maximizes the probability of the data set, assuming the model to be correct. By differentiating the likelihood of equation 17.3 with respect to θ, we find the ML estimate of θ to be $\theta_{ML} = k/n$. A classical confidence interval is based on the comparison of this ML estimate with other possible values of the parameter using *P*-values.

We refer to all such intervals, be they classical or Bayesian, as "confidence intervals"; ecologists do not use the Bayesian jargon "credible interval." To compare confidence intervals, we describe a "likelihood profile" approach. Likelihood profiles are being used increasingly by ecologists; a full description of likelihood

profiles can be found in Hilborne and Mangel (1997). The summary of likelihood profiles that follows illustrates the link between classical and Bayesian approaches represented by the likelihood function.

Within a classical context, a probability statement about an estimate requires some alternative hypothesis against which it can be compared. Because there might be many such alternatives, let's consider a broad range. This range is the basis for a classical confidence interval, which we obtain by constructing a likelihood profile. The method involves successively calculating two likelihoods for the same data set, one for each competing hypothesis about θ against the ML estimate, that is, the value obtaining most support from the data. The likelihood ratio (LR) is simply the ratio of the likelihoods of the two models,

$$LR(\theta_0, \theta_{ML}) = \frac{f(k \mid \theta_0)}{f(k \mid \theta_{ML})} = \left(\frac{\theta_0}{\theta_{ML}}\right)^k \left(\frac{1-\theta_0}{1-\theta_{ML}}\right)^{n-k}$$

The deviance is the test statistic. It is twice the difference in log likelihoods

$$D(\theta_0) = -2\ln\left(\frac{f(k \mid \theta_0)}{f(k \mid \theta_{ML})}\right) = -2\ln LR$$

and is distributed as χ^2 with 1 degree of freedom (there is one parameter at issue). The deviances increase (figure 17.3A) and associated *P*-values decrease (figure 17.3B) as the hypothesized value of θ deviates from the ML estimate ($\theta = 0.96$).

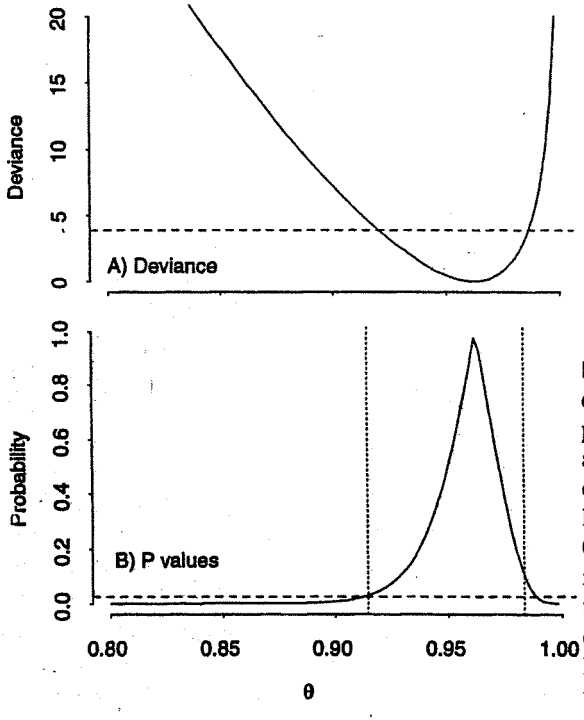

Figure 17.3 Classical confidence intervals for the example in figure 17.2A. The deviance (A) has a horizontal dashed line corresponding to likelihood profile values of $P = 0.05$. The plot of *P*-values (B) includes the $P = 0.025$ horizontal line and the 95% Bayesian confidence interval (vertical lines) from the example in figure 17.2A.

We might conclude from this example that the data allow us to reject, at the $\alpha = 0.05$ level, the hypothesis that θ lies outside the interval bounded by the horizontal dashed line at $P = 0.025$ in figure 17.3B. We can represent the same interval by the horizontal line through the deviance plot at 3.84 (figure 17.3A): the χ^2-value that yields $P = 0.95$ with 1 degree of freedom. For comparison, we include in figure 17.3B the Bayesian confidence interval obtained with a uniform prior in figure 17.2A (the vertical dashed lines in figure 17.3B).

How different are the interpretations derived from these approaches? The astute reader will note that the confidence intervals from the Bayesian (vertical lines in figure 17.3B) and the classical method (horizontal line in figure 17.3B) nearly coincide. Indeed, with large n, they converge. The lower limits are equivalent, and the upper limits differ slightly. If both methods yield the same confidence intervals, then how important is the distinction? From equation 17.1 (and 17.5), we note that a uniform prior (a reflection of prior ignorance about θ) means that the posterior is simply a normalized likelihood function. Because the likelihood and posterior bear the same shape (we cannot distinguish them in figure 17.2A), they contain the same information about θ. The posterior is completely controlled by the data, without prior bias. And the posterior (normalized likelihood) yields about the same confidence interval as the likelihood profile. This example is general; with large sample size, a noninformative prior produces a confidence interval that converges with the classical one.

Despite similarities, statisticians talk about these two confidence intervals in different ways. The classical confidence interval is taken to cover the fraction of repeated experiments in which the interval would contain the true value of the fixed parameter. If we were to conduct a large number of identical experiments on survival of trees that are subject to the same set of risks, our survival estimate would fall above the dashed lines in figure 17.3A in 95% of those experiments. The Bayesian confidence interval represents our belief that the random parameter spans a certain interval. Here, survival is viewed as random with a density given in figure 17.2A. There are cases where the two approaches can yield importantly different answers (e.g., Cousins 1995). However, from a practical standpoint, it is worth remembering that much of the time the confidence intervals nearly coincide.

17.3.3 An Informed Prior

For problems like tree survival, where estimates suffer from inadequate data, prior (external) knowledge about θ can sharpen our inference. Wyckoff and Clark (2000) incorporated estimates obtained from U.S. Forest Service (USFS) inventories as prior estimates of survivorship. The USFS data might not provide the best estimate for Wyckoff and Clark's (2000) study site, because the data come from a broad region, but they do represent a prior estimate of survival that might be worth combining with field data from their more restricted study area.

USFS data contained $k_0 = 137$ of $n_0 = 142$ surviving *Acer rubrum* trees from the region that includes the study area of Wyckoff and Clark (2000). The prior

density, likelihood function, and posterior density are compared in figure 17.2B. For convenience, the prior is taken to be a beta density,

$$f(\theta) = B(k_0, n_0) \tag{17.6}$$

Note that the parameters of this density are simply the numbers of total and surviving trees. Because the prior from USFS data in equation 17.6 (shown in figure 17.2C) contains far more information about θ than does the uniform one (figure 17.2A), the posterior in figure 17.2C is concentrated about the most probable estimate (still $\theta = 0.96$) to a greater degree than in figure 17.2A. The posterior splits the difference between prior and likelihood, because it incorporates information from both. The greater information that results from the informed prior is reflected in the narrower confidence intervals shown on the right-hand side of figure 17.2.

The beta-binomial example makes obvious the importance of sample size. In our example, the weight of the prior evidence ($n_0 = 142$) and of the new data ($n = 137$) are about the same. The exact solution for the posterior is the beta density with parameters obtained by summing the prior information and data:

$$f(\theta \mid k) = B(k_0 + k,\ n_0 + n - k_0 - k) \tag{17.7}$$

Because the parameters in equation 17.7 are simply the sums of total and surviving trees, it is clear that both contribute similar weight to the posterior. For a small sample size, an informed prior ($n_0 \gg n$) dominates the posterior; the likelihood (i.e., the data) has minimal impact. With increasing sample size ($n \gg n_0$), the likelihood dominates the prior, and the posterior approaches the likelihood. The example using a flat prior (figure 17.2A) is an extreme case, where the weight of the evidence is concentrated in the likelihood. Thus, the impact of the prior is felt most when sample size is low. With an increasing sample size, the posterior tends to normality, with the mean approaching the "true" value of θ_0, and the parameter variance is determined by the curvature of the likelihood surface at θ_0. Thus, with a large sample size, the likelihood alone can be used to estimate the mode and curvature. Provided that the prior assigns nonzero probability to the true value θ_0, the curvature increases with increasing n until the mass of the posterior is concentrated at the point θ_0.

One objection to Bayesian methods is that subjectivity may creep into the analysis through the choice of the prior. In the hope of reducing subjectivity, some practitioners recommend using a flat prior. As we have seen, this approach yields a posterior density of the parameter based on the data and on an initial belief that all values are equally probable. (The foregoing section explains similarities to a classical approach.) Although some strict subjectivist Bayesians might disagree, it is generally good practice to consider several different priors, representing different evaluations of outside information, use them each to compute a posterior, and compare the posteriors. Often the posteriors that result from different priors will be similar (Crome et al. 1996). Wyckoff and Clark (2000) determined how the survival estimates changed when using priors obtained from different data sources. In their analysis, changing the prior had little effect on the posterior, because there were not large discrepancies between the priors obtained

from different sources and the likelihood. But, in some cases, they can be qui different, meaning that people can disagree. Wolfson et al. (1996) provide an e ample of how sample size can be adjusted to ensure a decisive experiment outcome when different parties bring to a problem very different priors. A larg sample size might be needed to demonstrate an outcome that is at odds wi strong prior evidence.

17.3.5 Conjugacy

If a probability statement about parameters is the only objective, then a Bayesi analysis can often be done without resorting to the mathematical details behind, sa equation 17.7. Indeed, increasing complexity, such as in the example that follow demands a computer-intensive approach. Numerical techniques, such Mark Chain Monte Carlo (MCMC) simulation, are well suited to analyzing such mode and calculating posterior distributions for the parameters of interest. Gelman al. (1995) provide an introduction to these methods. The route to the posterior intractable, and the nonparametric nature of the posterior means that it is n readily transported from one application to the next.

Much ecological investigation is concerned with developing models for unde standing and prediction. Knowing the numerical techniques for extracting con dence intervals from high-dimensional posteriors is often not enough. The dev opment of minimal models that permit transparent error propagation and analys is a goal of ecological research (Hilborne and Mangel 1997; Burnham and And son 1998).

A special class of models is analytically tractable when the number of param ters is small and provides a powerful technique for data assimilation. It involv a special relationship between prior and likelihood termed *conjugacy*. A conjug prior-likelihood pair is one for which application of Bayes' rule results in a pos rior having the same form as the prior. Conjugate prior-likelihood pairs can found for many low-dimensional problems. The beta-binomial is a common e ample: the prior (equation 17.6) and posterior (equation 17.7) have the same for and only the parameter values are updated. There are a number of conjugate pa (we mention the invχ^2-Gaussian conjugate pair subsequently), and their use ways simplifies the analysis. Conjugacy is a valuable tool, because it permits exact result that can be updated repeatedly. For example, a model of forest d namics can be implemented in fully parametric form. A standard model of t sort using, say, a ML (point) estimate of survival probability does not reflect uncertainty described by figure 17.2B. The conjugate pair model allows us draw survival estimates directly from equation 17.7, thus propagating uncertai in the parameter estimate directly to the model output. Moreover, the next oc sion to update the data set requires only a change in the parameters of equati 17.7. Although their calculation requires some math, conjugate pairs provide most transparent view of the relationship between priors and posteriors, a when available, they provide a powerful way to assimilate data in ecologi models. Of course, in the many cases where a conjugate pair is not available, analysis must proceed numerically.

17.4 Bayesian Estimation in a Dynamic Model

The methods outlined previously can be extended to more complex ecological problems. Here we provide an example that uses the binomial sampling distribution to estimate the dynamics of seedlings. Ecologists typically study such dynamics by tagging every seedling in a study plot and, through censuses, determining survival probabilities. Such studies are so labor-intensive that few data sets exist (Clark et al. 1999). Moreover, the heavy loss of tags and failure to relocate seedlings necessitate far more complex statistical models than investigators actually use to analyze such data. The following example from Lavine et al. (in press) illustrates how the Bayesian approach can be implemented in a dynamic model to incorporate different types of error and, in the process, extract parameter estimates without the intensive labor required by the standard approach. The model is based on identification of only two classes of seedling age, and it uses local densities at each stage rather than individually tagged seedlings.

Tree seedlings can be conveniently separated into a first-year class and a >first-year class, which is presumably less susceptible to mortality risks. A dynamic model based on this two-stage classification is readily applied to field data, because the two classes of seedlings are distinguished by the presence of bud-scale scars on >first-year seedlings. The data consist of seedling densities of the two classes counted in 1-m^2 quadrats. First-year and >first-year seedlings are termed *New* (N) and *Old* (O), respectively. Each class has its own survival probability, θ_N and θ_O. The number of new seedlings entering the population in year j, N_j, is determined by input of new seeds. The number of old seedlings, O_j, is the sum of both new and old seedlings from last year that survived to this year. To simplify notation in the equations that follow we define the numbers of both classes that survivor to year j as

$$Y_j = \text{Bin}(N_{j-1}, \theta_N)$$
$$X_j = \text{Bin}(O_{j-1}, \theta_O)$$

The first of these equations says that the number of survivors is a random variable drawn from a binomial distribution with parameters N_{j-1} (the number of potential survivors) and θ_N (the probability that any one individual survives). The total number of old seedlings is the sum of old and new seedlings that survived from year $j-1$, $O_j = X_j + Y_j$. In other words, the number of old seedlings is the sum of two binomial variates, each with its own survival probability. Data from one of the 1-m^2 quadrates are shown in table 17.1.

Table 1 Numbers of first-year (new) and >first-year (old) *Acer rubrum* seedlings censused in quadrate 9 from Lavine et al. (in press)

Year	1993	1994	1995	1996	1997
Number of new seedlings, N_j	1	0	2	0	0
Number of old seedlings, O_j	1	1	1	1	0

In 1996 (call this year j), one old seedling was observed. We do not know whether this survivor was old or new in 1995 (year $j-1$). To estimate the two survival probabilities, we must enumerate the possibilities. As many as 1 or as few as 0 could have been old; the remaining 0 or 1 was new in 1995. From 1 old and 2 new in 1995, the probability (i.e., likelihood) that exactly 1 survived to become old seedlings in 1996 is

$$f(O_{1996} = 1 \mid \theta_O, \theta_N) = \sum_{x=0}^{1} \binom{O_{1995}}{x} \theta_O^x (1-\theta_O)^{O_{1995}-x}$$

$$\binom{N_{1995}}{O_{1996}-x} \theta_N^{O_{1996}-x} (1-\theta_N)^{N_{1995}-O_{1996}+x}$$

This likelihood is the sum of two binomial probabilities. The summation from $x=0$ to 1 adds the two ways we could observe one old seedling. If the old seedling was new last year ($x=0$ in the summation), we have the probability that the single old individual died (the first binomial) times the probability that one of the two ($O_j - x = 1 - 0 = 1$) $N_{j=1}$ survived. [Lavine et al. (in press) provide simple rules for obtaining the summation limits.] By adding to this value the probabilities that would apply if a single O_{j-1} survived ($x=1$), we obtain the total probability of obtaining the data. The likelihood function for the whole data set is the product over all Q quadrats and all T years,

$$f(\{O_{i,j}\} \mid \theta_O, \theta_N) = \prod_{i=1}^{Q} \prod_{j=1}^{T} L(O_{i,j} \mid \theta_O, \theta_N) \tag{17.8}$$

$$= \prod_{i=1}^{Q} \prod_{j=1}^{T} \sum_{x=x_{\min}}^{x_{\max}} \binom{O_{j-1}}{x} \theta_O^x (1-\theta_O)^{O_{j-1}-x} \binom{N_{j-1}}{O_j - x} \theta_N^{O_j - x} (1-\theta_N)^{N_{j-1}-O_j+x}$$

In this particular likelihood, we treat each plot in each year as independent. [Lavine et al. (in press) relax this assumption].

Because of the number of parameters involved, calculating a posterior can require some numerical tools. As written, the posterior can thus far be calculated exactly. Combining the likelihood of equation 17.8 with a flat prior results in a posterior density for θ_N and θ_O (figure 17.4A). The posterior shows how well each combination of (θ_N, θ_O) is supported by the evidence. To examine θ_N only, we integrate over θ_O to obtain the marginal density of θ_N:

$$f(\theta_N \mid O) = \int_0^1 f(\theta_N, \theta_O \mid O) d\theta_O$$

This integration is necessary because parameters in complex models can often be correlated (a type of ill conditioning we mention subsequently). Because the two parameters are largely independent (there is little sign of correlation in figure 17.4), the marginal density obtained from this integration might not be too different from a conditional density (at, say, the ML estimate of the other parameter).

In the real world, other sources of error require parameterization. Because not all seedlings will be found in all years, there is random "findability," which can be thought of as the probability that a seedling is counted at all. In particular,

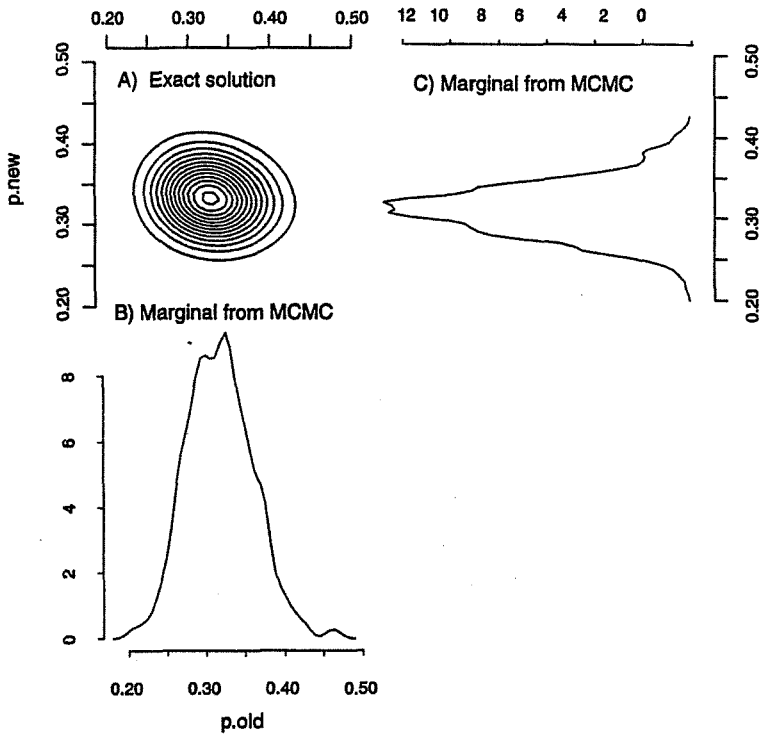

Figure 17.4 (A) Contour plot of the posterior density $f(\theta_N, \theta_O|O)$ shows that values near 0.32 are the best support for both parameters. Marginal posterior densities for the survival parameters θ_N (B) and θ_O (C) obtained by Gibbs sampling from the joint posterior density for the model that includes a findability error (Lavine et al. 2000) are in agreement with those obtained by the simpler model (A).

new seedlings can be small and hard to find. If, for example, there is a Poisson distribution of new seedlings with parameter (mean value) λ and a probability f that a new seedling is actually found, then the problem is too complex to pursue analytically. For the data set considered by Lavine et al. (in press), the marginal posteriors for the survival probabilities are shown in figure 17.4B,C. The marginal distributions are bumpy because they are obtained by numerically (Gibbs) sampling from a joint posterior having these extra parameters to accommodate additional sources of error. Gibbs sampling is a MCMC technique that simulates a posterior and, in the process, accomplishes the integration described previously (without actually integrating anything; see Gelman et al. (1995) for an introduction). The posteriors indicate that survival rates between 0.25 and 0.40 for both new and old seedlings are most likely. This result ran counter to our expectation that survival would be lower for new seedlings, but it may be explained by the fact that some seedlings may still emerge after the July censuses. Note too that the values do not greatly differ from those obtained from the simple model in figure 17.4A.

The biplots for parameter pairs show that there can be substantial correlation between some parameters (figure 17.5). Especially strong correlation occurs between the parameters λ and f. This negative correlation can be understood if we reconsider the verbal model of the foregoing paragraph. The number of new seedlings recorded in a census is the true number of seedlings (with mean value λ) times the probability that a given seedling is observed, f. Because neither quantity is directly observed, both parameters can trade off to yield a particular observed number of seedlings. The correlation evident in the biplot represents a corresponding ridge in the likelihood surface: the fit (i.e., the likelihood) for a high value of f and a low value of λ can be just as good as the fit for low f and high λ. The posterior densities in figure 17.4B,C integrate over such correlations, but it is still important to know that such correlations exist.

Few data sets have sufficient sample size and duration to estimate seedling survival rates. The majority of studies last a year or less, involve sampling from

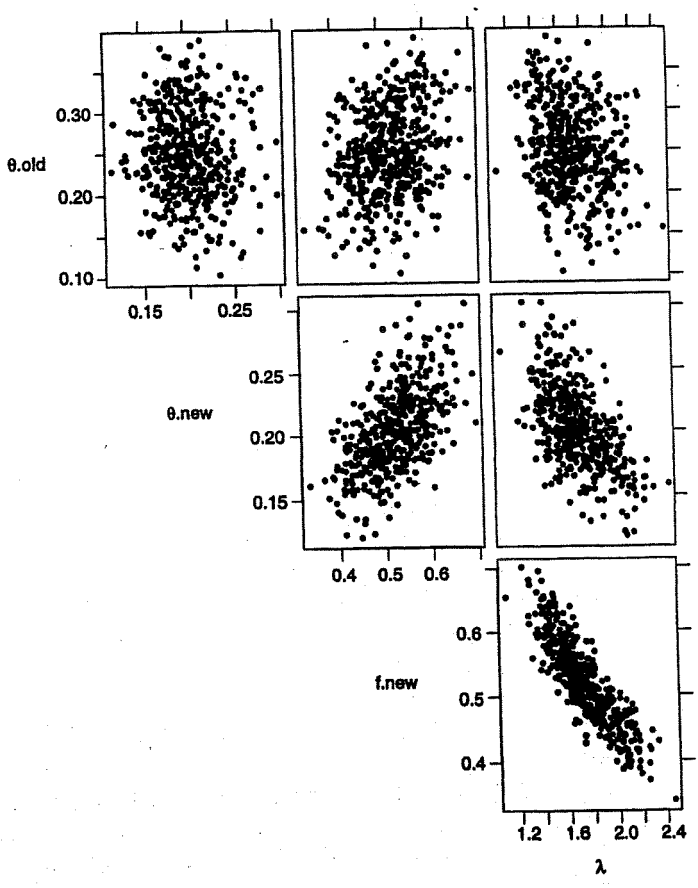

Figure 17.5 Parameter pairs for θ_O (θ.old), θ_N (θ.new), f(f.new), and λ (lambda) show tendency for correlation among some parameters.

a single stand, and examine effects on a single cohort (Clark et al. 1999b). The lack of adequate data for parameter estimation results from the intense labor required to census tagged seedlings. Comparing data sets obtained from tagged and untagged seedlings, Lavine et al. (in press) showed that this Bayesian approach provides only slightly less information than would the laborious practice of tagging all seedlings. Thus, the method makes it far easier to obtain much larger data sets.

The seedling example demonstrates the common challenge: estimation is based on a dynamic process and a particular observation can be obtained in different ways. By enumerating all of the ways by which a particular observation might arise (equation 17.8), we can accommodate far more complex problems than could be approached using a simple sampling distribution (e.g., equation 17.3). With increasing model complexity, the possibility of ill-conditioning increases, whereby the model asks for more information than the data contain. The parameter trade-offs evident here can be detected by calculating correlation coefficients between pairs of parameters or by examining biplots of the posterior (figure 17.5). Figure 17.4A indicates almost no correlation between θ_O and θ_N, but parameter correlations do arise between some parameters when the model is expanded to include other types of error (Lavine et al. in press). Although ill conditioning arose here in the context of a posterior density, the problem must be considered in any data modeling exercise, not just in Bayesian analysis.

17.5 Some Additional Ecological and Environmental Examples

17.5.1 Incorporating Different Types of Data

Population densities of bowhead whales are difficult to estimate, because the whales move from place to place and they are often underwater. The problems with counting whales motivated Raftery and Zeh (1993) to use a Bayesian analysis that accommodates counts of whales both seen and heard as they migrate past Point Barrow, Alaska. The likelihood function comes from the counts in the 1988 census. The prior takes into account the physical considerations related to locations of observers and sonar arrays, visibility, the physics of sonar location, and the knowledge of bowhead migratory behavior. The posterior that combines this information suggests that the most likely number of bowhead whales is about 7500, but that any size between 6500 and 9000 is reasonably well supported by the data. In listing the advantages of a Bayesian analysis, Raftery and Zeh (1993, pp. 166–168) state the following:

- "It enables us to use a realistic, scientifically relevant model, rather than forcing us to make artificially simple assumptions for the sake of mathematical tractability."
- "It permits us to incorporate the available external, or 'prior' information."
- "It makes elaboration and refinement of the underlying physical assumptions relatively straightforward."

- "It is very hard to develop a non-Bayesian approach that takes account of all the important sources of error."

17.5.2 External Evidence can Change the Inference

Global warming of the oceans could have a large impact on fisheries and coastlin management. Change in ocean temperature is difficult to document, because tem perature varies with depth, and data sets of the duration needed to document th temperature rise are hard to obtain. An analysis of temperature measurement along the 24.5°N transect at a depth of 1000 m in the Atlantic Ocean suggest tha a 0.1°C warming occurred during the time between two voyages completed dur ing the period 1957–1981 (Parilla et al. 1994). Lavine and Lozier (1999) reana lyzed the data using Bayesian methods that allowed them to determine the histori trend in ocean temperatures and to incorporate additional data. The Bayesia approach allowed Lavine and Lozier to consider data from other voyages tha were near the 24.5°N transect and thereby to reconstruct the temperature histor through time. The temperature history revealed by the Bayesian approach re vealed the following:

- 1957 was an unusually cold year in the historical record at 1000 m and, thus, the "trend" resulted in large part to the fortuitous timing of the first voyage,
- Temperatures of isopycnals (surfaces of constant density) are much more constant over time than temperatures at fixed depth.
- The temperature fluctuations are likely due to vertical movement of isopycnals up and down past the fixed depth rather than to a simple increasing trend.

Thus, incorporating the additional evidence brought perspective that changed ou interpretation of long-term change in ocean temperatures.

17.5.3 Parametric Empirical Bayes

Parametric empirical Bayes is a term applied to models for data that arise fro several sources of variability (see Ver Hoef 1996 for an ecological example). T understand why we say that parametric empirical Bayes is not truly Bayesian, w must discuss mixtures. To demonstrate both the utility of the method and its no Bayesian nature, we refer to a seed dispersal example of Clark et al. (1999 where both methods were used.

Ecologists have long suspected that seed shadows might have long, "fat" tail meaning that small numbers of seeds might be dispersed far from the pare plant (e.g., Portnoy and Willson 1993). Recent studies emphasizing how fat-taile kernels produce patterns of spread that differ qualitatively from traditional mode (Kot et al. 1996, Clark 1998, Lewis 1997, Clark et al. 2001) make it importa to determine whether fat-tailed dispersal is common. Traditional dispersal kernel such as Gaussian or exponential, have tails that approach zero rapidly. Unfort nately, kernels with fat tails are difficult to fit to data, and there have not bee decisive tests among competing models (i.e., those that assume fat-tailed kerne versus those that do not). Mechanistic models of dispersal are hard to apply

trees, because seeds emanate from broad and diffuse sources (tree crowns) and are released over time, as wind fields and animal dispersers vary. To determine whether seed dispersal data are best described by fat-tailed kernels, Clark et al. (1999a) used more empirical models and the method that is sometimes referred to as parametric empirical Bayes.

Both Bayesian and parametric empirical Bayes involve densities of a parameter $f(\theta)$ and a likelihood function $f(\theta|\text{data})$. In the beta-binomial example, no matter how uncertain we are about a parameter (summarized by the prior $f(\theta)$ in equation 17.4 or 17.6), the data themselves are assumed to have a binomial distribution (the likelihood $f(\text{data}|\theta)$ in equation 17.3). The prior expresses our uncertainty about θ. Uncertainty diminishes as data accumulate. The posterior becomes concentrated at the value of θ that describes the precise binomial distribution that "best" describes the data.

Instead of the binomial of equation 17.3, the dispersal example of Clark et al. (1999a) uses the likelihood for a Gaussian (normal) kernel, $f(\mathbf{r}|\theta) = N(\mathrm{r}|0, \theta^2)$, where θ is a dispersion parameter, and $\mathbf{r}$ is a vector of dispersal distances (i.e., the "data"). A conjugate prior for this Gaussian likelihood is an inverse chi-square density for the parameter θ, $f(\theta) = \text{Inv}\chi^2(u_0, p_0)$, which is shown as a dashed line in figure 17.6B. The parameters u_0 and p_0 determine the spread and shape of the density. The results are insensitive to the precise shape of this prior in this example because it is given low weight, with parameters $u_0 = 25$ and $p = 1$.

As in the beta-binomial example, application of Bayes' rule (ignoring the constant denominator in equation 17.2),

$$f(\theta \mid \mathbf{r}) \propto \text{likelihood} \times \text{prior} = f(\mathbf{r}|\theta) f(\theta; u_0, p_0)$$

yields a posterior having the same form ($\text{Inv}\chi^2$) as the prior,

$$f(\theta \mid \mathbf{r}) = \text{Inv}\chi^2\left(u_0 + \sum_{i=1}^{n} r_i^2,\ p_0 + n\right)$$

(the definition of conjugacy). From prior to posterior, only the parameter values change: the first parameter is increased by adding to it the squared observations, and the second parameter is increased by the sample size n. The combined effect is a posterior that is more "peaked" than the prior, which makes us more certain that we know which Gaussian kernel best describes the data. In figure 17.6B, the posterior lies between the prior and likelihood. It is much closer to the likelihood than the prior, because the prior is "weak." As with the beta-binomial example, our uncertainty about a parameter θ does not affect our assumption about the distribution of data (the likelihood is Gaussian).

Parametric empirical Bayes differs from true Bayes in that the variability in θ is assumed to affect the distribution of the data themselves and not just our understanding of model parameters. Suppose that the Gaussian density (defined by a value for θ) describes dispersal for a given seed released from a particular canopy location at a specific time. Because the canopy, seeds, and transport conditions all vary, there might be a different Gaussian distribution for each seed (represented by a density of values for θ). Unlike the Bayesian case, the density of

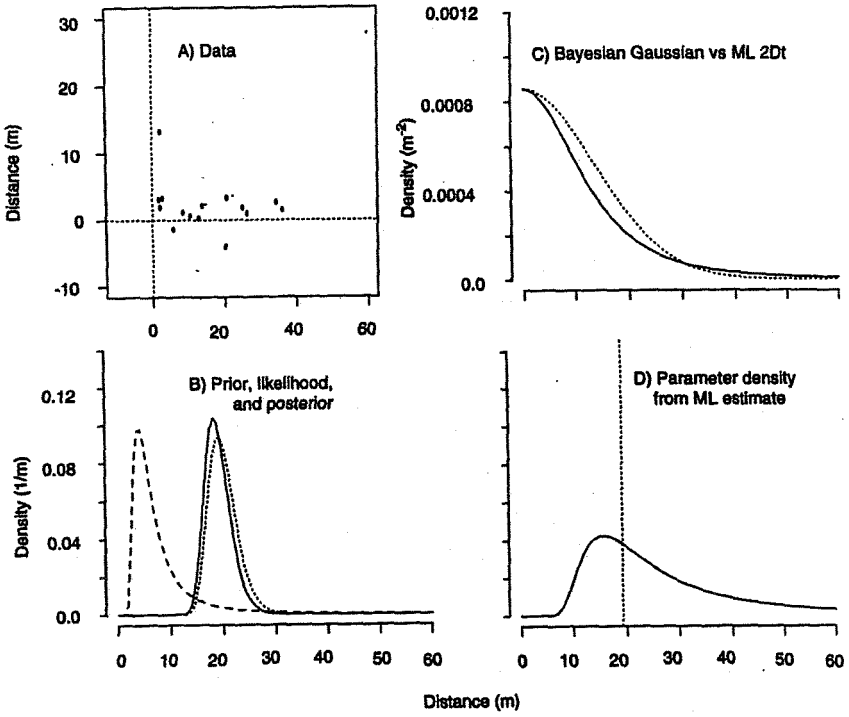

Figure 17.6 A comparison of Bayes and parametric empirical Bayes for seed dispersal data. (A) Dispersal locations of 15 *Fraxinus americana* seeds released from coordinate (0, 0). (B) Bayesian analysis. Symbolism follows figure 17.2. (C) Comparison of the ML dispersal kernel using the mixture model that assumes an Invχ^2 density of dispersal parameter values (solid line) and the Gaussian kernel corresponds to the most probable Bayesian estimate of θ. (D) The Invχ^2 density of θ for ML estimates of u and p corresponding to the 2Dt dispersal kernel in figure 17.7C.

θ-values represents variability in the data, not our range of belief about θ-values. We incorporate this variability into the likelihood itself, because it is part of the process that produces the data. Both r and θ are random variables, and the likelihood that includes both of their effects on the observations is a marginal density (mixture) obtained by integrating over the variability in θ:

$$f(r \mid u, p) = \int_0^\infty f(r \mid \theta) f(\theta; u, p) d\theta$$

This expression says that the probability (likelihood) of observing a given dispersal distance r, when r depends on a random variable θ, is their joint probability (their product), accumulated (integrated) over all possible values that θ might assume. Upon integration, the likelihood is Student's t with parameters u and p. The next step generally involves fitting the Student's t-parameters (u, p) directly to data (e.g, using maximum likelihood).

The analysis of Clark et al. (1999a) demonstrated the tendency of dispersal kernels to have long, fat tails (small values of the shape parameter p) (figure 17.6C), which implied that the density of θ values also had a long tail (figure 17.6D). The mixture model provided a test among competing models, which can be represented by different values of the parameter p. Results suggest that the dispersion parameter of the Gaussian density can be viewed as sometimes having large values that correspond to times of high winds or dispersal by animal vectors (the tail of figure 17.6D). This variability can produce a fat-tailed kernel (solid line in figure 17.6C), which, in turn, suggests the plausibility of rapid population spread (Clark 1998).

In summary, although the term *parametric empirical Bayes* sounds Bayesian, it does not involve priors and posteriors. With large sample size, the Bayesian posterior converges to a point mass centered at a single value of θ. This posterior, in turn, implies a Gaussian dispersal kernel, regardless of whether the data better support the fatter-tailed two-dimensional t-distribution 2Dt (which, in fact, they did in the analysis by Clark et al. 1999a). Parametric empirical Bayes assumes the data are distributed as 2Dt, because the variability in θ affects the actual process. The 2Dt does not converge to a Gaussian kernel with increased sample size, because the data are better described by the 2Dt. In other words, the posterior density of θ in figure 17.6B would become increasing peaked with additional data, whereas we can expect the density of θ in figure 17.6D to retain its spread.

17.5.4 Classical Significance versus Bayesian Support

Crome et al. (1996) compared a classical intervention analysis and a Bayesian approach to assess the impacts of logging on recapture rates of birds and small mammals in eastern Queensland. The likelihood of the data set is the product of two lognormal distributions, the joint probability of observing a set of differences between the logged and unlogged sites before and after logging. This is the Before-After-Control-Impact-Pairs (BACIP) model of Stewart-Oaten et al. (1986, 1992; see chapter 9). The Bayesian analysis included three priors, representing expectations about logging impact that ranged from a 25% reduction to a 25% increase in capture rates. Although the mean values differed by these percentages, the priors possessed broad overlap and thus did not represent large differences in perspective.

The classical analysis showed so few significant results as to be unhelpful—because the notion of no logging effect in this context is silly. It would be difficult to convince any ecologist that birds and mammals failed to notice that the trees had vanished. A nonsignificant result does not alter this view; rather, it points to the need for larger sample sizes to obtain a "significant" result.

By contrast, Bayesian confidence intervals help sharpen our understanding of the logging impact. Due to broadly overlapping priors, the posteriors obtained for a given species did not show large differences. The authors refer to these similarities as examples of "concensus" among those bearing different prior views. Posteriors suggest the degrees to which different species responded to the intervention and how those responses differed among habitats.

17.6 Getting the Analysis Done

Unlike most classical statistical methods, the availability of software for Bayesian analysis is limited. Most practitioners program their own models. This is most easily done using a high-level language, such as S-Plus, or specialized Bayesian software, such as BUGS [http://www.mrc-bsu.cam.ac.uk/bugs/Welcome.html] for graphical models, or BATS, for time series [http://www.stat.duke.edu/~mw/bats.html]. More complicated problems require programming in low-level languages, such as C++ or Fortran.

17.7 Conclusion

A Bayesian approach allows us to incorporate external information in the interpretation of experimental or observational data. This approach can place data in perspective of prior insights, it can provide for probability statements in situations that do submit to simple, classical frameworks, and it can minimize sample sizes and study durations necessary to arrive at experimental outcomes. This chapter scratches the surface of a broad and complex topic—one that cannot be ignored regardless of philosophical leanings, because all ecologists must judge for themselves the increasing number of Bayesian studies in the ecological literature.

Acknowledgments For their helpful comments on the manuscript, we thank B. Beckage, J. Gurevitch, J. HilleRisLambers, J. Lynch, J. McLachlan, A. Pringle, C. Saunders, S. Scheiner, and two anonymous reviewers.