1

CONCEPTUAL AND PHILOSOPHICAL CONSIDERATIONS IN ECOLOGY AND STATISTICS

A bird in the hand
is worth two in the bush ....
.... if p = 0.5

Much of contemporary ecological theory, conservation biology, and natural resource management is concerned with variation in the abundance or occurrence of species. Variation may exist over space or time and is often associated with measurable differences in environmental characteristics. Indeed, ecology has been defined as the study of spatial and temporal variation in abundance and distribution of species (Krebs, 2001). In this regard, a variety of problems in ecology require inferences about species abundance or occurrence. Examples include estimation of population size (or density), assessment of species' range and distribution, and identification of landscape-level characteristics that influence occurrence or abundance. Other examples include studies of processes that affect the dynamics of populations, metapopulations, or communities (i.e., survival, recruitment, dispersal, and interactions among species or individuals).

A common problem in the study of populations or communities is that a census (i.e., complete enumeration) of individuals is rarely attainable, given the size of the region that must be surveyed. Consequently, conclusions about populations or communities must be inferred from samples. An additional complication is that individuals exposed to sampling may not be detected. Detection problems are especially acute in surveys of animals in which detection failures can be produced by a variety of sources (e.g., differences in behavior or coloration of individuals, differences in observers' abilities, etc.). In response to these problems, clever sampling protocols and statistical methods of analysis are commonly used in surveys of animal populations. These protocols and analytical methods, whose origins may

be traced to the efforts of Dahl (1919), Lincoln (1930) and Leopold (1933) (also see discussions by Hayne (1949) and Le Cren (1965)), include capture–recapture sampling (Cormack, 1964; Jolly, 1965; Seber, 1965), distance sampling (Burnham and Anderson, 1976), and the statistical methods for analyzing such data. One of the first major syntheses of these early efforts is provided by Seber (1982). Updates of his synthesis are still regarded as general references for sampling and inference in animal populations (Seber, 1992; Schwarz and Seber, 1999). A number of books have also been published recently on the subjects of sampling biological populations and statistical modeling and inference in ecology (Borchers et al., 2002; Williams et al., 2002; Gotelli and Ellison, 2004; Buckland et al., 2004b; MacKenzie et al., 2006; Clark and Gelfand, 2006). In addition to these syntheses, countless monographs and review articles on the same topics have appeared in the primary literature.

In the face of this recent proliferation, what more can we possibly offer? We offer a comprehensive treatment of modeling strategies for many different classes of ecological inference problems, ranging from classical populations of individuals to spatially organized community systems (metacommunities). However, our primary conceptual contribution is the development of a principled approach to modeling and inference in ecological systems based on hierarchical models. We adopt a strict focus on the use of parametric inference and probability modeling, which yields a cohesive and generic approach for solving a large variety of problems in population, metapopulation, community and metacommunity systems. Hierarchical models allow an explicit and formal representation of the data into constituent models of the *observations* and of the underlying ecological or *state process*. The model of the ecological process of interest (the 'process model') describes variation (spatial, temporal, etc.) in the ecological process that is the primary object of inference. This process is manifest in a state variable, which is typically unobservable (or partially so). An example would be animal abundance or occurrence at some point in space and time. In contrast, the model of the observations (the 'observation model') contains a probabilistic description of the mechanisms that produce the observable data. In ecology, this often involves an explicit characterization of detection bias.

1.1 SCIENCE BY HIERARCHICAL MODELING

This description of models by explicit observation and state process components is now a fairly conventional approach to statistical modeling in ecology, where the term 'state-space' model is widely used. In fact, the term state-space might be more common in the prevailing ecological literature. Some recent examples include De Valpine and Hastings (2002), Buckland et al. (2004a), Jonsen et al. (2005),Viljugrein et al. (2005), Newman et al. (2006), and Dennis et al. (2006). Our view of hierarchical modeling is that it is not merely a technical approach

to model formulation, or a method of variance accounting. Rather, hierarchical modeling is a much broader conceptual framework for doing science. By focusing our thinking on conceptually and scientifically distinct components of a system, it helps clarify the nature of the inference problem in a mathematically and statistically precise way. Thus, while hierarchical models yield a cohesive treatment of many technical issues (components of variance, combining sources of data, 'scale'), they also foster the fundamental (to science) activities of 'model building' and inference. Our colleague L. Mark Berliner advocates this view in application of hierarchical models to geophysical problems. For this reason he has, on several occasions, made a distinction between *scientific modeling* – based on hierarchical models in which the underlying physical or biological process is manifest as one component of the model – and *statistical modeling*, in which this distinction is not made explicitly. The conceptual and practical distinction between these two approaches provides, in large part, the motivation for and content of this book.

Many of the practical benefits of hierarchical modeling are technical – accounting for sources of variance, different kinds of data, 'scales' of observation and many other factors. But the conceptual benefits of hierarchical modeling are both profound and profoundly subtle, and we elucidate and develop supporting arguments for this view in the following chapters of this book, using many classes of models that are widely used in ecology. For example, in Chapter 3 we discuss what is probably the simplest but most widely used class of statistical models in ecology – models for 'occurrence' which are commonly formulated in terms of simple logistic regression. We demonstrate in several subsequent chapters how relatively simple hierarchical models for occurrence provide solutions to problems of fundamental importance in ecology. In Chapter 4, we discuss models of 'occupancy and abundance.' The linkage between occupancy and abundance is expressed naturally using a hierarchical model. In Chapter 12, we show how models of animal community structure are formulated naturally as 'multi-species' models of occurrence. We provide many more examples in other chapters.

1.1.1 Example: Modeling Replicated Counts

We focus briefly on a particular example (which is described in some generality in Chapter 8) that provides an insightful illustration of the profound subtlety alluded to above. This interesting hierarchical model arises in the context of spatially-indexed sampling of a species. Suppose N_i is the (unobserved) 'local abundance' or population size of individuals on spatial sample unit $i = 1, 2, \ldots, M$. Suppose further that y_i is the *observed* count of individuals at sample unit i. A common model for such data is to assume (or assert) that individuals within the population are sampled independently of one another, in which case y_i is binomial with index

N_i and parameter p (commonly interpreted as detection probability). In this case, the N_i 'parameters' are unobserved. As such, it would be natural in many settings to impose a model on them (e.g., thinking of them as random effects). For example, we might suppose that the local abundance parameters are realizations of Poisson random variables, with parameter λ. We have in this case a two-stage hierarchical model composed of a binomial observation model and a Poisson 'process' model, which we denote in compact notation as follows:

$$f(y_i|N_i) = \text{Bin}(y_i|N_i, p)$$
$$g(N_i|\lambda) = \text{Po}(N_i|\lambda)$$

This is a degenerate hierarchical model in the sense that parameters are confounded – that is, the marginal distribution of y is Poisson with mean $p\lambda$, and unique estimates of p and λ cannot be obtained – but it seems like a sensible construction of an ecological sampling problem when sample units are spatially referenced and counts of organisms are obtained on each sample unit. (See Section 8.3 for an application of this model.)

The degeneracy of the model is easily resolved by adding a little bit more information in the form of replicate counts. That is, suppose each local population is sampled $J = 2$ times, yielding counts y_{ij} which we assume, as before, are $\text{Bin}(N_i, p)$ outcomes. In this case, the hierarchical model is not degenerate (Royle, 2004c; see Chapter 8), and we can focus on developing distinct models for the observation process (the binomial sampling model) and for the ecological process (in the form of $g(N|\theta)$).

To illustrate the benefits of this hierarchical construction, let's consider an analysis of counts of harbor seals observed in Prince William Sound, Alaska (PWS). Annual surveys of harbor seals have been conducted at 25 locations throughout PWS following the March 1989 spill of roughly 40 million liters of crude oil by the T/V Exxon Valdez. The surveyed locations include both heavily oiled sites and sites that were less affected by the spill. A primary objective of the survey is to monitor interannual changes in the size of the harbor seal population at these sites and to estimate any trend in seal abundances over time. Harbor seals are counted at each sample location on several days (usually 7 to 10) of each year using low-flying aircraft. Sampling is conducted during the molting season (August to September) when seals spend more time out of the water and are more susceptible to detection. The details of sampling and analyses of the data collected prior to 2000 are reported elsewhere (Frost et al., 1999; Boveng et al., 2003; Ver Hoef and Frost, 2003). In this example, we analyze the counts of harbor seals observed during 1990 to 2002 at 12 sample locations, half of which were heavily oiled sites.

Let y_{ijt} denote the number of harbor seals counted during the jth replicate visit to the ith sample location in year t. A variety of site- and time-specific covariates,

such as time of day (`time`), day of year (`day`), and time since low tide (`tide`) are thought to influence the detectability of harbor seals in this survey (Frost et al., 1999; Ver Hoef and Frost, 2003); therefore, in our binomial model of the observed count, $y_{ijt} \sim \text{Bin}(N_{it}, p_{ijt})$, we assume that the detection probability p_{ijt} depends on these covariates as follows

$$\begin{aligned}\text{logit}(p_{ijt}) = \alpha_{0t} &+ \alpha_1 \texttt{tide}_{ijt} + \alpha_2 \texttt{tide}^2_{ijt} + \alpha_3 \texttt{time}_{ijt} + \alpha_4 \texttt{time}^2_{ijt} \\ &+ \alpha_5 \texttt{day}_{ijt} + \alpha_6 \texttt{day}^2_{ijt}\end{aligned}$$

which allows for peaks in detectability with each covariate. We also assume that the average detectability of seals varied among years by specifying random variation in the intercept parameters as follows: $\alpha_{0t} \sim \text{N}(\mu_\alpha, \sigma^2_\alpha)$. No covariates of harbor seal abundance were observed in this survey, so our model of abundance is relatively simple, $N_{it}|\lambda_t \sim \text{Po}(\lambda_t)$, yet it allows seal abundance to vary among sample locations and years. A reduced-parameter version of this model, wherein we assume $\log \lambda_t = \log \lambda_0 + rt$, allows us to specify a trend in population abundance in terms of an intercept parameter λ_0 and a slope parameter r. Thus, we can estimate the rate of exponential growth ($r > 0$) or decline ($r < 0$) in the mean abundance of seals over time.

On fitting the hierarchical model with fixed-year effects on abundance (i.e., a model without explicit trend), we estimated the mean abundance of seals for each year of the survey (Figure 1.1). Evidently, abundance was relatively low in 1990, the year after the oil spill, but then increased markedly in 1991. However, harbor seal abundance appears to have declined steadily during the next 11 years. On fitting the hierarchical model with a trend in abundance, we estimate a rate of decline of $\hat{r} = -0.07$ per year ($SE = 0.0043$). In other words, the mean abundance of harbor seals at these locations is estimated to have decreased by 54 percent between 1991 and 2002.

The seal problem is typical of a much broader class of problems in ecology that involve modeling spatially- and temporally-indexed counts. We have demonstrated that under this common design of spatially and temporally replicate counts of individuals, a very simple hierarchical construction permits a formal rendering of the model into constituent observation and state process components. That is, the mean count is decomposed into two components, one for abundance λ and another for detection p, that have obvious interpretations in the context of the ecological sampling and inference problems. Furthermore, each of these components can be modeled separately allowing auxiliary information (in the form of covariates) to be used in a more meaningful way – i.e., in manner that is consistent with how we think such things distinctly influence *observation* and *process*.

This simple hierarchical model for the counts seems obvious and intuitive, given the context. However, the conventional approach favored by ecologists and

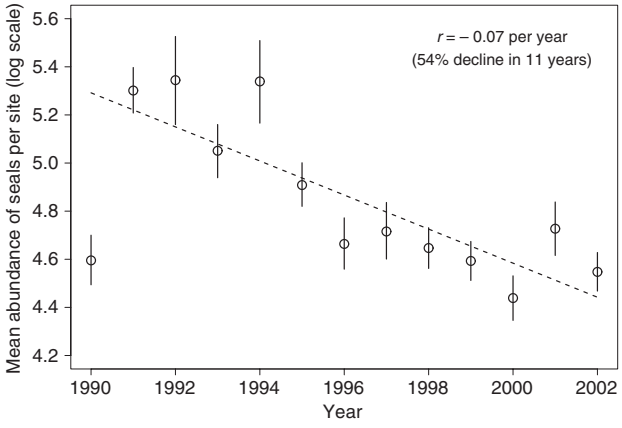

Figure 1.1. Estimates of mean seal abundance per site ($\log \lambda_t$). The circles are the estimates of the fixed year effects. The dotted line indicates the estimated trend in mean abundance between 1991 and 2002.

statisticians alike is to perform a regression (sometimes a Poisson regression) on the counts y_{ijt}, in which the site effects and replicate effects are all modeled as having the same influence on $\mathrm{E}(y)$, e.g., using a generalized linear mixed model (GLMM). In contrast, the hierarchical model is focused on describing the conditional mean of the observations, say $\mathrm{E}(y|N)$ where N is the abundance process. We provide additional conceptual and technical context of this in Section 1.4.3. What this means practically, in the seal example, is that the mean count of seals is decomposed into an abundance component and a detection component – two components that relate to different 'real' processes. Whereas, in a model for $\mathrm{E}(y)$ (e.g., using a GLMM), the variation due to these distinct processes is considered in the aggregate with no conceptual distinction between them – they both are considered equivalently in models for the mean count of the observations. Some might regard the conventional approach (GLMM) as possessing both an observation and state process model; however, the 'process' is implicit and lacks a scientific interpretation. Does a random site effect correspond to heterogeneity in the count due to detection or due to abundance? Or something else? It's difficult to say, given the ambiguity of the model. While this distinction might seem subtle or esoteric, consider the problem of making predictions of abundance. Using the hierarchical model, we can make a prediction from the abundance component of the hierarchical model which, under the model, provides predictions of abundance that are free of observation effects. Conversely, making a prediction of the marginal mean $\mathrm{E}(y)$ based on a GLMM – a statistical model – then you can predict the expected count that would be

observed under whatever sampling method was used to collect the original data. Of what biological relevance is *that*? By ignoring the ecological interpretation and the sampling context induced by viewing the replicate samples as being relevant to the same (local) population of individuals, the conventional approach yields a 'solution' to the inference problem that is more difficult to interpret. We therefore favor a more *science-based* approach to the construction of hierarchical models, as opposed to conventional approaches, which seem to be mostly statistical or descriptive in nature.

1.2 ECOLOGICAL SCALES OF ORGANIZATION

Ecological systems are fundamentally hierarchical systems in which we encounter hierarchies of organization and spatial and temporal scale. These scales often correspond to 'levels' in a hierarchical model. In this book, we address inference at a number of different ecological scales and organizational systems, including populations, metapopulations, communities, and metacommunities (Table 1.1).

Much of quantitative population ecology is concerned with populations of *individuals* – or rather characteristics of populations, such as population size. When the system state is allowed to propagate over time, we introduce population dynamic attributes such as survival and recruitment, and we seek inferences about probabilities or rates or factors that influence them. We can extend the population demographic framework up one hierarchical level to the case of multiple, spatially organized populations. Such systems are commonly referred to as *metapopulations*, the modern conceptual formulation being due to Levins (1969) and Hanski (1998). In metapopulation systems, it is not the size of the population that is relevant so much as the spatial structure in local population attributes and dynamics. Thus, population size becomes spatially indexed, $N(s)$ and one is interested in summaries of $N(s)$ such as mean, variance, and percentiles including $\Pr(N(s) > 0)$, 'occupancy' (see Chapter 3). The metapopulation analogs to individual survival and recruitment are local extinction (analogous to mortality, the complement of survival) and local colonization, which are local population or 'patch-level' attributes, describing transition events from the state $(N(s) = 0)$ to $(N(s) > 0)$, etc. An important focus of applications of metapopulation concepts is the manner in which extinction and colonization dynamics are related to such things as patch area, distance between patches, and other features of the landscape. Metapopulation extinction and colonization are closely related to individual-level dynamics. For example, for a local population to go extinct, all of the individuals have to die. Conversely, for a patch to become colonized, at least one individual must disperse into that patch from a surrounding patch, etc.

Table 1.1. Ecological scales of organization discussed in this book. Each of these systems has some notion of a 'size' parameter, N that is often the quantity of interest in static systems. In dynamic systems, there are analogs of survival and recruitment, which are usually described by extinction and colonization parameters in metapopulation and community systems.

	Static system	Dynamic system
Population of individuals	N = population size	ϕ = survival γ = recruitment
Population of populations (metapopulation)	$N(s)$ $\psi(s) = \Pr(N(s) > 0)$	$1 - \phi$ = extinction γ = colonization
Population of species (metacommunity)	N = species richness	ϕ, γ
Population of communities (metacommunity)	$N(s)$, $\psi_i(s)$	ϕ, γ

Population- and metapopulation-level concepts extend up a level of organizational complexity to systems consisting of a population (or populations) of *species*. A single population of species is a community, and spatially organized communities then are metacommunities (Leibold et al., 2004). At the community level, there is also a 'population size' parameter – usually referred to as species richness (see Chapters 6 and 12), which is the number of distinct species in the community. Species richness is a parameter of some concern in conservation biology, biogeography and, naturally, community ecology. Considerable interest exists in using patterns of species richness to inform management or conservation activities. You can examine many ecological journals and find maps of species richness (e.g., Jetz and Rahbek, 2002; Orme et al., 2005). In temporally dynamic community systems, there are also analogs of survival and recruitment that describe changes in the pool of species. These are usually referred to as local extinction (the analog of the complement of survival) and local colonization (Nichols et al., 1998a,b).

Population, metapopulation, community and metacommunity systems often share similar statistical characterizations in the sense that sampling these systems results in replicate binary data on individuals, or sites, or species, or species and sites. Regardless of the system context, we are basically interested in modeling how (or at what rate) 0s become 1s and *vice versa* – as well as how many 'all zeros' there are. Thus, the formulation of models, at least at the level of the observations, is broadly similar and not much more complicated than logistic regression. The equivalence (or perhaps duality) between models of these various systems has been exploited historically in many different contexts to solve new problems using existing methods. For example, Burnham and Overton (1979) used classical models

for estimating the size of a closed population of individuals to estimate species richness – the size of a community. Nichols and Karanth (2002) made the linkage between estimating occupancy and closed population size (in the former the zeros are observed, but not so in the latter) which we exploited in Royle et al. (2007a) to aid in the analysis of certain classes of closed population models. Barbraud et al. (2003) exploited this duality between population-level and metapopulation-level systems in the context of an 'open' system – a capture–recapture type model for dynamics of waterbird colonies.

While this technical duality between statistical formulations of models is very exciting to statisticians, the correspondence between ecological levels of organization and the levels of a hierarchical, statistical model is highly relevant to modeling and inference in ecological problems. For example, we might have parameters that are indexed by individual and then a model that governs variation (in parameter values) among individuals within a population; or (as in the seal example) parameters that are indexed by local spatially-indexed populations, combined with a model that governs variation among local populations; or parameters that are indexed by species, which are governed by a model that describes variation among species within a community. Thus, there is a conceptual correspondence between ecological scales of organization and hierarchical levels of organization in the statistical models that we consider in this book.

1.3 SAMPLING BIOLOGICAL SYSTEMS

There are two fundamental considerations common to almost all biological sampling problems regardless of scale or level of organization. These two issues provide considerable motivation for the hierarchical modeling framework that we advocate in this book.

The first issue is that of imperfect detection, or detection bias, which has a number of relevant manifestations. Individual animals in a population are detected imperfectly, species are detected imperfectly, and one might falsely conclude that a species is absent from a site that was surveyed. 'Detectability' of individuals or species is fundamental to assessing rarity and it can be closely linked to demographic processes (see Chapter 4). In contemporary animal sampling, some attention to the issue of detectability is the common thread that ties together a diverse array of sampling methods, protocols, models, and inference strategies. We feel that this issue is important in the context of hierarchical models because it is an important element of the observation process.

The second issue is that sample units are spatially indexed and they almost always represent a subset or sample of all spatial units. Spatial sampling has two consequences that concern hierarchical models. First, it forces us to confront

how to combine or aggregate information across space. We confront this issue repeatedly in this book, in which we consistently adopt a model-based approach to solving this problem. Second, spatial attribution induces a component of variation – spatial variation in the state variable – that is directly relevant to most inference problems. This spatial variance component is distinct from variance due to imperfect observation, and it has a profound influence on certain inference problems.

We can see the relevance of imperfect detection and spatial sampling by a formal description of the origin of variance in observations. Suppose that a population is composed of a spatially-indexed collection of subpopulations, of sizes $N_1, N_2, \ldots,$. Suppose we sample the populations and observe n_i individuals at location i. The 'observation variance' is the variation in n_i *conditional* on N_i. That is, the variance in the observed counts under repeated sampling of *the same* population – i.e., holding N_i fixed. This is a function of the sampling method typically, such that the better the method (more money and more time), then the closer n_i is to N_i. One common way in which this source of variation is manifest is in terms of the detection probability parameter of a binomial distribution, as described in the next subsection. The second source of variation that is relevant to some inference problems is the variation in N_i across 'replicate' populations, or spatial sample units. This source of variation is most relevant when the scope of inference is extended to populations beyond those that were sampled, i.e., when making predictions.

1.3.1 Detectability or Detection Bias

The notion of detection bias as a result of incomplete sampling is an important concept that is pervasive in almost all aspects of contemporary animal sampling, and it provides a conceptual unification of many apparently disparate sampling methods, protocols, and models (Seber, 1982; Williams et al., 2002). Detection bias has also been shown to be relevant in the sampling of sessile organisms such as plants (Kéry and Gregg, 2003; Kéry, 2004; Kéry et al., 2006).

The basic problem of imperfect detection is most commonly described using a binomial sampling argument.[1] That is, we suppose that a population composed of N individuals is sampled, and that individuals appear in the sample (i.e., are detected) independently of one another with probability π. Then the number of animals observed in the sample, n, is the outcome of a binomial random variable with sample size N and parameter π. Thus, it must be that $n \leq N$. Binomial sampling gives rise to one of the fundamental equations of animal sampling:

$$\mathrm{E}[n] = \pi \times N.$$

[1] This is a natural and concise framework for dealing with this problem, but certainly competing views are reasonable for modeling nuisance variation induced by detectability.

We refer to the general phenomenon of bias in sample quantities (e.g., $\mathrm{E}[n] \leq N$) as detection bias. Usually this parameter π is a function of other parameters. It is the probability that an individual appears in the sample at all; thus, if sampling is conducted twice, and detection is independent, then $\pi = 1 - (1-p)^2$ where p is the 'per sample' probability of detection.

One reason that 'detectability' is important is that ecological concepts and scientific hypotheses are formulated in terms of ecological state variables, e.g., abundance, or N, not in terms of the difficulty with which animals are detected. Imperfect detection *induces* a component of variation that is strictly nuisance variation and does not usually correspond to any kind of phenomenon of direct scientific or ecological relevance. This observational variance can obscure our ability to measure ecological processes of interest. Thus, formal attention to variation induced by imperfect detection yields conceptual clarity and strengthens ecological context. A practical reason to be concerned with detectability is that it provides a measure of sampling efficacy – how well animals are being counted. All things being equal, we are inclined to favor a method that counts more individuals. This could be motivated by means of a formal statistical power argument as well. For example, the power to detect differences (e.g., over time) would typically be maximized at $p = 1$.

The importance of detectability is sometimes hotly debated by ecologists and statisticians working in ecology. The approach of basically tossing the issue out the window is as prevalent as the view that detectability must be accounted for or else inferences are completely invalid. Although detectability can be addressed using models that do not explicitly account for a binomial parameter p (models of this nature are sometimes referred to by terms such as relative abundance and abundance indices, etc.), we believe that detection bias, in the form of detection probability, is an important and useful conceptual formulation of the observation process in ecological systems.

In the vast majority of situations, we consider the problem only of 'false absence', in which we fail to observe some individuals or falsely conclude a species doesn't occur. Double counting, misclassification, and 'false positives' are manifestations of imperfect detection that have received considerably less formal attention, but some recent attention to this problem can be found in genetic sampling (Lukacs and Burnham, 2005a,b; Yoshizaki, 2007) and occupancy models (Royle and Link, 2006).

1.3.2 Spatial Sampling and Spatial Variation

That sample units are spatially referenced has important consequences for inference about abundance, occurrence, and other demographic quantities. It forms the other

critical consideration motivating the adoption of a hierarchical modeling framework for inference.

We can understand the components of variance clearly by application of standard rules relating marginal and conditional variance (Casella and Berger, 2002, Chapter 4). In particular, the total variance of the observations, $n(s)$, is the sum of two parts: The *observation variance*, which is the variation in what was observed given what you would *like to have observed*, say $\text{Var}(n(s)|N(s))$ (the binomial variance component) and secondly, the *process variance* – the variance of what you would *like to have observed* across replicate spatial sample units, say $\text{Var}_s(N(s))$. Formally, the marginal variance of $n(s)$ is related to the other two quantities according to:

$$\begin{aligned}\text{Var}(n(s)) &= \text{E}\left\{\text{Var}[n(s)|N(s)]\right\} + \text{Var}\left\{\text{E}[n(s)|N(s)]\right\} \\ &= \text{E}\left\{\pi(1-\pi)N(s)\right\} + \text{Var}\left\{\pi N(s)\right\} \\ &= \pi(1-\pi)\text{E}_s[N(s)] + \pi^2\text{Var}_s[N(s)].\end{aligned}$$

The first component here is essentially a spatial average of the binomial observation variance – the expected observation variance, whereas the second part is proportional to the spatial variance. Note that as detection improves ($\pi \to 1$), the spatial variance component dominates; whereas, as detection gets worse ($\pi \to 0$), the observation (binomial) variance dominates. Thus, π induces a compromise between these two sources of variation, and we naturally prefer π close to 1. In practice, we have to estimate one or more parameters comprising π that are related to our ability to sample a population or community. As a result, this induces a third component of variance – *estimation variance*.

Thus, in most practical sampling and estimation problems, we have to be concerned with the following components of variance (and sometimes others):

- *Estimation variance* – $\text{Var}(\hat{\pi})$ – variance due to estimating parameters.
- *Observation variance* – $\text{Var}(n|N)$ – variance due to sampling a population of individuals
- *Spatial* or *process variance* – $\text{Var}(N)$ – variance due to sampling a population of populations.

This hierarchical decomposition of variance meshes conceptually with the broader theme of hierarchical models. Regardless of one's view on modeling and inference, importance of detectability, focus on relative abundance or absolute abundance, these different sources of variation exist. Hierarchical models facilitate a formal, model-based accounting for components of variation. However, it is not necessary to adopt a formal model-based solution to properly address the components of variance problem. Classical distance sampling (Buckland et al., 2001) for estimating density provides a good analysis of the components of variance. In the context of

distance sampling, it can be shown that the 'total variance' of an estimate of density (or abundance) is the sum of three pieces corresponding to observation variance, process variance, and estimation variance.

The components of variance are often very important in making predictions of the ecological process. Given an estimate $\hat{N}$ of population size for some spatially-indexed population, suppose we wish to predict the population size of a 'similar' population, down the road, in a similar habitat, and in an area of the same size. Estimation usually provides the sampling variance – the variance of the estimator conditional on N – $\mathrm{Var}(\hat{N}|N)$. Then, the same iterated conditional variance formula yields the marginal variance

$$\mathrm{Var}(\hat{N}) = \mathrm{Var}(\hat{N}|N) + \mathrm{Var}(N)$$

which is the variance that applies to a prediction at some 'new' population that was not sampled.

In practice, there are usually other components of variance to consider. For example, in many applications there is also temporal variation that affects inference (Link and Nichols, 1994). We have not addressed this explicitly here because when we are confronted with a temporally dynamic system, our model almost always contains explicit parameters that describe the sources of temporal variation in the system (e.g., survival and recruitment).

1.4 ECOLOGICAL INFERENCE PARADIGMS

Quantitative ecology is practiced by scientists with diverse views and approaches to the analysis of data from ecological systems. In our typology of approaches to the analysis of problems in quantitative ecology, we recognize three more or less distinctive philosophical or conceptual approaches. We refer to these as the observation-driven, process-driven, and the hierarchical view. These are not philosophies, such as Bayesian or frequentist, but rather are somewhat more conceptual views about the analysis of data from ecological systems. On one hand, the observation-driven view focuses on developing models for the data as functions of basic structural parameters. The process-driven view deemphasizes the observation process in favor of building elegant, often complex, models of the underlying ecological process. The hierarchical view that we advocate is a conceptual compromise between these other two views.

1.4.1 The Observation-driven View

In quantitative ecology, there has been a strong focus historically on quantities that are relevant mostly to the sampling apparatus, such as the fraction of the population detected during a sample – quantities that are largely unrelated to any ecological process. This direct focus on modeling and estimation of the observation process characterizes what we call the observation-driven view, and this is the classical view that seems to dominate the segment of statistical ecology concerned with sampling and estimation in biological populations. The key feature of the observation-driven analysis is the prominence of the observation model which, in practical situations, is manifest as an elaborate model of detection probability under a binomial sampling model. Often, there is no attention to the ecological process of interest, except as it relates to an 'adjustment' of sample data based on some estimate of detectability.

As a technical matter, in the observation-driven view, methods are characterized by partial likelihood or conditional estimators of quantities that are relevant. In this approach, the biological quantities of interest are usually formally removed from the model (usually by 'conditioning'), so that attention can be focused on estimating nuisance parameters. As a conceptual example, it is common to treat N as a nuisance parameter by removing it from the likelihood by conditioning on n. This leads to so-called 'conditional likelihood' methods (which we discuss in Chapters 5 and 6) in which the estimator of N is of the form

$$\hat{N} = \frac{n}{\hat{\pi}}.$$

Of course, there is not really anything fundamentally wrong with this estimator of N *per se*. However, as a philosophical matter, removal of the object of inference from the likelihood so that nuisance parameters can be estimated is at odds with statistical norms. Moreover, this approach is philosophically unappealing. More importantly, the basic strategy is ad hoc and becomes self-defeating when the objectives are extended to include not just estimation of N, but developing models, especially predictive models, of the parameters that were removed from the likelihood.

This classical, observation-driven view has been fostered to a large extent by the advent of software that performs certain limited types of analyses, converting sample data to estimates of N or other things. Indeed, it seems that many studies or analyses are motivated largely by the procedures available in existing software. To achieve broader objectives, the practitioner must plug results into a second-stage procedure. The *procedures* are hierarchical (or perhaps 'multi-step' would be a better descriptor), but the *model* is not hierarchical. It is this multi-step approach that gives rise to the phrase 'statistics on statistics', which is almost always an ad hoc attempt at hierarchical modeling. By that, we mean there is usually an

improper accounting for variance in the second-stage analysis since the first-stage estimates are treated as data in the second-stage analysis, frequently ignoring error due to estimation.

A typical 'hierarchical procedure' might resemble the following: consider a metacommunity system wherein surveys are carried out at a number of landscape patches. Suppose the goal is to model sources of variation in species richness among patches, e.g., how does the composition of the landscape, patchiness, urbanization, etc. affect species richness? One way in which this has been done (repeatedly) in the literature is to obtain $\hat{N}$ for each patch using some software package, and then conduct a further analysis of the $\hat{N}$'s. For example, fit a regression model relating $\hat{N}$ to covariates. There are a number of problems with this approach – most having to do with improper accounting for uncertainty – components of variance – and the extensibility of the procedure. On the other hand, using hierarchical models, we can achieve a coherent framework for inference – one that properly accounts for sources of variance, and one that is flexible and extensible.

1.4.2 The Process-Driven View

A second philosophy to modeling and inference in ecology is what we refer to as the process-driven view. This is the conceptual opposite of the observation-driven view in the sense that the modeling, estimation, and inference are typically focused on the process component of the model, to the exclusion of any consideration of the observation process. Often, the process-driven approach to a problem involves complex but elegant mathematical models that have been developed based solely on conceptual considerations.

The basic construct is pervasive in models of invasive species and disease systems, which usually involve very complex models of system dynamics, such as invasive spread and density-dependent mechanisms. Another area dominated by the process-driven view are ecosystem models, which are largely 'organizational charts' of trophic levels. A final example can be found in many community ecology analyses, in which counts of species (and other summaries) are viewed as direct measurements of the state process. There was a period in the 1970s and 1980s when considerable research went into 'quantifying' biodiversity, as if species at some point in space and time could simply be tabulated and individuals enumerated. This approach is still the dominant approach in biodiversity research. Another type of problem is the construction of complex models of population dynamics, (e.g., based on Leslie matrix models) that use estimates of various parameters (often of varying quality and relevance) as if they were the truth. This is sometimes justified by an explanation such as 'estimates were obtained from the literature', as if the fact that they had to be estimated from data is irrelevant.

The process-driven view seems to be dominated by mathematicians and engineers, who have a lot of skill with devising models of things but who are less concerned with how observations are obtained in the field. Process-driven applications are characterized by WYSIWYG analyses – analysis of the data as if "what you see is what you get." This can't possibly be the case in practical field situations – nor even in so-called 'field experiments.' In a sense the process-driven view is the natural manner in which scientists want to approach the study of biological or physical processes. The problem is that there can be important effects of the observation process on inference, and this must be considered in the analysis.

1.4.3 The Philosophical Middle Ground: The Hierarchical View

There is a conceptual middle-ground to the observation-driven and process-driven views that is based on hierarchical models. The hierarchical view shares elements with both. Like the observation-driven approach, hierarchical models admit formal consideration of the observation process. Like the process-driven approach, hierarchical models also possess a component model representing the ecological process that is the primary focus of the scientific problem.

We adopt the conceptual definition of a hierarchical model from Berliner (1996). Hierarchical models are composed of a model for the observations – the 'data' – that is conditional on some underlying ecological process that is the focus of inference. A second model – the process model – describes the dynamics of the ecological process. Finally, we may require additional structure to relate the parameters of the observation or process model, in some cases. As a technical, statistical matter, inference is based on the joint distribution, which is the product of submodels. That is (Berliner, 1996)

$$[\text{data}|\text{process}, \text{parameters}][\text{process}|\text{parameters}][\text{parameters}]$$

The first component, [data|process, parameters], describes the observation process – the conditional distribution of the data given the state process and parameters. The second component, [process|parameters], describes the state process as we are able to characterize it based on our understanding of the system under study – without regard to 'data' or sampling considerations. Finally, we sometimes express explicit model assumptions about parameters of these two models, and that is the term [parameters].

The existence of the process model is central to the hierarchical modeling view. We recognize two basic types of hierarchical models. First is the hierarchical model that contains an *explicit* process model, which describes realizations of an actual ecological process (e.g., abundance or occurrence) in terms of explicit biological

quantities. The second type is a hierarchical model containing an *implicit* process model. The implicit process model is commonly represented by a collection of random effects that are often spatially and or temporally indexed. Usually the implicit process model serves as a surrogate for a real ecological process, but one that is difficult to characterize or poorly informed by the data (or not at all). In this book, we typically deal with hierarchical models having an explicit process model, and we prefer such models when it is possible to characterize them for a given problem.

The distinction was evident in the seal example described in Section 1.1.1. We formulated a model in which the process model described spatial and temporal variation in N, the population size of seals. This spatial model for N represents an explicit process model. We noted in Section 1.1.1 that it is more common in statistical analyses of such data to adopt an analysis based on generalized linear mixed models (GLMMs) or similar methods containing a collection of random effects describing unstructured spatial or temporal variation in the mean of the observations, $\mathrm{E}(y)$. We would call this 'random effects' model an implicit process model because the 'process' (the random effects) lack an explicit biological interpretation.

The correspondence (or lack thereof) between implicit and explicit representations comes down to *interpretation* of the implied marginal moment structure of the observations. While the marginal structure may be statistically equivalent (but not necessarily), the interpretation of the parameters is distinctly different between the two approaches. For example, the model for replicate counts (used in the seal example) in which N is Poisson, implies a certain within-group or intra-class correlation (Royle, 2004a). In particular, the covariance between replicate counts of the same local population can be shown to be

$$\mathrm{Cov}(y_{i1}, y_{i2}) = p^2\lambda,$$

where p is the probability of detection, a component of the observation model, and λ is the mean of N across sites – a component of the process model. In adopting an implicit process model formulation, we would just introduce a variance component into the model to account for this correlation, say $\sigma_{12} = p^2\lambda$ and estimate σ_{12}. This is essentially what a GLMM does. Conversely, under the explicit hierarchical model, we retain the distinction between the observation and process models and develop distinct models for both p and λ (as we did in the seal example) – both of which have precise interpretations in the context of the sampling and inference problem at hand. In contrast, the marginal covariance parameter σ_{12} is just that regardless of whether we are sampling seal populations in Alaska, taking repeated measurements of pigs in Iowa, or counting dental patients in Ann Arbor. We are not typically concerned about the marginal structure of the observations if our

hierarchical model has a coherent formulation that is meaningful in the context of the scientific problem. Where possible, we prefer formulations of hierarchical models in terms of a real ecological process.

1.5 THE ROLE OF PROBABILITY AND STATISTICS

Statistics is a discipline concerned with learning from data. Given a set of observations, we wish to make evidentiary conclusions about some phenomenon or process. The theory of statistics provides the conceptual and methodological framework for doing this. Probability is instrumental to this endeavor in two important respects. First, probability models provide the basis for describing variation in things we can and cannot observe (i.e., variation in observable data and in ecological processes). Second, we use probability to express uncertainty in our conclusions, such as inferences about model parameters and latent variables (things we wish we could observe) of interest.

1.5.1 Probability as the Basis for Modeling Ecological Processes

A model is an abstraction of a phenomenon, process or system. Often, the model takes the form of a mathematical expression, or a probability distribution, but could as well be a graphical description, a pie chart, even a verbal description. Regardless of form, the model represents an abstraction from which we hope to learn about a system or that might elucidate some component of that system. Ecological science is largely concerned with developing models of biological systems and the observation of those systems. In this book, we exclusively adopt probability models as the basis for describing both the observation process, the means by which data are obtained, and of the underlying ecological process.

Classical applied statistics, as taught at most universities, is largely focused on the mechanics of converting data to estimates and p-values. This approach is very procedure-oriented, and there is little attention to broader concepts of model construction – i.e., the use of probability models as the basis for modeling ecological processes. In this book, we try to emphasize the construction of probability models across a broad spectrum of ecological problems.

1.5.2 Probability as the Basis for Inference

Formal inference is the other major use of probability in statistics. By inference, we mean confronting models with data in order to estimate parameters (i.e., fit

the model), to carry out some kind of inference (hypothesis test, model selection, model evaluation), to make predictions (Clark et al., 2001), as is done commonly in the construction of species distribution maps (Guisan and Zimmermann, 2000), and even to provide guidance on how to sample the underlying process in an efficient manner (Wikle and Royle, 2005).

While virtually everyone agrees on the use of probability as a tool for building models, there is considerable, sometimes excited, debate over how probability should be used in the conduct of inference. There are at least two schools of thought on this matter. The classical view, which is based on the frequentist idea of a hypothetical collection of repeated samples or experiments, uses probability in many different ways, but never to make direct probability statements about model parameters. In contrast, the Bayesian approach uses probability to make direct probability statements about all unknown quantities in a model.

1.6 STATISTICAL INFERENCE PARADIGMS: BAYESIAN AND FREQUENTIST

The term 'hierarchical modeling' has become almost synonymous with 'Bayesian' in recent years, or at least a good deal of papers on 'Bayesian analysis' also have 'hierarchical model' in the title or key words (e.g. Wikle, 2003; Hooten et al., 2003; Clark, 2003). While hierarchical models are often conveniently analyzed by Bayesian methods, the mode of analysis and inference really stands independent of the formulation of the model. As we have tried to stress, hierarchical modeling is mostly concerned with model *construction*. As such, we freely adopt Bayesian and non-Bayesian analyses of hierarchical models. Although we believe that the analysis of hierarchical models can sometimes be accomplished effectively using classical, non-Bayesian methods, we also believe that Bayesian analysis of hierarchical models is more natural and has some conceptual and practical advantages, which we will describe subsequently and throughout the rest of this book.

The main practical distinction between Bayesian and non-Bayesian treatments of hierarchical models comes down to how latent variables (random effects) are treated. Bayesians put prior distributions on all unknown quantities and use basic probability calculus in conjunction with simulation methods (known as Markov chain Monte Carlo (MCMC), see Link et al., 2002) to characterize the posterior distribution of parameters and random effects by Monte Carlo simulation. The non-Bayesian removes the random effects from the model by integration. This approach works reasonably well in a lot of problems. It does not work as well when the structure of the latent variable model or dependence among latent variables is complex, and it does not always provide a cohesive framework for inference about random effects.

1.6.1 Dueling Caricatures of Bayesianism

We find that much is lost in the transition of classically trained ecologists to the Bayesian paradigm of inference. Or perhaps much confusion is injected into the debate. Ecologists, when first confronted with Bayesianism from an *opposing* view, are often taught that Bayesian analysis is impractical because it is hard to do the calculations, or because model selection is difficult, or because your results depend on the prior. Since you don't usually know anything about the priors, then your inference is not objective. Therefore, Bayesianism is not science. Moreover, priors are not invariant to transformation of the parameters. Therefore, what looks uninformative for θ could very well be informative for $g(\theta)$. Such naive arguments are used by some to explain Bayesianism while, at the same time, debunking it.

Conversely, when confronted with Bayesianism from a *proponent* of Bayesian methods, the ecologist often gets an equal-but-opposite parody in which the profound difference in these two views is reduced to a caricature having to do with Bayesian confidence intervals having a more desirable interpretation, something like this:

> Whereas a frequentist will say "the probability that the interval $[a, b]$ contains the fixed, but unknown, value of θ is 0.95," the Bayesian will say "the probability that θ is in the interval $[a, b]$ is 0.95."

Aside from not being very insightful, it leaves most practical scientists wondering "So what?" Our experience is that ecological scientists do not really care about how confidence intervals should be interpreted. The other part of the pro-Bayesian parody has to do with how great Bayesian methods are for using *prior information* (because, of course, they have prior distributions and the frequentist doesn't!). Unfortunately, prior information isn't a problem that most ecologists have, or even want to have for that matter – or perhaps want to admit. While it is true that the Bayesian framework admits this generalization of the inference problem, it has not proved to be terribly useful in ecology (but see McCarthy and Masters, 2005 for an exception).

In the context of these opposing parodies of Bayesianism, it is useful to reflect on the cynical (but, we believe, accurate) view of Lindley (1986), who remarked "What most statisticians have is a parody of the Bayesian argument, a simplistic view that just adds a woolly prior to the sampling-theory paraphernalia. They look at the parody, see how absurd it is, and thus dismiss the coherent approach as well." While we basically agree with this view, we also emphasize that this book is not about Bayesianism, although we tend to embrace the basic tenets of Bayesianism for analysis and inference in hierarchical models.

1.6.2 Our Bayesian Parody

It would be impossible to elaborate here in a meaningful way on the distinction between classical statistics based on frequentist inference and Bayesian analysis. We don't really want to be guilty of caricaturizing Bayesianism. Several recent books address Bayesian analysis and even Bayesian analysis in Ecology (Clark and Gelfand, 2006; McCarthy, 2007). On the other hand, we probably need to address it briefly, as follows.

In classical statistics, one does not condition on the observed data but rather entertains the notion of replicate realizations (i.e., hypothetical data) and evaluates properties of estimators by averaging over these unobserved things. Conversely, in the Bayesian paradigm, the Bayesian conditions on data, since its the only thing known for certain. So the frequentist will evaluate some procedure, say an estimator, $\hat{\theta}$, that is a function of x, say $\hat{\theta}(x)$, by averaging over realizations of x. The nature of $\hat{\theta}$ becomes somewhat important to the frequentist way of life – there are dozens of rules and procedures for cooking up various flavors of $\hat{\theta}$. On the other hand, the Bayesian will fix x, and base inference on the conditional probability distribution of θ given x, which is called the posterior distribution of θ. For this reason, Bayes is, conceptually, completely objective – inference is *always* based on the posterior distribution. But, therein lies also the conflict. To compute the posterior distribution, the Bayesian has to prescribe a prior distribution for θ, and this is a model *choice*. Fortunately, in practice, this is usually not so difficult to do in a reasonably objective fashion. As such, we view this as a minor cost for being able to exploit probability calculus to yield a coherent framework for modeling and inference in any situation.

There are good philosophical and practical reasons to adopt the Bayesian framework for inference, in general. For example, the Bayesian paradigm is ideal for inference about latent variables and functions of latent variables. Classically, this problem is attacked using a partially Bayesian idea – calculation of the 'conditional posterior distribution' upon which the Best Unbiased Predictor (see Section 2.6) is based. This resolves the main problem (obtaining the point estimate) but creates an additional problem (characterizing uncertainty). A Bayesian formulation of the problem produces a coherent solution to both components of the inference problem. While this point might seem fairly minor, the Bayesian implementation even for very complex models (e.g., non-normal, nonlinear) is conceptually and, using modern methods of computation, practically accessible. An important benefit of Bayesian inference is its relevance to small sample situations (or rather, finite samples). We think that it is under-appreciated by ecologists that frequentist inference is asymptotic, or at least the practical relevance of asymptotic procedures is not often considered. Finally, an important benefit of a Bayesian approach to the analysis

of hierarchical models is a transparent accounting for all sources of variation in an estimate or a prediction.

1.7 PARAMETRIC INFERENCE

This book is primarily about model construction, using probability as a basis for modeling and the use of probability as the basis for formal inference. As we have noted, we readily adopt conventional Bayesian and non-Bayesian methods for the analysis of hierarchical models. While the two inference frameworks have important (and profound) technical and conceptual differences, they are unified under the rubric of parametric inference. Thus parametric or 'model-based' inference constitutes one of the overarching themes of this book. That is, our inference is conditional on a prescribed model. This yields a flexible, cohesive framework for inference, and generic procedures having many desirable properties (see Chapter 2).

At many universities, ecologists are taught principles of 'survey sampling' in their statistical curricula. In classical survey sampling or design-based sampling, samples are drawn according to some probabilistic rule, and the properties of estimators are derived from that rule by which the sample is drawn. The actual 'data-generating' process is completely irrelevant. No matter how pathological the data are, the method of sampling *induces* certain desirable properties on parameter estimators. Unfortunately, these desirable properties don't necessarily apply to your data set, obtained under your design. Rather, for example, estimators are unbiased in the sense that, when averaged over all possible samples, the estimator will equal the target population parameter.

A common rationale for reliance on design-based ideas (or at least motivating the use of such methods) is that it will then be 'robust' to parametric model assumptions. However, there is nothing about design-based sampling theory that suggests robustness to arbitrary, unstated model assumptions. Little (2004) gave a good example that turns out to be highly relevant to several procedures used in animal ecology (see Chapter 6). The Horvitz–Thompson estimator (HTE) (Thompson, 2002, Ch. 6) is a widely used procedure for unequal probability sampling. Given a set of observations y_i where observation y_i has probability ψ_i of appearing in the sample, consider estimating the total of a population of N units

$$T = \sum_{i=1}^{N} y_i.$$

Then, the HTE of T based on a sample of size n is $T_{ht} = \sum_{i=1}^{n} y_i/\psi_i$. The HTE has some appealing properties when evaluated from a design-based perspective. Little

(2004) notes that the HTE does have a model-based justification, the estimator arises under the weighted regression model

$$y_i = \beta\psi_i + \psi_i\epsilon_i,$$

where ψ_i is the sample inclusion probability and ϵ_i are iid errors with mean 0 and variance σ^2. This leads to $\hat{\beta} = T_{ht}/n$ where n is the sample size. Little concluded: "This analysis suggests that the HTE is likely to be a good estimator when [this model] is a good description of the population, and may be inefficient when it is not." In other words, the HTE is a good (perhaps great?) procedure when inclusion probabilities are relevant to the data-generating mechanism. However, if the inclusion probabilities are not related to the data-generating mechanism, then the HTE may not be that good at all, Little gave an example, the "elephant example," in which they were basically unrelated. This correspondence between design-based and model-based views is not as widely appreciated as we believe it should be.

In classical statistics we are taught elements of finite population sampling. And so, naturally, these notions pervade contemporary statistical ecology as well. But models are more fundamentally relevant to many, if not most, inference problems in ecology. There are always important uncontrollable, unaccounted for sources of variation or practical matters associated with the conduct of field studies that force us to use models to accommodate things that simply could not have been anticipated or controlled for *a priori*. We do have a need for elements of sampling, but often we rely on these things because they give us faith in our models or protect us from severe departures from our model.

1.7.1 Parametric Inference and The Nature of Assumptions

In ecology, we often cannot enumerate sample frames, randomly sample, observe the state variable, or sufficiently control our environment. The consequence of this is that properties of estimators and inference procedures must be, to a large extent, inherited from *parametric inference theory*. What this means, practically, is that we pick a model, we fit that model, and we can compute variances, posterior distributions, p-values, whatever we want under the assumption that the prescribed model is the data-generating model (some would say 'correct' or 'true' model). The unifying conceptual thread of Bayesian and classical frequentist inference is that they are both largely frameworks for the conduct of parametric inference, in which we specify a model, then make an inference that is conditional on the model.

The issue of model-based parametric inference is much bigger than just the distinction between it and other paradigms (i.e., design-based inference). We believe

that much of the antagonism between Bayesians and frequentists comes down to anxiety over parametric inference theory vs. these 'other' concepts or philosophies. Biologists favor design-based sampling, at least in part, because the idea of a well-defined set of procedures that can be applied to any problem is reassuring to many. Moreover, one doesn't really have to understand these procedures to apply them. On the other hand, using parametric inference, the specter of 'sensitivity to model' can always be used as a criticism of any analysis. Results are conditional on the model being correctly specified. We accept this as coming with the territory of a generic, flexible, and cohesive framework for inference.

One of the most famous quotes in statistics is that by G.E.P. Box, who remarked: "all models are wrong but some are useful..." which must be the most cited statistical quote of all time[2]. There is a school of thought contending that one must have the right model, or at least the best model achievable, or else you can't do inference. But really these paradigms – parametric inference and model selection – are mutually exclusive to a large extent. You can't carry out a big model selection procedure and then rely on parametric inference in any meaningful way. Consistent with this "all models are wrong ..." sentiment, we view the objective of statistics, at least in part, as being the development of useful models. These (i.e., useful models) exist independent of the quality of data, even of the reasonableness of assumptions since, in our view, the reasonableness of any assumption is completely subjective and debatable. We will further explore model selection and assessment in Chapter 2.

1.7.2 The Hierarchical Rube Goldberg Device

Often, researchers compensate for this reliance on parametric assumptions by building huge elaborate models that obscure what is going on. Now statisticians routinely engage in the development of complex models under the guise of 'hierarchical modeling,' seemingly for the sole purpose of introducing complexity – hierarchical modeling for the sake of hierarchical modeling. Such models are big and conceptually beautiful, and the motives are usually pragmatic, but they can be difficult to understand 'statistically.' It is not easy to criticize such efforts because the mere effort of criticism would be a research project unto itself. The end result is to render a problem unassailable, unrepeatable, unfalsifiable, and beyond comprehension.

In this regard, Dennis Lindley remarked in his book "Understanding Uncertainty" (Lindley, 2006):

[2]Besides "Lies, damned lies, and statistics" – by Disraeli.

> There are people who rejoice in the complicated saying, quite correctly, that the real world is complicated and that it is unreasonable to treat it as if it was simple. They enjoy the involved because it is so hard for anyone to demonstrate that what they are saying can be wrong, whereas in a simple argument, fallacies are more easily exposed.

The point being, in the words of our colleague (Link, 2003), "Easily assailable but clearly articulated assumptions ought always to be preferable." Simplicity is a virtue. Not necessarily procedural simplicity, but conceptual simplicity – and clearly assailable assumptions.

1.8 SUMMARY

Hierarchical modeling is something of a growth industry in statistics (and ecology), partially due to the advent of practical Bayesian methods, which we believe foster the adoption of hierarchical models, and partially due to the conceptual advantage of the hierarchical formulation of models in scientific disciplines. Our focus on hierarchical modeling in this book is not about the choice of inference method (Bayesian or frequentist); instead, we focus on providing pragmatic, but principled, solutions to inference problems encountered in ecological studies. Hierarchical models represent a compromise – the conceptual middle ground – between two distinctive approaches to the conduct of ecological science, approaches that we have referred to as the observation and process-driven views. As such, hierarchical models contain explicit representations of both the observation process and also the ecological process of scientific relevance.

Many ecological problems yield naturally to a hierarchical construction because it allows for the formulation of models in terms of the ecological process of interest, the thing that is interesting to most ecologists, while at the same time dealing formally with imperfect observation of the state process. For example, many problems that we encounter have a natural hierarchical structure that is induced by ecological scale: individuals within populations, populations of populations, species within communities, communities of communities. This structure often induces or coincides with components of a hierarchical model, as we will see in subsequent chapters of this book. Secondly, there is almost always a natural and distinct observation component of the model that represents our inability to count individuals. Variation induced in the data by imperfect observation is hardly ever the focus of scientific inquiry, but this source of variation almost always has a profound influence on inference, so must formally be accounted for by statistical procedures. Hierarchical models facilitate and formalize the manner in which the observation process is handled and integrated with the process model. While this clear technical and conceptual distinction between sources of variation in data is a nice feature of hierarchical

models, we believe that it is secondary to the main benefit of hierarchical modeling – that it fosters an emphasis on model construction and elucidates the fundamental nature of inference, whether it be about processes, parameters, or predictions. Thus, in our view, hierarchical modeling fosters *scientific modeling*.

2

ESSENTIALS OF STATISTICAL INFERENCE

In the previous chapter we described our philosophy related to statistical inference. Namely, we embrace a model-based view of inference that focuses on the construction of abstract, but hopefully realistic and useful, statistical models of things we can and cannot observe. These models always contain stochastic components that express one's assumptions about the variation in the observed data and in any latent (unobserved) parameters that may be part of the model. However, statistical models also may contain deterministic or structural components that are usually specified in terms of parameters related to some ecological process or theory. Often, one is interested in estimating these parameters given the available data.

We have not yet described how one uses statistical models to estimate parameters, to conduct an inference (hypothesis test, model selection, model evaluation), or to make predictions. These subjects are the focus of this chapter. Inference, by definition, is an inductive process where one attempts to make general conclusions from a collection of specific observations (data). Statistical theory provides the conceptual and methodological framework for expressing uncertainty in these conclusions. This framework allows an analyst to quantify his/her conclusions or beliefs probabilistically, given the evidence in the data.

Statistical theory provides two paradigms for conducting model-based inference: the classical (frequentist) approach and the Bayesian view. As noted in the previous chapter, we find both approaches to be useful and do not dwell here on the profound, foundational issues that distinguish the two modes of inference. Our intent in this chapter is to describe the basis of both approaches and to illustrate their application using inference problems that are likely to be familiar to many ecologists.

We do not attempt to present an exhaustive coverage of the subject of statistical inference. For that, many excellent texts, such as Casella and Berger (2002), are available. Instead, we provide what we regard as essentials of statistical inference in a manner that we hope is accessible to many ecologists (i.e., without using too much mathematics). The topics covered in this chapter are used throughout the book, and it is our belief that they will be useful to anyone engaged in scientific research.

2.1 PRELIMINARIES

2.1.1 Statistical Concepts

Before we can begin a description of model-based inference, we need a foundation for specifying statistical models. That foundation begins with the idea of recognizing that any observation may be viewed as a realization (or outcome) of a stochastic process. In other words, chance plays a part in what we observe.

Perhaps the simplest example is that of a binary observation, which takes one of two mutually exclusive values. For example, we might observe the death or survival of an animal exposed to a potentially lethal toxicant in an experimental setting. Similarly, we might observe the outcomes, 'mated' or 'did not mate', in a study of reproductive behavior. Other binary outcomes common in surveys of animal abundance or occurrence are 'present/absent' and 'detected/not detected.' In all of these examples, there is an element of chance in the observed outcome, and we need a mechanism for specifying this source of uncertainty.

Statistical theory introduces the idea of a *random variable* to represent the role of chance. For example, let Y denote a random variable for a binary outcome, such as death or survival. We might codify a particular outcome, say y, using $y = 0$ for death and $y = 1$ for survival. Notice that we have used uppercase to denote the random variable Y and lowercase to denote an observed value y of that random variable. This notation is a standard practice in the field of statistics. The random variable Y is a theoretical construct, whereas y is real data.

A fully-specified statistical model provides a precise, unambiguous description of its random variables. By that, we mean that the model specifies the probability (or probability density) for *every* observable value of a random variable, i.e., for every possible outcome. Let's consider a model of the binary random variable Y as an example. Suppose $p = \Pr(Y = 1)$ denotes the probability of success (e.g., survival) in a single trial or unit of observation. Since Y is binary, we need only specify $\Pr(Y = 0)$ to complete the model. A fully-specified model requires

$$\Pr(Y = 1) + \Pr(Y = 0) = 1.$$

Given our definition of p, this equation implies $\Pr(Y = 0) = 1 - p$. Thus, we can express a statistical model of the observable outcomes ($y = 0$ or $y = 1$) succinctly as follows:

$$\Pr(Y = y) = p^y(1 - p)^{1-y}, \tag{2.1.1}$$

where $p \in [0, 1]$ is a formal parameter of the model.

Equation (2.1.1) is an example of a special kind of function in statistics, a *probability mass function* (pmf), which is used to express the probability distribution

of a discrete-valued random variable. In fact, Eq. (2.1.1) is the pmf of a Bernoulli distributed random variable. A conventional notation for pmfs is $f(y|\theta)$, which is intended to indicate that the probability of an observed value y depends on the parameter(s) θ used to specify the distribution of the random variable Y. Thus, using our example of a binary random variable, we would say that

$$f(y|p) = p^y(1-p)^{1-y} \tag{2.1.2}$$

denotes the pmf for an observed value of the random variable Y, which has a Bernoulli distribution with parameter p. As summaries of distributions, pmfs honor two important restrictions:

$$f(y|\theta) \geq 0 \tag{2.1.3}$$

$$\sum_y f(y|\theta) = 1, \tag{2.1.4}$$

where the summation is taken over every observable value of the random variable Y.

The probability distribution of a continuous random variable, which includes an infinite set of observable values, is expressed using a *probability density function* (pdf). An example of a continuous random variable might be body weight, which is defined on the set of all positive real numbers. The notation used to indicate pdfs is identical to that used for pmfs; thus, $f(y|\theta)$ denotes the pdf of the continuous random variable Y. Of course, pdfs honor a similar set of restrictions:

$$f(y|\theta) \geq 0 \tag{2.1.5}$$

$$\int_{-\infty}^{\infty} f(y|\theta)\,\mathrm{d}y = 1. \tag{2.1.6}$$

In practice, the above integral need only be evaluated over the range of observable values (or support) of the random variable Y because $f(y|\theta)$ is zero elsewhere (by definition). We describe a variety of pdfs in the next section.

Apart from their role in specifying models, pmfs and pdfs can be used to compute important summaries of a random variable, such as its mean or variance. For example, the mean or *expected value* of a discrete random variable Y with pmf $f(y|\theta)$ is

$$\mathrm{E}(Y) = \sum_y y f(y|\theta).$$

Similarly, the mean of a continuous random variable is defined as

$$\mathrm{E}(Y) = \int_{-\infty}^{\infty} y f(y|\theta)\,\mathrm{d}y.$$

The expectation operator $\mathrm{E}(\cdot)$ is actually defined quite generally and applies to *functions of random variables*. Therefore, given a function $g(Y)$, its expectation is computed as

$$\mathrm{E}[g(Y)] = \sum_y g(y) f(y|\theta)$$

if Y is discrete-valued and as

$$\mathrm{E}[g(Y)] = \int_{-\infty}^{\infty} g(y) f(y|\theta)\, \mathrm{d}y$$

if Y is a continuous random variable. One function of particular interest defines the variance,

$$\mathrm{Var}(Y) = \mathrm{E}[(Y - \mathrm{E}(Y))^2].$$

In other words, the variance is really just a particular kind of expectation. These formulae may seem imposing, but we will see in the next section that the means and variances associated with some common distributions can be expressed in simpler forms. We should not forget, however, that these simpler expressions are actually derived from the general definitions provided above.

2.1.2 Common Distributions and Notation

The construction of fully-specified statistical models requires a working knowledge of some common distributions. In Tables 2.1 and 2.2 we provide a list of discrete and continuous distributions that will be used throughout the book.

For each distribution we provide its pmf (or pdf) expressed as a function of y and the (fixed) parameters of the distribution. We also use these tables to indicate our choice of notation for the remainder of the book. We find it convenient to depart slightly from statistical convention by using lower case to denote both a random variable and its observed value. For example, we use $y \sim \mathrm{N}(\mu, \sigma^2)$ to indicate that a random variable Y is normally distributed with mean μ, variance σ^2, and pdf $f(y|\mu, \sigma)$. We also find it convenient to represent pmfs and pdfs using both bracket and shorthand notations. For example, we might represent the pmf of a binomially distributed random variable Y in either of 3 equivalent ways:

- $f(y|N, p)$
- $[y|N, p]$
- $\mathrm{Bin}(y|N, p)$.

Table 2.1. Common distributions for modeling discrete random variables.

Distribution	Notation	Probability mass function	Mean and variance
Poisson	$y \sim \text{Po}(\lambda)$	$f(y\|\lambda) = \exp(-\lambda)\lambda^y/y!$	$\text{E}(y) = \lambda$
	$[y\|\lambda] = \text{Po}(y\|\lambda)$	$y \in \{0, 1, \ldots\}$	$\text{Var}(y) = \lambda$
Bernoulli	$y \sim \text{Bern}(p)$	$f(y\|p) = p^y(1-p)^{1-y}$	$\text{E}(y) = p$
	$[y\|p] = \text{Bern}(y\|p)$	$y \in \{0, 1\}$	$\text{Var}(y) = p(1-p)$
Binomial	$y \sim \text{Bin}(N, p)$	$f(y\|N, p) = \binom{N}{y}p^y(1-p)^{N-y}$	$\text{E}(y) = Np$
	$[y\|N, p]$ $= \text{Bin}(y\|N, p)$	$y \in \{0, 1, \ldots, N\}$	$\text{Var}(y) = Np(1-p)$
Multinomial	$\boldsymbol{y} \sim \text{Multin}(N, \boldsymbol{p})$	$f(\boldsymbol{y}\|N, \boldsymbol{p}) =$ $\binom{N}{y_1 \cdots y_k}p_1^{y_1}p_2^{y_2}\cdots p_k^{y_k}$	$\text{E}(y_j) = Np_j$
	$[\boldsymbol{y}\|N, \boldsymbol{p}]$ $= \text{Multin}(\boldsymbol{y}\|N, \boldsymbol{p})$	$\times(1-p_\bullet)^{N-y_\bullet}$	$\text{Var}(y_j) = Np_j(1-p_j)$
		$y_j \in \{0, 1, \ldots, N\}$	$\text{Cov}(y_i, y_j) = -Np_ip_j$
Negative-binomial	$y \sim \text{NegBin}(\lambda, \alpha)$	$f(y\|\lambda, \alpha) =$ $\frac{\Gamma(y+\alpha)}{y!\,\Gamma(\alpha)}\left(\frac{\lambda}{\alpha+\lambda}\right)^y\left(\frac{\alpha}{\alpha+\lambda}\right)^\alpha$	$\text{E}(y) = \lambda$
	$[y\|\lambda, \alpha]$ $= \text{NegBin}(y\|\lambda, \alpha)$	$y \in \{0, 1, \ldots\}$	$\text{Var}(y) = \lambda + \lambda^2/\alpha$
Beta-binomial	$y \sim \text{BeBin}(N, \alpha, \beta)$	$f(y\|N, \alpha, \beta)$ $= \binom{N}{y}\frac{\Gamma(\alpha+y)\Gamma(N+\beta-y)}{\Gamma(\alpha+\beta+N)}$	$\text{E}(y) = N\alpha/(\alpha+\beta)$
	$[y\|N, \alpha, \beta]$ $= \text{BeBin}(y\|N, \alpha, \beta)$	$\times\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$	$\text{Var}(y)$ $= N\frac{\alpha\beta(\alpha+\beta+N)}{(\alpha+\beta)^2(\alpha+\beta+1)}$
		$y \in \{0, 1, \ldots, N\}$	

The bracket and shorthand notations are useful in describing hierarchical models that contain several distributional assumptions. The bracket notation is also useful when we want to convey probabilities or probability densities without specific reference to a particular distribution. For example, we might use $[y|\theta]$ to denote the probability of y given a parameter θ without reference to a particular distribution. Similarly, we might use $[y]$ to denote the probability of y without specifying a particular distribution or its parameters.

We adhere to the common practice of using a regular font for scalars and a bold font for vectors or matrices. For example, in our notation $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)'$ indicates a $n\times1$ vector of scalars. A prime symbol is used to indicate the transpose of a matrix or vector. There is one exception in which we deviate from this notational convention. We sometimes use θ with regular font to denote a model parameter

Table 2.2. Common distributions for modeling continuous random variables.

Distribution	Notation	Probability density function	Mean and variance
Normal	$y \sim \mathrm{N}(\mu, \sigma^2)$	$f(y\|\mu,\sigma)$ $= \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$	$\mathrm{E}(y) = \mu$
	$[y\|\mu,\sigma] = \mathrm{N}(y\|\mu,\sigma^2)$	$y \in \mathbb{R}$	$\mathrm{Var}(y) = \sigma^2$
Multivariate	$\boldsymbol{y} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	$f(\boldsymbol{y}\|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ $= (2\pi)^{-p/2}\|\boldsymbol{\Sigma}\|^{-1/2}$	$\mathrm{E}(\boldsymbol{y}) = \boldsymbol{\mu}$
normal	$[\boldsymbol{y}\|\boldsymbol{\mu}, \boldsymbol{\Sigma}] = \mathrm{N}(\boldsymbol{y}\|\boldsymbol{\mu}, \boldsymbol{\Sigma})$	$\times \exp\left(-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}\right.$ $\left.(\boldsymbol{y}-\boldsymbol{\mu})\right)\, \boldsymbol{y} \in \mathbb{R}^p$	$\mathrm{Var}(\boldsymbol{y}) = \boldsymbol{\Sigma}$
Uniform	$y \sim \mathrm{U}(a, b)$	$f(y\|a,b) = 1/(b-a)$	$\mathrm{E}(y) = (a+b)/2$
	$[y\|a,b] = \mathrm{U}(y\|a,b)$	$y \in [a,b]$	$\mathrm{Var}(y) = (b$ $- a)^2/12$
Beta	$y \sim \mathrm{Be}(\alpha,\beta)$	$f(y\|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}$ $\times (1-y)^{\beta-1}$	$\mathrm{E}(y) = \alpha/(\alpha+\beta)$
	$[y\|\alpha,\beta] = \mathrm{Be}(y\|\alpha,\beta)$	$y \in [0,1]$	$\mathrm{Var}(y) = \alpha\beta/\{(\alpha$ $+\beta)^2(\alpha+\beta+1)\}$
Dirichlet	$\boldsymbol{y} \sim \mathrm{Dir}(\boldsymbol{\alpha})$	$f(\boldsymbol{y}\|\boldsymbol{\alpha})$ $= \frac{\Gamma(\alpha_1+\cdots+\alpha_k)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_k)} y_1^{\alpha_1-1}\cdots$	$\mathrm{E}(y_j)$ $= \alpha_j/\sum_{l=1}^k \alpha_l$
	$[\boldsymbol{y}\|\boldsymbol{\alpha}] = \mathrm{Dir}(\boldsymbol{y}\|\boldsymbol{\alpha})$	$y_k^{\alpha_k-1} y_j \in [0,1];$ $\sum_{j=1}^k y_j = 1$	$\mathrm{Var}(y_j)$ $= \frac{\alpha_j(-\alpha_j+\sum_l \alpha_l)}{(\sum_l \alpha_l)^2(1+\sum_l \alpha_l)}$
Gamma	$y \sim \mathrm{Gamma}(\alpha,\beta)$	$f(y\|\alpha,\beta)$ $= \frac{\beta^\alpha}{\Gamma(\alpha)} y^{\alpha-1}\exp(-\beta y)$	$\mathrm{E}(y) = \alpha/\beta$
	$[y\|\alpha,\beta] = \mathrm{Gamma}(y\|\alpha,\beta)$	$y \in \mathbb{R}^+$	$\mathrm{Var}(y) = \alpha/\beta^2$

that may be a scalar *or* a vector depending on the context. To avoid confusion, in these cases, we state explicitly that θ is possibly vector-valued.

2.1.3 Probability Rules for Random Variables

Earlier we noted that a statistical model is composed of one or more random variables. In fact, in most inferential problems it would be highly unusual to observe the value of only one random variable because it is difficult to learn much from a sample of size one. Therefore, we need to understand the 'rules' involved in modeling multiple outcomes.

Since outcomes are modeled as observed values of random variables, it should come as no surprise that the *laws of probability* provide the foundation for modeling multiple outcomes. To illustrate, let's consider the joint distribution of only two random variables. Let (y, z) denote a vector of two discrete random variables. The *joint pmf* of (y, z) is defined as follows:

$$f(y, z) = \Pr(Y = y, Z = z),$$

where we suppress the conditioning on the parameter(s) needed to characterize the distribution of (y, z). The notation for the *joint pdf* of two continuous random variables is identical (i.e., $f(y, z)$). Now suppose we want to calculate a *marginal pmf* or *marginal pdf* for each random variable. If y and z are discrete-valued, the marginal pmf is calculated by summation:

$$f(y) = \sum_z f(y, z)$$
$$f(z) = \sum_y f(y, z).$$

If y and z are continuous random variables, their marginal pdfs are computed by integration:

$$f(y) = \int_{-\infty}^{\infty} f(y, z)\,\mathrm{d}z$$
$$f(z) = \int_{-\infty}^{\infty} f(y, z)\,\mathrm{d}y$$

Statistical models are often formulated in terms of *conditional* outcomes; therefore, we often need to calculate conditional probabilities, such as $\Pr(Y = y|Z = z)$ or $\Pr(Z = z|Y = y)$. Fortunately, *conditional pmfs* and *conditional pdfs* are easily calculated from joint and marginal distribution functions. In particular, the conditional pmf (or pdf) of y given z is

$$f(y|z) = \frac{f(y, z)}{f(z)}.$$

Likewise, the conditional pmf (or pdf) of z given y is

$$f(z|y) = \frac{f(y, z)}{f(y)}.$$

The above formulae may not seem useful now, but they will be used extensively in later chapters, particularly in the construction of hierarchical models. However,

one immediate use is in evaluating the consequences of independence. Suppose random variables y and z are assumed to be *independent*; thus, knowing the value of z gives us no additional information about the value of y and vice versa. Then, by definition, the joint pmf (or pdf) of the the vector (y, z) equals the product of the marginal pmfs (pdfs) as follows:

$$f(y, z) = f(y)f(z).$$

Now, recalling the definition of the conditional pmf (pdf) of y given z and substituting the above expression yields

$$\begin{aligned} f(y|z) &= \frac{f(y, z)}{f(z)} \\ &= \frac{f(y)f(z)}{f(z)} \\ &= f(y). \end{aligned}$$

Therefore, the conditional probability (or probability density) of y given z is identical to the marginal probability (or probability density) of y. This result confirms that knowledge of z provides no additional information about y when the random variables y and z are independent.

One well-known application of the laws of probability is called the *law of total probability*. To illustrate, let's consider two discrete random variables, y and z, and suppose z has n distinct values, $z_1, z_2, \ldots, z_n$. The law of total probability corresponds to the expression required for calculating the marginal probability $f(y)$ given only the conditional and marginal pmfs, $f(y|z)$ and $f(z)$, respectively. We know how to calculate $f(y)$ from the joint pmf $f(y, z)$:

$$f(y) = \sum_{z=z_1}^{z_n} f(y, z)$$

By definition of conditional probability, $f(y, z) = f(y|z)f(z)$. Therefore, substituting this expression into the right-hand side of the above equation yields

$$\begin{aligned} f(y) &= \sum_{z=z_1}^{z_n} f(y|z)f(z) \\ &= f(y|z_1)f(z_1) + f(y|z_2)f(z_2) + \cdots + f(y|z_n)f(z_n) \end{aligned}$$

which is the law of total probability. In this equation $f(y)$ can be thought of as a weighted average (or expectation) of the conditional probabilities where the weights correspond to the marginal probabilities $f(z)$. The law of total probability is often used to remove latent random variables from hierarchical models so that the parameters of the model can be estimated. We will see several examples of this type of marginalization in subsequent chapters.

2.2 THE ROLE OF APPROXIMATING MODELS

Inference begins with the data observed in a sample. These data may include field records from an observational study (a survey), or they may be outcomes of a designed experiment. In either case the observed data are manifestations of at least two processes: the sampling or experimental procedure, i.e., the process used to collect the data, *and* the ecological process that we hope to learn about.

Proponents of model-based inference base their conclusions on one or more *approximating models* of the data. These models should account for the observational (or data-gathering) process and for the underlying ecological process. Let's consider a simple example. Suppose we randomly select n individual animals from a particular location and measure each animal's body mass y_i $(i = 1, \ldots, n)$ for the purpose of estimating the mean body mass of animals at that location. In the absence of additional information, such as the age or size of an individual, we might be willing to entertain a relatively simple model of each animal's mass, e.g., $y_i \sim \mathrm{N}(\mu, \sigma^2)$. This model has two parameters, μ and σ, and we cannot hope to estimate them from the body mass of a single individual; therefore, an additional modeling assumption is required. Because animals were selected at random to obtain a representative sample of those present, it seems reasonable to assume *mutual independence* among the n measurements of body mass. Given this additional assumption, an approximating model of the sample data is

$$[y_1, y_2, \ldots, y_n|\mu, \sigma] = \prod_{i=1}^{n} \mathrm{N}(y_i|\mu, \sigma^2). \tag{2.2.1}$$

Therefore, the joint pdf of observed body masses is modeled as a product of identical marginal pdfs (in this case $\mathrm{N}(y|\mu, \sigma^2)$). We describe such random variables as *independent and identically distributed* (abbreviated as *iid*) and use the notation, $y_i \overset{iid}{\sim} \mathrm{N}(\mu, \sigma^2)$, as a compact summary of these assumptions.

Now that we have formulated a model of the observed body masses, how is the model used to estimate the mean body mass of animals that live at the sampled locations? After all, that is the real inferential problem in our example. To answer this question, we must recognize the connection between the parameters of the model and the scientifically relevant estimand, the mean body mass of animals in the population. In our model-based view each animal's body mass is a random variable, and we can prove that $\mathrm{E}(y_i) = \mu$ under the assumptions of the approximating model. Therefore, in estimating μ we solve the inference problem.

This equivalence may seem rather obvious; however, if we had chosen a *different* approximating model of body masses, the mean body mass would necessarily be specified in terms of *that* model's parameters. For example, suppose the sample of n animals includes both males and females but we don't observe the sex of each

individual. If we know that males and females of similar ages have different average masses and if we have reason to expect an uneven sex ratio, then we might consider a mixture of 2 normal distributions as an approximating model:

$$y_i \overset{iid}{\sim} p\mathrm{N}(\mu_m, \sigma^2) + (1-p)\mathrm{N}(\mu_f, \sigma^2),$$

where p denotes the unknown proportion of males in the population and μ_m and μ_f denote the mean body masses of males and females, respectively. Under the assumptions of this model we can show that $\mathrm{E}(y_i) = p\mu_m + (1-p)\mu_f$; therefore, to solve the inference problem of estimating mean body mass, we must estimate the model parameters, p, μ_m, and μ_f.

We have used this example to illustrate the crucial role of modeling in inference problems. The parameters of a model specify the theoretical properties of random variables, and we can use those properties to deduce how a model's parameters are related to one or more scientifically relevant estimands. Often, but not always, these estimands may be formulated as summaries of observations, such as the sample mean. In these cases it is important to remember that such summary statistics may be related to the parameters of a model, but the two are not generally equivalent.

2.3 CLASSICAL (FREQUENTIST) INFERENCE

In the previous section we established the role of modeling in inference. Here we describe classical procedures for estimating model parameters and for using the estimates to make some kind of inference or prediction.

Let $\boldsymbol{y} = (y_1, \ldots, y_n)$ denote a sample of n observations. Suppose we develop an approximating model of $\boldsymbol{y}$ that contains a (possibly vector-valued) parameter θ. The model is a formal expression of the processes that are assumed to have produced the observed data. In classical inference the model parameter θ is assumed to have a fixed, but unknown, value. The observed data $\boldsymbol{y}$ are regarded as a single realization of the stochastic processes specified in the model. Similarly, any summary of $\boldsymbol{y}$, such as the sample mean $\bar{y}$, is viewed as a random outcome.

Now suppose we have a procedure or method for estimating the value of θ given the information in the sample, i.e, given $\boldsymbol{y}$. In classical statistics such procedures are called *estimators* and the result of their application to a particular data set yields an *estimate* $\hat{\theta}$ of the fixed parameter θ. Of course, different estimators can produce different estimates given the same set of data, and considerable statistical theory has been developed to evaluate the operating characteristics of different estimators (e.g., bias, mean squared error, etc.) in different inference problems. However, regardless of the estimator chosen for analysis, classical inference views the estimate $\hat{\theta}$ as a random outcome because it is a function of $\boldsymbol{y}$, which also is regarded as a random outcome.

To make inferences about θ, classical statistics appeals to the idea of hypothetical outcomes under *repeated sampling*. In other words, classical statistics views $\hat{\theta}$ as a single outcome that belongs to a distribution of estimates associated with hypothetical repetitions of an experiment or survey. Under this view, the fixed value θ and the assumptions of the model represent a mechanism for generating a random, hypothetical sequence of data sets and parameter estimates:

$$(\boldsymbol{y}_1, \hat{\theta}_1), (\boldsymbol{y}_2, \hat{\theta}_2), (\boldsymbol{y}_3, \hat{\theta}_3), \ldots$$

Therefore, probability statements about θ (i.e., inferences) are made with respect to the distribution of estimates of θ that could have been obtained in repeated samples.

For this reason those who practice classical statistics are often referred to as *frequentists*. In classical statistics the role of probability in computing inferences is based on the relative frequency of outcomes in repeated samples (experiments or surveys). Frequentists never use probability directly as an expression of degrees of belief in the magnitude of θ. Probability statements are based entirely on the hypothetical distribution of $\hat{\theta}$ generated under the model and repeated sampling. We will have more to say about the philosophical and practical differences that separate classical statistics and Bayesian statistics later in this chapter.

2.3.1 Maximum Likelihood Estimation

We have not yet described an example of a model-based estimator, i.e., a procedure for estimating θ given the observed data $\boldsymbol{y}$. One of the most widely adopted examples in all of classical statistics is the *maximum likelihood estimator* (MLE), which can be traced to the efforts of Daniel Bernoulli and Johann Heinrich Lambert in the 18th century (Edwards, 1974). However, credit for the MLE is generally awarded to the brilliant scientist, Ronald Aylmer Fisher, who in the early 20th century fully developed the MLE for use in inference problems (Edwards, 1992). Among his many scientific contributions, Fisher invented the concept of *likelihood* and described its application in point estimation, hypothesis testing, and other inference problems. The concept of likelihood also has connections to Bayesian inference. Therefore, we have chosen to limit our description of classical statistics to that of likelihood-based inference.

Let's assume, without loss of generality, that the observed data $\boldsymbol{y}$ are modeled as continuous random variables and that $f(\boldsymbol{y}|\theta)$ denotes the joint pdf of $\boldsymbol{y}$ given a model indexed by the parameter θ. In many cases the observations are mutually

independent so that the joint pdf can be expressed as a product of individual pdfs as follows:

$$f(\boldsymbol{y}|\theta) = \prod_{i=1}^{n} g(y_i|\theta).$$

However, the concept of likelihood applies equally well to samples of dependent observations, so we will not limit our notation to cases of independence.

To define the MLE of θ, Fisher viewed the joint pdf of $\boldsymbol{y}$ as a function of θ,

$$L(\theta|\boldsymbol{y}) \equiv f(\boldsymbol{y}|\theta),$$

which he called the *likelihood function*. The MLE of θ, which we denote by $\hat{\theta}$, is defined as the particular value of θ that maximizes the likelihood function $L(\theta|\boldsymbol{y})$. Heuristically, one can think of $\hat{\theta}$ as the value of θ that is most likely given the data because $\hat{\theta}$ assigns the highest chance to the observations in $\boldsymbol{y}$. Fisher always intended the likelihood $L(\theta|\boldsymbol{y})$ to be interpreted as a measure of *relative* support for different hypotheses (i.e., different values of θ); he never considered the likelihood function to have an absolute scale, such as a probability. This distinction provides an important example of the profound differences between the classical and Bayesian views of inference, as we shall see later (Section 2.4).

The MLE of θ is, by definition, the solution of an optimization problem. In some cases this problem can be solved analytically using calculus, which allows the MLE to be expressed in closed form as a function of $\boldsymbol{y}$. If this is not possible, numerical methods of optimization must be used to compute an approximation of $\hat{\theta}$; however, these calculations can usually be done quite accurately and quickly with modern computers and optimization algorithms.

2.3.1.1 Example: estimating the probability of occurrence (analytically)

Suppose a survey is designed to estimate the average occurrence of an animal species in a region. As part of the survey, the region is divided into a lattice of sample units of uniform size and shape, which we will call 'locations', for simplicity. Now imagine that n of these locations are selected at random and that we are able to determine with certainty whether the species is present ($y = 1$) or absent ($y = 0$) at each location.[1]

On completion of the survey, we have a sample of binary observations $\boldsymbol{y} = (y_1, \ldots, y_n)$. Because the sample locations are selected at random, it seems

[1] Observations of animal occurrence are rarely made with absolute certainty when sampling natural populations; however, we assume certainty in this example to keep the model simple.

reasonable to assume that the observations are mutually independent. If we further assume that the probability of occurrence is identical at each location, this implies a rather simple model of the data:

$$y_i \overset{iid}{\sim} \text{Bern}(\psi), \tag{2.3.1}$$

where ψ denotes the probability of occurrence.

We could derive the MLE of ψ based on Eq. (2.3.1) alone; however, we can take a mathematically equivalent approach by considering the implications of Eq. (2.3.1) for a summary of the binary data. Let $v = \sum_{i=1}^{n} y_i$ denote the total number of sample locations where the species is present. This definition allows the information in $\boldsymbol{y}$ to be summarized as the frequency of ones, v, and the frequency of zeros, $n - v$. Given the assumptions in Eq. (2.3.1), the probability of any particular set of observations in $\boldsymbol{y}$ is

$$\psi^v (1 - \psi)^{n-v} \tag{2.3.2}$$

for $v = 0, 1, \ldots, n$. However, to formulate the model in terms of v, we must account for the total number of ways that v ones and $n - v$ zeros could have been observed, which is given by the combinatorial

$$\binom{n}{v} = \frac{n!}{v!(n-v)!}. \tag{2.3.3}$$

Combining Eqs. (2.3.2) and (2.3.3) yields the total probability of observing v ones and $n - v$ zeros independent of the ordering of the binary observations in $\boldsymbol{y}$:

$$f(v|\psi) = \binom{n}{v} \psi^v (1 - \psi)^{n-v}. \tag{2.3.4}$$

The astute reader will recognize that $f(v|\psi)$ is just the pmf of a binomial distribution with index n and parameter ψ (see Table 2.1).

This simple model provides the likelihood of ψ given v (and the sample size n)

$$L(\psi|v) = \psi^v (1 - \psi)^{n-v}, \tag{2.3.5}$$

where the combinatorial term has been omitted because $\binom{n}{v}$ does not involve ψ. (Note that the combinatorial term may be ignored given Fisher's definition of likelihood because $\binom{n}{v}$ does not involve ψ and only contributes a multiplicative constant to the likelihood.) The MLE of ψ is the value of ψ that maximizes

$L(\psi|v)$. Because $L(\psi|v)$ is non-negative for admissible values of ψ, the MLE of ψ also maximizes

$$\log L(\psi|v) = v \log \psi + (n - v) \log(1 - \psi). \tag{2.3.6}$$

This results stems from the fact that the natural logarithm is a one-to-one, monotone function of its argument. The MLE of ψ is the solution of either of the following equations,

$$\frac{\mathrm{d}L(\psi|z)}{\mathrm{d}\psi} = 0 \quad \text{or} \quad \frac{\mathrm{d}\log L(\psi|z)}{\mathrm{d}\psi} = 0$$

and equals $\hat{\psi} = v/n$. Therefore, the MLE of ψ is equivalent to the sample mean of the binary observations: $\bar{y} = (1/n)\sum_{i=1}^{n} y_i = v/n$.

2.3.1.2 Example: estimating the probability of occurrence (numerically)

In the previous example we were able to derive $\hat{\psi}$ in closed form; however, in many estimation problems the MLE cannot be obtained as the analytic solution of a differential equation (or system of differential equations for models with $\geq$2 parameters). In such problems the MLE must be estimated numerically.

To illustrate, let's consider the previous example and behave as though we could not have determined that $\hat{\psi} = v/n$. Suppose we select a random sample of $n = 5$ locations and observe a species to be present at only $v = 1$ of those locations. How do we compute the MLE of ψ numerically?

One possibility is a brute-force calculation. Because ψ is bounded on the interval $[0, 1]$, we can evaluate the log-likelihood function in Eq. (2.3.6) for an arbitrarily large number of ψ values that span this interval (e.g., $\epsilon, 2\epsilon, \ldots, 1 - \epsilon$, where $\epsilon > 0$ is an arbitrarily small, positive number). Then, we identify $\hat{\psi}$ as the particular value of ψ with the highest log-likelihood (see Figure 2.1). Panel 2.1 contains the **R** code needed to do these calculations and yields $\hat{\psi} = 0.2$, which is assumed to be correct within $\pm 10^{-6}$ $(= \epsilon)$. In fact, the answer is exactly correct, since $v/n = 1/5 = 0.2$.

Another possibility for computing a numerical approximation of $\hat{\psi}$ is to employ an optimization algorithm. For example, **R** contains two procedures, `nlm` and `optim`, which can be used to find the *minima* of functions of arbitrary form. Of course, these procedures can also be used to find maxima by simply changing the sign of the objective function. For example, *maximizing* a log-likelihood function is equivalent to *minimizing* a *negative* log-likelihood function. In addition to defining the function to be minimized, **R**'s optimization algorithms require a starting point. Ideally, the starting point should approximate the solution. Panel 2.2 contains an example of **R** code that uses `optim` to do these calculations and produces an estimate of $\hat{\psi}$ that is very close to the correct answer. Note, however, that **R** also reports a warning message upon fitting the model. During the optimization `optim` apparently

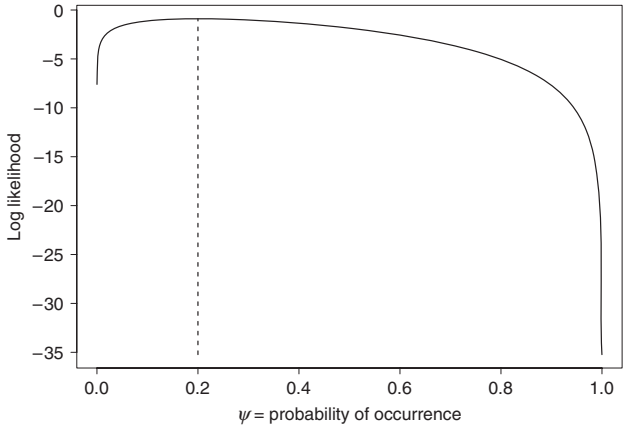

Figure 2.1. Log likelihood for a binomial outcome ($v = 1$ successes in $n = 5$ trials) evaluated over the admissible range of ψ values. Dashed vertical line indicates the MLE.

```
> v=1
> n=5
> eps=1e-6
> psi = seq(eps, 1-eps, by=eps)
> logLike = v*log(psi) + (n-v)*log(1-psi)
> psi[logLike==max(logLike)]
[1] 0.2
```

Panel 2.1. R code for brute-force calculation of MLE of ψ.

attempted to evaluate the function `dbinom` using values of ψ that were outside the interval $[0, 1]$. In this case the message can be ignored, but it's an indication that computational improvements are possible. We will return to this issue later.

2.3.1.3 Example: estimating parameters of normally distributed data

In Section 2.2 we described an example where body mass was measured for each of n randomly selected animals. As a possible model, we assumed that the body masses in the sample were independent and identically distributed as follows: $y_i \overset{iid}{\sim} \mathrm{N}(\mu, \sigma^2)$. Here, we illustrate methods for computing the MLE of the parameter vector (μ, σ^2).

Given our modeling assumptions, the joint pdf of the observed body masses $\boldsymbol{y}$ is a product of identical marginal pdfs, $\mathrm{N}(y_i|\mu,\sigma^2)$, as noted in Eq. (2.2.1). Therefore, the likelihood function for these data is

$$L(\mu,\sigma^2|\boldsymbol{y}) = (2\pi\sigma^2)^{-n/2}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2\right). \tag{2.3.7}$$

To find the MLE, we need to solve the following set of simultaneous equations:

$$\frac{\partial \log L(\mu,\sigma^2)}{\partial \mu} = 0$$
$$\frac{\partial \log L(\mu,\sigma^2)}{\partial \sigma^2} = 0.$$

It turns out that an analytical solution exists in closed form. The MLE is $(\hat{\mu},\hat{\sigma}^2) = (\bar{y}, \frac{n-1}{n}s^2)$, where $\bar{y}$ and s^2 denote the sample mean and variance, respectively, of the observed body masses. Note that $\hat{\sigma}^2$ is strictly less than s^2, the usual (unbiased) estimator of σ^2; however, the difference becomes negligible as sample size n increases.

Suppose we could not have found the MLE of (μ,σ^2) in closed form. In this case we need to compute a numerical approximation. We could try the brute-force approach used in the earlier example, but in this example we would need to evaluate the log-likelihood function over 2 dimensions. Furthermore, the parameter space is unbounded ($\mathbb{R}\times\mathbb{R}^+$), so we would need to restrict evaluations of the log-likelihood to be in the vicinity of the MLE, which we do not know!

It turns out that the brute-force approach is seldom feasible, particularly as the number of model parameters becomes large. Therefore, numerical methods of

```
> v=1
> n=5
> neglogLike = function(psi) -dbinom(v, size=n, prob=psi, log=TRUE)
> fit = optim(par=0.5, fn=neglogLike, method='BFGS')
Warning messages: 1: NaNs produced in: dbinom(x, size, prob, log) 2:
NaNs produced in: dbinom(x, size, prob, log)
>
> fit$par
[1] 0.2000002
>
```

Panel 2.2. R code for numerically maximizing the likelihood function to approximate $\hat{\psi}$.

optimization are often used to compute an approximation of the MLE. To illustrate, suppose we observe the following body masses of 10 animals:

$$\boldsymbol{y} = (8.51, 4.03, 8.20, 4.19, 8.72, 6.15, 5.40, 8.66, 7.91, 8.58)$$

which have sample mean $\bar{y} = 7.035$ and sample variance $s^2 = 3.638$. Panel 2.3 contains **R** code for approximating the MLE and yields the estimates $\hat{\mu} = 7.035$ and $\hat{\sigma}^2 = 3.274$, which are the correct answers.

2.3.2 Properties of MLEs

Maximum likelihood estimators have several desirable properties. In this section we describe these properties and illustrate their consequences in inference problems. In particular, we show how MLEs are used in the construction of confidence intervals.

2.3.2.1 Invariance to reparameterization

If $\hat{\theta}$ is the MLE of θ, then for any one-to-one function $\tau(\theta)$, the MLE of $\tau(\theta)$ is $\tau(\hat{\theta})$. This invariance to reparameterization can be extremely helpful in computing

```
> y = c(8.51, 4.03, 8.20, 4.19, 8.72, 6.15, 5.40, 8.66, 7.91, 8.58)
>
> neglogLike = function(param) {
+ mu = param[1]
+ sigma = exp(param[2])
+-sum(dnorm(y,mean=mu,sd=sigma, log=TRUE))
+}
>
> fit = optim(par=c(0,0), fn=neglogLike, method='BFGS')
> fit$par
[1] 7.0350020 0.5930949
>
> exp(fit$par[2])^2
[1] 3.274581
>
```

Panel 2.3. **R** code for numerically maximizing the likelihood function to approximate $(\hat{\mu}, \log \hat{\sigma})$.

MLEs by numerical approximation. For example, let $\gamma = \tau(\theta)$ and suppose we can compute $\hat{\gamma}$ that maximizes $L_1(\gamma|\boldsymbol{y})$ easily; then, by the property of invariance we can deduce that $\hat{\theta} = \tau^{-1}(\hat{\gamma})$ maximizes $L_2(\theta|\boldsymbol{y})$ without actually computing the solution of $\mathrm{d}L_2(\theta|\boldsymbol{y})/\mathrm{d}\theta = 0$, which may involve numerical difficulties.

To illustrate, let's consider the problem of estimating the probability of occurrence ψ that we described in Section 2.3.1.1. The *logit* function, a one-to-one transformation of ψ, is defined as follows:

$$\text{logit } (\psi) = \log\left(\frac{\psi}{1-\psi}\right)$$

and provides a mapping from the domain of ψ ($[0, 1]$) to the entire real line. Let $\theta = \text{logit}(\psi)$ denote a reparameterization of ψ. We can maximize the likelihood of θ given v to obtain $\hat{\theta}$ and then calculate $\hat{\psi}$ by inverting the transformation as follows:

$$\begin{aligned}\hat{\psi} &= \text{logit}^{-1}(\hat{\theta}) \\ &= 1/(1 + \exp(-\hat{\theta})).\end{aligned}$$

We will use this inversion often; therefore, in the remainder of this book we let $\text{expit}(\theta)$ denote the function $\text{logit}^{-1}(\theta)$ as a matter of notational convenience.

Panel 2.4 contains **R** code for computing $\hat{\theta}$ by numerical optimization and for computing $\hat{\psi}$ by inverting the transformation. Notice that the definition of the **R** function `neglogLike` is identical to that used earlier (Panel 2.2) except that we have substituted `theta` for `psi` as the function's argument and `expit(theta)` for `psi` in the body of the function. Therefore, the extra coding required to compute ψ on the logit scale is minimal. Notice also in Panel 2.4 that in maximizing the likelihood function of θ, **R** did not produce the somewhat troubling warning messages that appeared in maximizing the likelihood function of ψ (cf. Panel 2.2). The default behavior of **R**'s optimization functions, `optim` and `nlm`, is to provide an *unconstrained* minimization wherein no constraints are placed on the magnitude of the argument of the function being minimized. In other words, if the function's argument is a vector of p components, their value is assumed to lie anywhere in $\mathbb{R}^p$. In our example the admissible values of θ include the entire real line; in contrast, the admissible values of ψ are confined to a subset of the real line ($[0, 1]$).

The lesson learned from this example is that when using *unconstrained* optimization algorithms to maximize a likelihood function of p parameters, one should typically try to formulate the likelihood so that the parameters are defined in $\mathbb{R}^p$. The invariance of MLEs always allows us to back-transform the parameter estimates if that is necessary in the context of the problem.

2.3.2.2 Consistency

Suppose the particular set of modeling assumptions summarized in the joint pdf $f(\boldsymbol{y}|\theta)$ is true, i.e., the approximating model of the data $\boldsymbol{y}$ correctly describes the process that generated the data. Under these conditions, we can prove that $\hat{\theta}$, the MLE of θ, converges to θ as the sample size n increases, which we denote mathematically as follows:

$$\hat{\theta} \to \theta \quad \text{as } n \to \infty.$$

Although the assumptions of an approximating model are unlikely to hold exactly, it is reassuring to know that with enough data, the MLE is guaranteed to provide the 'correct' answer.

```
> v=1
> n=5
> expit = function(x) 1/(1+exp(-x))
>
> neglogLike = function(theta) -dbinom(v, size=n, prob=expit(theta), log=TRUE)
> fit = optim(par=0, fn=neglogLike, method='BFGS')
>
> fit$par
[1] -1.386294
>
> expit(fit$par)
[1] 0.2
```

Panel 2.4. R code for numerically maximizing the likelihood function to estimate $\hat{\theta} = \text{logit}(\hat{\psi})$.

2.3.2.3 Asymptotic normality

As in the previous section, suppose the particular set of modeling assumptions summarized in the joint pdf $f(\boldsymbol{y}|\theta)$ is true. If, in addition, a set of 'regularity conditions' that have to do with technical details[2] are satisfied, we can prove the following limiting behavior of the MLE of θ:

In a hypothetical set of repeated samples with θ fixed and with $n \to \infty$,

$$(\hat{\theta} - \theta) \mid \theta \sim \text{N}(0, [I(\hat{\theta})]^{-1}), \tag{2.3.8}$$

where $I(\hat{\theta}) = -\frac{\partial^2 \log L(\theta|\boldsymbol{y})}{\partial\theta\partial\theta}|_{\theta=\hat{\theta}}$ is called the *observed information*.

[2]Such as identifiability of the model's parameters and differentiability of the likelihood function. See page 516 of Casella and Berger (2002) for a complete list of conditions.

If θ is a vector of p parameters, then $I(\hat{\theta})$ is a $p \times p$ matrix called the *observed information matrix*.

According to Eq. (2.3.8), the distribution of the discrepancy, $\hat{\theta}-\theta$, obtained under repeated sampling is approximately normal with mean zero as $n \to \infty$. Therefore, $\hat{\theta}$ is an *asymptotically unbiased* estimator of θ. Similarly, Eq. (2.3.8) implies that the inverse of the observed information provides the estimated *asymptotic variance* (or *asymptotic covariance matrix*) of $\hat{\theta}$.

The practical utility of asymptotic normality is evident in the construction of $100(1-\alpha)$ percent *confidence intervals* for θ. For example, suppose θ is scalar-valued; then in repeated samples, the *random* interval

$$\hat{\theta} \pm z_{1-\alpha/2}([I(\hat{\theta})]^{-1})^{1/2} \tag{2.3.9}$$

'covers' the fixed value θ $100(1-\alpha)$ percent of the time, provided n is sufficiently large. Here, $z_{1-\alpha/2}$ denotes the $(1-\alpha/2)$ quantile of a standard normal distribution. Note that Eq. (2.3.9) does *not* imply that any *individual* confidence interval includes θ with probability $1-\alpha$. This misinterpretation of the role of probability is an all-too-common occurrence in applications of statistics. An individual confidence interval either includes θ or it doesn't. A correct probability statement (or inference) refers to the proportion of confidence intervals that include θ in a hypothetical, infinitely long sequence of repeated samples. In this sense $1-\alpha$ is the probability (relative frequency) that an interval constructed using Eq. (2.3.9) includes the fixed value θ.

Example: estimating the probability of occurrence

As an illustration, let's compute a 95 percent confidence interval for ψ, the probability of occurrence, that was defined earlier in an example (Section 2.3.1.1). The information is easily derived using calculus:

$$\frac{\mathrm{d}^2 \log L(\psi|v)}{\mathrm{d}\psi^2} = I(\psi) = \frac{n}{\psi(1-\psi)}.$$

The model has only one parameter ψ; therefore, we simply take the reciprocal of $I(\psi)$ to compute its inverse. Substituting $\hat{\psi}$ for ψ yields the 95 percent confidence interval for ψ:

$$\hat{\psi} \pm 1.96\sqrt{\frac{\hat{\psi}(1-\hat{\psi})}{n}}.$$

Suppose we had not been able to derive the observed information or to compute its inverse analytically. In this case we would need to compute a numerical approximation of $[I(\hat{\psi})]^{-1}$. Panel 2.5 contains the **R** code for computing the MLE of

ψ and a 95 percent confidence interval having observed only $v = 1$ occupied site in a sample of $n = 5$ sites. As before, we estimate $\hat{\psi} = 0.20$. A numerical approximation of $I(\hat{\psi})$ is computed by adding `hessian=TRUE` to the list of `optim`'s arguments. After rounding, our **R** code yields the following 95 percent confidence interval for ψ: $[-0.15, 0.55]$. This is the correct answer, but it includes negative values of ψ, which don't really make sense because ψ is bounded on $[0, 1]$ by definition.

One solution to this problem is to compute a confidence interval for $\theta = \text{logit}(\psi)$ and then transform the upper and lower confidence limits back to the ψ scale (see Panel 2.6). This approach produces an asymmetrical confidence interval for ψ ($[0.027, 0.691]$), but the interval is properly contained in $[0, 1]$.

Another solution to the problem of nonsensical confidence limits is to use a procedure which produces limits that are invariant to reparameterization. We will describe such procedures in the context of hypothesis testing (see Section 2.5). For now, we simply note that confidence limits computed using these procedures and those computed using Eq. (2.3.9) are asymptotically equivalent. The construction of intervals based on Eq. (2.3.9) is far more popular because the confidence limits are relatively easy to compute. In contrast, the calculation of confidence limits based on alternative procedures is more challenging in many instances.

Before leaving our example of interval estimation for ψ, let's examine the influence of sample size. Suppose we had examined a sample of $n = 50$ randomly selected

```
> v=1
> n=5
> neglogLike = function(psi) -dbinom(v, size=n, prob=psi, log=TRUE)
> fit = optim(par=0.5, fn=neglogLike, method='BFGS', hessian=TRUE)
Warning messages: 1: NaNs produced in: dbinom(x, size, prob, log) 2:
NaNs produced in: dbinom(x, size, prob, log)
>
> fit$par
[1] 0.2000002
>
> psi.mle = fit$par
> psi.se = sqrt(1/fit$hessian)
> zcrit = qnorm(.975)
> c(psi.mle-zcrit*psi.se,  psi.mle+zcrit*psi.se)
[1] -0.1506020  0.5506024
>
```

Panel 2.5. **R** code for computing a 95 percent confidence interval for ψ.

locations (a tenfold increase in sample size) and had observed the species to be present at $v = 10$ of these locations. Obviously, our estimate of ψ is unchanged because $\hat{\psi} = 10/50 = 0.2$. But how has the uncertainty in our estimate changed? Earlier we showed that $I(\psi) = n/(\psi(1-\psi))$. Because $\hat{\psi}$ is identical in both samples, we may conclude that the observed information in the sample of $n = 50$ locations is ten times higher than that in the sample of $n = 5$ locations, as shown in Table 2.3. In fact, this is easily illustrated by plotting the log-likelihood functions for each sample (Figure 2.2). Notice that the curvature of the log-likelihood function in the vicinity of the MLE is greater for the larger sample ($n = 50$). This is consistent with the differences in observed information because $I(\hat{\psi})$ is the negative of the second derivative of $\log L(\psi|v)$ evaluated at $\hat{\psi}$, which essentially measures the curvature of $\log L(\psi|v)$ at the MLE. The log-likelihood function decreases with distance from $\hat{\psi}$ more rapidly in the sample of $n = 50$ locations than in the sample of $n = 5$ locations; therefore, we might expect the estimated precision of $\hat{\psi}$ to be higher in the larger sample. This is exactly the case; the larger sample yields a narrower confidence interval for ψ. Table 2.3 and Figure 2.2 also illustrate the effects of parameterizing the log-likelihood in terms of $\theta = \text{logit}(\psi)$. The increase in observed information associated with the larger sample is the same (a tenfold increase), and this results in a narrower confidence interval. The confidence limits for ψ based on the asymptotic normality of $\hat{\theta}$ are not identical to those based on the asymptotic normality of $\hat{\psi}$, as mentioned earlier; however, the discrepancy between the two confidence intervals is much lower in the larger sample.

```
> v=1
> n=5
> expit = function(x) 1/(1+exp(-x))
>
> neglogLike = function(theta) -dbinom(v, size=n, prob=expit(theta), log=TRUE)
> fit = optim(par=0, fn=neglogLike, method='BFGS', hessian=TRUE)
>
> theta.mle = fit$par
> theta.se = sqrt(1/fit$hessian)
> zcrit = qnorm(.975)
> expit(c(theta.mle-zcrit*theta.se, theta.mle+zcrit*theta.se))
[1] 0.02718309 0.69104557
>
```

Panel 2.6. R code for computing a 95 percent confidence interval for ψ by back-transforming the lower and upper limits of $\theta = \text{logit}(\psi)$.

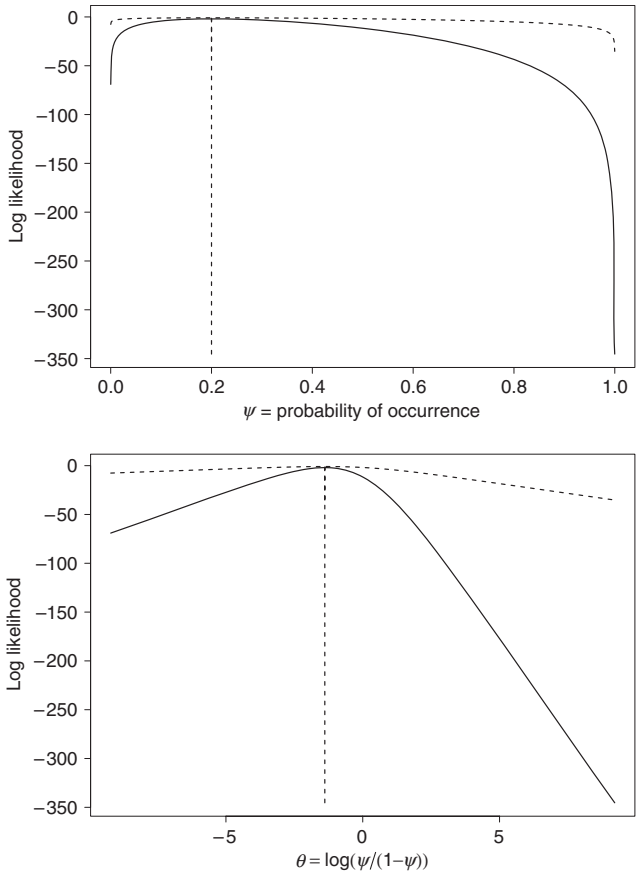

Figure 2.2. Comparison of log-likelihood functions for binomial outcomes based on different sample sizes, $n = 5$ (dashed line) and $n = 50$ (solid line), and different parameterizations, ψ (upper panel) and $\text{logit}(\psi)$ (lower panel). Dashed vertical line indicates the MLE, which is identical in both samples.

Table 2.3. Effects of sample size n and parameterization on 95 percent confidence intervals for ψ.

n	$I(\hat{\psi})$	$[I(\hat{\psi})]^{-1}$	95% C.I. for ψ	$I(\hat{\theta})$	$[I(\hat{\theta})]^{-1}$	95% C.I. for $\psi = \text{expit}(\theta)$
5	31.25	0.0320	[-0.15, 0.55]	0.8	1.250	[0.03, 0.69]
50	312.50	0.0032	[0.09, 0.31]	8.0	0.125	[0.11, 0.33]

Example: estimating parameters of normally distributed data
We conclude this section with a slightly more complicated example to illustrate the construction of confidence intervals when the model contains two or more parameters. In an earlier example, the body weights of n animals were modeled as $y_i \overset{iid}{\sim} \mathrm{N}(\mu, \sigma^2)$, and we determined that the MLE is $(\hat{\mu}, \hat{\sigma}^2) = (\bar{y}, \frac{n-1}{n} s^2)$, where $\bar{y}$ and s^2 denote the sample mean and variance, respectively, of the observed body weights.

Suppose we want to compute a 95 percent confidence interval for the model parameter μ. To do this, we rely on the asymptotic normality of MLEs stated in Eq. (2.3.8). Let $\theta = (\mu, \sigma^2)$. It turns out that $[I(\hat{\theta})]^{-1}$ may be expressed in closed form

$$[I(\hat{\theta})]^{-1} = \begin{bmatrix} \frac{s^2}{n} \cdot \frac{n-1}{n} & 0 \\ 0 & \frac{s^4}{n^2} \cdot \frac{2(n-1)^2}{n} \end{bmatrix}.$$

Therefore, the asymptotic normality of MLEs justifies the following approximation

$$\begin{bmatrix} \hat{\mu} - \mu \\ \hat{\sigma}^2 - \sigma^2 \end{bmatrix} \sim \mathrm{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \frac{s^2}{n} \cdot \frac{n-1}{n} & 0 \\ 0 & \frac{s^4}{n^2} \cdot \frac{2(n-1)^2}{n} \end{bmatrix} \right)$$

which implies that the 95 percent confidence interval for μ may be computed as follows:

$$\bar{y} \pm 1.96 \sqrt{\frac{s^2}{n} \cdot \frac{n-1}{n}}.$$

Applying this formula to the sample of $n = 10$ body weights yields a confidence interval of $[5.913, 8.157]$.

If we had not been able to derive $[I(\hat{\theta})]^{-1}$ in closed form, we still could have computed it by numerical approximation. For example, Panel 2.7 contains the **R** code for computing the MLE of μ and its 95 percent confidence interval. **R** contains several functions for inverting matrices, including `chol2inv` and `solve`. In Panel 2.7 we use the function `chol` to compute the Cholesky decomposition of $I(\hat{\theta})$ and then `chol2inv` to compute its inverse. This procedure is particularly accurate because $I(\hat{\theta})$ is a positive-definite symmetric matrix (by construction).

2.4 BAYESIAN INFERENCE

In this section we describe the Bayesian approach to model-based inference. To facilitate comparisons with classical inference procedures, we will apply the Bayesian approach to some of the same examples used in Section 2.3.

Let $\boldsymbol{y} = (y_1, \ldots, y_n)$ denote a sample of n observations, and suppose we develop an approximating model of $\boldsymbol{y}$ that contains a (possibly vector-valued) parameter θ. As in classical statistics, the approximating model is a formal expression of the processes that are assumed to have produced the observed data. However, in the Bayesian view the model parameter θ is treated as a random variable and the approximating model is elaborated to include a probability distribution for θ that specifies one's beliefs about the magnitude of θ *prior to having observed the data*. This elaboration of the model is therefore called the *prior distribution*.

In the Bayesian view, computing an inference about θ is fundamentally just a probability calculation that yields the probable magnitude of θ given the assumed

```
> y = c(8.51, 4.03, 8.20, 4.19, 8.72, 6.15, 5.40, 8.66, 7.91, 8.58)
>
> neglogLike = function(param) {
+ mu = param[1]
+ sigma = exp(param[2])
+-sum(dnorm(y,mean=mu,sd=sigma, log=TRUE))
+}
>
> fit = optim(par=c(0,0), fn=neglogLike, method='BFGS', hessian=TRUE)
>
> fit$hessian
              [,1]          [,2]
[1,]  3.053826e+00 -1.251976e-05
[2,] -1.251976e-05  2.000005e+01
>
> covMat = chol2inv(chol(fit$hessian))
> covMat
             [,1]         [,2]
[1,] 3.274581e-01 2.049843e-07
[2,] 2.049843e-07 4.999987e-02
>
> mu.mle = fit$par[1]
> mu.se = sqrt(covMat[1,1])
> zcrit = qnorm(.975)
>
> c(mu.mle-zcrit*mu.se, mu.mle+zcrit*mu.se)
[1] 5.913433 8.156571
```

Panel 2.7. R code for computing a 95 percent confidence interval for μ.

prior distribution and given the evidence in the data. To accomplish this calculation, the observed data $\boldsymbol{y}$ are assumed to be *fixed* (once the sample has been obtained), and all inferences about θ are made with respect to the fixed observations $\boldsymbol{y}$. Unlike classical statistics, Bayesian inferences do not rely on the idea of hypothetical repeated samples or on the asymptotic properties of estimators of θ. In fact, probability statements (i.e., inferences) about θ are *exact* for any sample size under the Bayesian paradigm.

2.4.1 Bayes' Theorem and the Problem of 'Inverse Probability'

To describe the principles of Bayesian inference in more concrete terms, it's convenient to begin with some definitions. Let's assume, without loss of generality, that the observed data $\boldsymbol{y}$ are modeled as continuous random variables and that $f(\boldsymbol{y}|\theta)$ denotes the joint pdf of $\boldsymbol{y}$ given a model indexed by the parameter θ. In other words, $f(\boldsymbol{y}|\theta)$ is an approximating model of the data. Let $\pi(\theta)$ denote the pdf of an assumed *prior distribution* of θ. Note that $f(\boldsymbol{y}|\theta)$ provides the probability of the data given θ. However, once the data have been collected the value of $\boldsymbol{y}$ is known; therefore, to compute an inference about θ, we really need the probability of θ given the evidence in the data, which we denote by $\pi(\theta|\boldsymbol{y})$.

Historically, the question of how to compute $\pi(\theta|\boldsymbol{y})$ was called the 'problem of inverse probability.' In the 18th century Reverend Thomas Bayes (1763) provided a solution to this problem, showing that $\pi(\theta|\boldsymbol{y})$ can be calculated to update one's prior beliefs (as summarized in $\pi(\theta)$) using the laws of probability[3]:

$$\pi(\theta|\boldsymbol{y}) = \frac{f(\boldsymbol{y}|\theta)\pi(\theta)}{m(\boldsymbol{y})}, \tag{2.4.1}$$

where $m(\boldsymbol{y}) = \int f(\boldsymbol{y}|\theta)\pi(\theta)\,\mathrm{d}\theta$ denotes the marginal probability of $\boldsymbol{y}$. Eq. (2.4.1) is known as Bayes' theorem (or Bayes' rule), and $\theta|\boldsymbol{y}$ is called the *posterior distribution* of θ to remind us that $\pi(\theta|\boldsymbol{y})$ summarizes one's beliefs about the magnitude of θ *after having observed the data*. Bayes' theorem provides a coherent, probability-based framework for inference because it specifies how prior beliefs about θ can be converted into posterior beliefs in light of the evidence in the data.

Close ties obviously exist between Bayes' theorem and likelihood-based inference because $f(\boldsymbol{y}|\theta)$ is also the basis of Fisher's likelihood function (Section 2.3.1). However, Fisher was vehemently opposed to the 'theory of inverse probability',

[3]Based on the definition of conditional probability, we know $[\theta|\boldsymbol{y}] = [\boldsymbol{y}, \theta]/[\boldsymbol{y}]$ and $[\boldsymbol{y}|\theta] = [\boldsymbol{y}, \theta]/[\theta]$. Rearranging the second equation yields the joint pdf, $[\boldsymbol{y}, \theta] = [\boldsymbol{y}|\theta][\theta]$, which when substituted into the first equation produces Bayes' rule: $[\theta|\boldsymbol{y}] = ([\boldsymbol{y}|\theta][\theta])/[\boldsymbol{y}]$.

as applications of Bayes' theorem were called in his day. Fisher sought inference procedures that did not rely on the specification of a prior distribution, and he deliberately used the term 'likelihood' for $f(\boldsymbol{y}|\theta)$ instead of calling it a probability. Therefore, it is important to remember that although the likelihood function is present in both inference paradigms (i.e., classical and Bayesian), dramatic differences exist in the way that $f(\boldsymbol{y}|\theta)$ is used and interpreted.

2.4.1.1 Example: estimating the probability of occurrence

Let's reconsider the problem introduced in Section 2.3.1.1 of computing inferences about the probability of occurrence ψ. Our approximating model of v, the total number of sample locations where the species is present, is given by the binomial pmf $f(v|\psi)$ given in Eq. (2.3.4). A prior density $\pi(\psi)$ is required to compute inferences about ψ from the posterior density

$$\pi(\psi|v) = \frac{f(v|\psi)\pi(\psi)}{m(v)},$$

where $m(v) = \int_0^1 f(v|\psi)\pi(\psi)\,\mathrm{d}\psi$. It turns out that $\pi(\psi|v)$ can be expressed in closed form if the prior $\pi(\psi) = \mathrm{Be}(\psi|\mathrm{a},\mathrm{b})$ is assumed, where the values of a and b are fixed (by assumption). To be specific, this choice of prior implies that the posterior distribution of ψ is $\mathrm{Be}(\mathrm{a}+v, \mathrm{b}+n-v)$. Thus, the prior and posterior distributions belong to the same class of distributions (in this case, the class of beta distributions). This equivalence, known as *conjugacy*, identifies the beta distribution as the *conjugate prior* for the success parameter of a binomial distribution. We will encounter other examples of conjugacy throughout this book. For now, let's continue with the example.

Suppose we assume prior indifference in the magnitude of ψ. In other words, before observing the data, we assume that all values of ψ are equally probable. This assumption is specified with a $\mathrm{Be}(1,1)$ prior ($\equiv\mathrm{U}(0,1)$ prior) and implies that the posterior distribution of ψ is $\mathrm{Be}(1+v, 1+n-v)$. It's worth noting that the mode of this distribution equals v/n, which is equivalent to the MLE of ψ obtained in a classical, likelihood-based analysis. Now suppose a sample of $n = 5$ locations contains only $v = 1$ occupied site; then the $\mathrm{Be}(2,5)$ distribution, illustrated in Figure 2.3, may be used to compute inferences for ψ. For example, the posterior mean and mode of ψ are 0.29 and 0.20, respectively. Furthermore, we can compute the $\alpha/2$ and $1-\alpha/2$ quantiles of the $\mathrm{Be}(2,5)$ posterior and use these to obtain a $100(1-\alpha)$ percent *credible interval* for ψ. (Bayesians use the term, 'credible interval', to distinguish it from the frequentist concept of a confidence interval.) For example, the 95 percent credible interval for ψ is $[0.04, 0.64]$.

We use this example to emphasize that a Bayesian credible interval and a frequentist confidence interval have completely different interpretations. The Bayesian

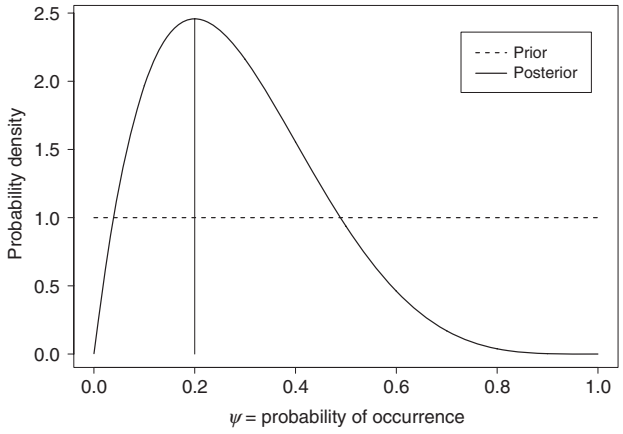

Figure 2.3. Posterior distribution for the probability of occurrence assuming a uniform prior. Vertical line indicates the posterior mode.

credible interval is the result of a probability calculation and reflects our posterior belief in the probable range of ψ values given the evidence in the observed data. Thus, we might choose to summarize the analysis by saying, "the probability that ψ lies in the interval $[0.04, 0.64]$ is 0.95." In contrast, the probability statement associated with a confidence interval corresponds to the proportion of confidence intervals that contain the fixed, but unknown, value of ψ in an infinite sequence of hypothetical, repeated samples (see Section 2.3.2). The frequentist's interval therefore requires considerably more explanation and is far from a direct statement of probability. Unfortunately, the difference in interpretation of credible intervals and confidence intervals is often ignored in practice, much to the consternation of many statisticians.

2.4.2 Pros and Cons of Bayesian Inference

Earlier we mentioned that one of the virtues of Bayesian inference is that probability statements about θ are *exact* for any sample size. This is especially meaningful when one considers that a Bayesian analysis yields the entire posterior pdf of θ, $\pi(\theta|\boldsymbol{y})$, as opposed to a single point estimate of θ. Therefore, in addition to computing summaries of the posterior, such as its mean $\mathrm{E}(\theta|\boldsymbol{y})$ or variance $\mathrm{Var}(\theta|\boldsymbol{y})$, *any function* of θ can be calculated while accounting for all of the posterior uncertainty in θ. The benefits of being able to manage errors in estimation in this way are especially evident in computing inferences for latent parameters of hierarchical

models, as we will illustrate in Section 2.6, or in computing predictions that depend on the estimated value of θ.

Specification of the prior distribution may be perceived as a benefit or as a disadvantage of the Bayesian mode of inference. In scientific problems where prior information about θ may exist or can be elicited (say, from expert opinion), Bayes' theorem reveals precisely how such information may be used when computing inferences for θ. In other (or perhaps most) scientific problems, little may be known about the probable magnitude of θ in advance of an experiment or survey. In these cases an *objective* approach would be to use a prior that places equal (or nearly equal) probability on all values of θ. Such priors are often called 'vague' or 'non-informative.' A problem with this approach is that priors are not invariant to transformation of the parameters. In other words a prior that is 'non-informative' for θ can be quite informative for $g(\theta)$, a one-to-one transformation of θ.

One solution to this problem is to develop a prior that is both non-informative *and* invariant to transformation of its parameters. A variety of such 'objective priors', as they are currently called (see Chapter 5 of Ghosh et al. (2006)), have been developed for models with relatively few parameters. Objective priors are often improper (that is, $\int \pi(\theta)\,\mathrm{d}\theta = \infty$); therefore, if an objective prior is to be used, the analyst must prove that that the resulting posterior distribution is proper (that is, $\int f(\boldsymbol{y}|\theta)\pi(\theta)\,\mathrm{d}\theta < \infty$). Such proofs often require considerable mathematical expertise, particularly for models that contain many parameters.

A second solution to the problem of constructing a non-informative prior is to identify a particular parameterization of the model for which a uniform (or nearly uniform) prior makes sense. Of course, this approach is possible only if we are able to assign scientific relevance and context to the model's parameters. We have found this approach to be useful in the analysis of ecological data, and we use this approach throughout the book.

Specification of the prior distribution can be viewed as the 'price' paid for the exactness of inferences computed using Bayes' theorem. When the sample size is low, the price of an exact inference may be high. As the size of a sample increases, the price of an exact inference declines because the information in the data eventually exceeds the information in the prior. We will return to this tradeoff in the next section, where we describe some asymptotic properties of posteriors.

2.4.3 Asymptotic Properties of the Posterior Distribution

We have noted already that the Bayesian approach to model-based inference has several appealing characteristics. In this section we describe additional features that are associated with computing inferences from large samples.

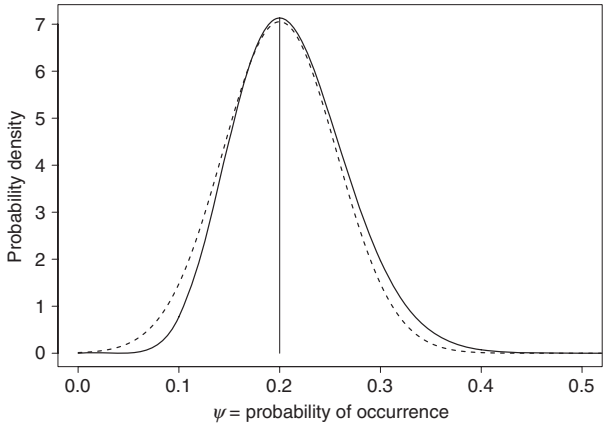

Figure 2.4. A normal approximation (dashed line) of the posterior distribution of the probability of occurrence (solid line). Vertical line indicates the posterior mode.

2.4.3.1 Approximate normality

Let $[\theta|\boldsymbol{y}]$ denote the posterior distribution of θ given an observed set of data $\boldsymbol{y}$. If a set of 'regularity conditions' that have to do with technical details, such as identifiability of the model's parameters and differentiability of the posterior density function $\pi(\theta|\boldsymbol{y})$, are satisfied, we can prove that as sample size $n \to \infty$,

$$(\theta - \hat{\theta}) \mid \boldsymbol{y} \sim \mathrm{N}(0, [I(\hat{\theta})]^{-1}), \tag{2.4.2}$$

where $\hat{\theta}$ is the posterior mode and $I(\hat{\theta}) = -\frac{\partial^2 \log \pi(\theta|\boldsymbol{y})}{\partial\theta\partial\theta}|_{\theta=\hat{\theta}}$ is called the *generalized observed information* (Ghosh et al., 2006). The practical utility of this limiting behavior is that the posterior distribution of θ can be approximated by a normal distribution $\mathrm{N}(\hat{\theta}, [I(\hat{\theta})]^{-1})$ if n is sufficiently large. In other words, when n is large, we can expect the posterior to become highly concentrated around the posterior mode $\hat{\theta}$.

Example: estimating the probability of occurrence

Recall from Section 2.4.1.1 that the posterior mode for the probability of occurrence was $\hat{\psi} = v/n$ when a $\mathrm{Be}(1,1)$ prior was assumed for ψ. It is easily proved that $[I(\hat{\psi})]^{-1} = \hat{\psi}(1-\hat{\psi})/n$ given this choice of prior; therefore, according to Eq. (2.4.2) we can expect a $\mathrm{N}(\hat{\psi}, \hat{\psi}(1-\hat{\psi})/n)$ distribution to approximate the true posterior, a $\mathrm{Be}(1+v, 1+n-v)$ distribution, when n is sufficiently large. Figure 2.4 illustrates that the approximation holds very well for a sample of $n = 50$ locations, of which $z = 10$ are occupied.

The asymptotic normality of the posterior (indicated in Eq. (2.4.2)) is an important result because it establishes formally that the relative importance of the prior distribution must decrease with an increase in sample size. To see this, note that $I(\theta)$ is the sum of two components, one due to the likelihood function $f(\boldsymbol{y}|\theta)$ and another due to the prior density $\pi(\theta)$:

$$\begin{aligned} I(\theta) &= -\frac{\partial^2 \log \pi(\theta|\boldsymbol{y})}{\partial\theta\partial\theta} \\ &= -\frac{\partial^2 \log f(\boldsymbol{y}|\theta)}{\partial\theta\partial\theta} - \frac{\partial^2 \log \pi(\theta)}{\partial\theta\partial\theta}. \end{aligned} \tag{2.4.3}$$

As n increases, only the magnitude of the first term on the right-hand side of Eq. (2.4.3) increases, whereas the magnitude of the second term, which quantifies the information in the prior, remains constant. An important consequence of this result is that we can expect inferences to be insensitive to the choice of prior if we have enough data. On the other hand, if the sample size is relatively small, the prior distribution may be a critical part of the model specification.

2.4.4 Modern Methods of Bayesian Computation

Thus far, we have illustrated the Bayesian approach to inference using a rather simple model (binomial likelihood and conjugate prior), where the posterior density function $\pi(\theta|\boldsymbol{y})$ could be expressed in closed form. However, in many (perhaps most) cases of scientific interest, an approximating model of the data will be more complex, often involving many parameters and multiple levels of parameters. In such cases it is often difficult or impossible to calculate the normalizing constant $m(\boldsymbol{y})$ accurately because the calculation requires a p-dimensional integration if the model contains p distinct parameters. Therefore, the posterior density is often known only up to a constant of proportionality

$$\pi(\theta|\boldsymbol{y}) \propto f(\boldsymbol{y}|\theta)\pi(\theta).$$

This computational impediment is one the primary reasons why the Bayesian approach to inference was not widely used prior to the late 20th century. Opposition by frequentists, many of whom strongly advocated classical inference procedures, is another reason. In 1774, Pierre Simon Laplace developed a method for computing a large-sample approximation of the normalizing constant $m(\boldsymbol{y})$, but this procedure is applicable only in cases where the unnormalized posterior density is a smooth function of θ with a sharp maximum at the posterior mode (Laplace, 1986). Refinements of Laplace's method have been developed (Ghosh et al., 2006,

Section 4.3.2), but these refinements, as with Laplace's method, lack generality in their range of application.

A recent upsurge in the use of Bayesian inference procedures can be attributed to the widespread availability of fast computers and to the development of efficient algorithms, known collectively as *Markov chain Monte Carlo* (MCMC) samplers. These include the Gibbs sampler, the Metropolis–Hastings algorithm, and others (Robert and Casella, 2004). The basic idea behind these algorithms is to compute an arbitrarily large sample from the posterior distribution $\theta|\boldsymbol{y}$ without actually computing its normalizing constant $m(\boldsymbol{y})$. Given an arbitrarily large sample of the posterior, the posterior density $\pi(\theta|\boldsymbol{y})$ can be approximated quite accurately (say, using a histogram or a kernel-density smoother). In addition, any function of the posterior, such as the marginal mean, median, or standard deviation for an individual component of θ, can be computed quite easily without actually evaluating the integrals implied in the calculation.

2.4.4.1 Gibbs sampling

One of the most widely used algorithms for sampling posterior distributions is called the *Gibbs sampler*. This algorithm is a special case of the *Metropolis–Hastings* algorithm, and the two are often used together to produce efficient hybrid algorithms that are relatively easy to implement.

The basic idea behind Gibbs sampling (and other MCMC algorithms) is to produce a random sample from the *joint* posterior distribution of ≥ 2 parameters by drawing random samples from a sequence of *full conditional* posterior distributions. The motivation for this idea is that while it may be difficult or impossible to draw a sample directly from the joint posterior, drawing a sample from the full conditional distribution of each parameter is often a relatively simple calculation.

Let's illustrate the Gibbs sampler for a model that includes 3 parameters: θ_1, θ_2, and θ_3. Given a set of data $\boldsymbol{y}$, we wish to compute an arbitrarily large sample from the joint posterior $[\theta_1, \theta_2, \theta_3|\boldsymbol{y}]$. We assume that the full-conditional distributions

$$[\theta_1|\theta_2, \theta_3, \boldsymbol{y}]$$
$$[\theta_2|\theta_1, \theta_3, \boldsymbol{y}]$$
$$[\theta_3|\theta_1, \theta_2, \boldsymbol{y}]$$

of the model's parameters are relatively easy to sample. To begin the Gibbs sampler, we assign an arbitrary set of initial values to each parameter, i.e., $\theta_1 = \theta_1^{(0)}$, $\theta_2 = \theta_2^{(0)}$, $\theta_3 = \theta_3^{(0)}$, where the superscript in parentheses denotes the order in the sequence of random draws. The Gibbs sampling algorithm proceeds as follows:

Step 1 Draw $[\theta_1^{(1)} \sim \theta_1|\theta_2^{(0)}, \theta_3^{(0)}, \boldsymbol{y}]$.

Step 2 Draw $\theta_2^{(1)} \sim [\theta_2|\theta_1^{(1)}, \theta_3^{(0)}, \boldsymbol{y}]$.

Step 3 Draw $\theta_3^{(1)} \sim [\theta_3|\theta_1^{(1)}, \theta_2^{(1)}, \boldsymbol{y}]$.

This completes one iteration of the Gibbs sampler and generates a new set of parameter values, say $\boldsymbol{\theta} = \boldsymbol{\theta}^{(1)}$ where $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)$. Steps 1 to 3 are then repeated using the values of $\boldsymbol{\theta}$ from the previous iteration to obtain $\boldsymbol{\theta}^{(2)}, \boldsymbol{\theta}^{(3)}, \ldots$. Typically several thousand iterations of the Gibbs sampler will be required to obtain an accurate sample from the joint posterior $\boldsymbol{\theta}|\boldsymbol{y}$.

The Gibbs sampling algorithm produces a Markov chain, $\boldsymbol{\theta}^{(0)}, \boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \boldsymbol{\theta}^{(3)}, \ldots$, whose stationary distribution is equivalent to the joint posterior $\boldsymbol{\theta}|\boldsymbol{y}$, provided a set of technical, regularity conditions are satisfied. Consequently, to obtain a sample from $\boldsymbol{\theta}|\boldsymbol{y}$, the beginning of the Markov chain, which is often called the *burn-in*, is typically discarded. To obtain a random sample of approximately independent draws from $\boldsymbol{\theta}|\boldsymbol{y}$, the remainder of the Markov chain must be subsampled (say, choosing every kth draw where $k > 1$) because successive draws $\boldsymbol{\theta}^{(t)}$ and $\boldsymbol{\theta}^{(t+1)}$ are not independent. Alternatively, one may retain only the last draw of the Markov chain and repeat the entire process n times (to obtain a posterior sample of size n) by using n different starting values for $\boldsymbol{\theta}^{(0)}$. This alternative procedure, though originally proposed by Gelfand and Smith (1990), is seldom used or even necessary. An extensive literature is devoted to assessing the convergence of Markov chains to stationarity. Because proofs of convergence are difficult to establish for complex models, a variety of diagnostics have been developed to assess convergence empirically by subsampling a few independently initialized Markov chains (Robert and Casella, 2004, Chapter 12). Contemporary Bayesian analyses routinely use such diagnostics to assess whether an MCMC sampler has been run long enough to produce an accurate sample from the joint posterior.

2.4.4.2 Example: estimating survival of larval fishes

We extend an example described by Arnold (1993) to illustrate the Gibbs sampler. Suppose an experiment is conducted to estimate the daily probability of survival for a single species of larval fish. In the experiment larval fishes are added to each of $n = 5$ replicate containers by pouring the contents (i.e., water and fishes) of a common source into each container. In practice, the contents of the source are carefully mixed, and equal volumes are added to each of the 5 replicate containers. This procedure is used because handling of individual fishes is thought to reduce their survival.

An important aspect of the experiment is that the number of fishes added to each replicate container is not known precisely, and only the survivors in each replicate can be enumerated. Fortunately, the number of fishes in the source

container is known to be 250, so we may assume that the average number of fishes initially added to each replicate container is 50 (=250/5). At the end of a 24-hour incubation period, the numbers of surviving fishes in the 5 replicate containers are $\boldsymbol{y} = (15, 11, 12, 5, 12)$.

To formulate a model of these counts, we define a latent parameter N_i for the number of fishes initially added to the ith replicate container; then we assume $y_i|N_i, \phi \sim \text{Bin}(N_i, \phi)$, where ϕ denotes the daily probability of survival of each fish. Given the design of the experiment, it is reasonable to assume $N_i \sim \text{Po}(\lambda)$, where $\lambda = 50$ denotes the average number of fishes initially added to each replicate container. These modeling assumptions imply that the marginal pmf of the observed counts can be expressed in closed form because

$$\begin{aligned}[y_i|\phi] &= \sum_{N_i=y_i}^{\infty} \text{Bin}(y_i|N_i, \phi)\ \text{Po}(N_i|\lambda) \\ &= \text{Po}(y_i|\lambda\phi)\end{aligned}$$

(proof omitted). To complete the model for a Bayesian analysis, we assume a non-informative $\text{Be}(a, b)$ prior for ϕ (wherein $a = b = 1$) and independence among replicate observations. These assumptions yield the following posterior density function

$$[\phi|\boldsymbol{y}] = \frac{(\prod_{i=1}^{n}[y_i|\phi])\,\text{Be}(\phi|a, b)}{m(\boldsymbol{y})}, \tag{2.4.4}$$

where $m(\boldsymbol{y})$ is the normalizing constant. After a bit of algebra, it can be shown that the posterior pdf in Eq. (2.4.4) is equivalent to

$$[\phi|\boldsymbol{y}] = c^{-1}\phi^{a-1+\sum_{i=1}^{n} y_i}(1-\phi)^{b-1}\exp(-n\lambda\phi), \tag{2.4.5}$$

where

$$c = \int_0^1 \phi^{a-1+\sum_{i=1}^{n} y_i}(1-\phi)^{b-1}\exp(-n\lambda\phi)\,\mathrm{d}\phi \tag{2.4.6}$$

is the constant of integration needed to make $[\phi|\boldsymbol{y}]$ a proper probability density function. However, the definite integral in Eq. (2.4.6) cannot be evaluated in closed form; therefore, c must be computed by numerical integration.

In this example c can be computed quite accurately because the integral is only one-dimensional. However, as an alternative, let's examine how Gibbs sampling may be used to compute a sample from the marginal posterior of $\phi|\boldsymbol{y}$. Instead of eliminating $\boldsymbol{N} = (N_1, \ldots, N_n)$ from the model, we will use the Gibbs algorithm to compute a sample from the joint posterior distribution of $\boldsymbol{N}$ and ϕ; then we simply ignore the simulated values of $\boldsymbol{N}$ to obtain a sample from $\phi|\boldsymbol{y}$. In this

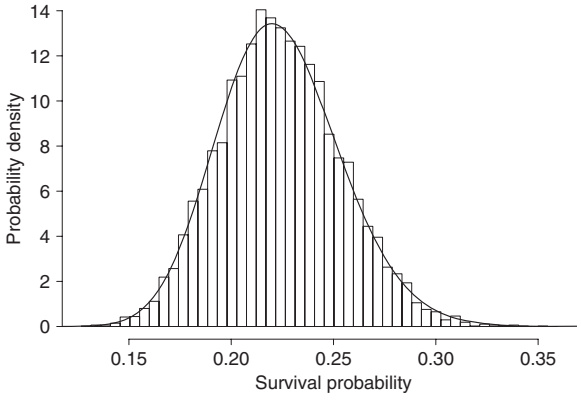

Figure 2.5. Posterior distribution of survival probability ϕ (solid line) and an approximation of the distribution obtained by Gibbs sampling (histogram).

way, the marginal posterior distribution of ϕ is obtained implicitly without actually integrating the joint posterior density $[\boldsymbol{N}, \phi|\boldsymbol{y}]$.

The full conditional distributions needed to compute a Gibbs sample of the joint posterior $[\boldsymbol{N}, \phi|\boldsymbol{y}]$ are given by

$$\begin{aligned} \phi|\boldsymbol{N}, \boldsymbol{y} &\sim \text{Be}\left(a + \sum_{i=1}^{n} y_i, b + \sum_{i=1}^{n} N_i - y_i\right) \\ N_i|\phi, y_i &\sim y_i + \text{Po}(\lambda(1-\phi)) \quad i = 1, \ldots, n \end{aligned}$$

both of which are standard distributions that are easy to sample. In fact, we computed a sample of 10000 random draws from the joint posterior by running the Gibbs sampler for 100000 iterations and then retaining every fifth draw of the last 50000 iterations. A histogram of the simulated values of ϕ provides an excellent approximation of the posterior density computed using Eq. (2.4.5) (Figure 2.5).

Now suppose the full conditional distributions had been more difficult to sample or that we simply did not want to bother with developing an efficient algorithm. In either case we might choose to use the **WinBUGS** software, which provides an implementation of the Gibbs sampler that does not require the user to specify a set of full conditional distributions. In fact, all that **WinBUGS** requires is a statement of the distributional assumptions used in the model. For example, the section of Panel 2.8 labeled `native WinBUGS code` is all that **WinBUGS** needs to compute the joint posterior of $\boldsymbol{N}$ and ϕ in our example. We prefer to access **WinBUGS** remotely while working in **R**. The remaining sections of Panel 2.8 provide **R** code for assigning values to $\boldsymbol{y}$ and to the prior parameters and for using

the **R** library, **R2WinBUGS** (Sturtz et al., 2005), to direct the calculations. We will use **WinBUGS** extensively throughout the book to illustrate how Bayesian methods may be used to fit models of varying complexity.

2.5 HYPOTHESIS TESTING

The development of formal rules for testing scientific hypotheses has a long history in classical statistics. The application of these rules is especially useful in making evidentiary conclusions (inferences) from *designed experiments*, where the scientist exercises some degree of control over the relevant sources of uncertainty in the observed outcomes. In a well-designed experiment, hypothesis testing can be used to establish *causal relationships* between experimental outcomes and the systematic sources of variation in those outcomes that are manipulated as part of the design.

Hypothesis testing also can be applied in *observational studies* (surveys), where the scientist does *not* have direct control over the sources of uncertainty being tested. In such studies hypothesis testing may be used to assess whether estimated levels of association (or correlation) between one or more observable outcomes are 'statistically significant' (i.e., are unlikely to have occurred by chance given a particular significance level α of the test). However, in observational studies hypothesis testing cannot be used to determine whether a significant association between outcomes is the result of a coincidence of events or of an underlying causal relationship. Therefore, it can be argued that hypothesis testing is more useful scientifically in the analysis of designed experiments.

We find hypothesis testing to be useful because it provides a general context for the description of some important inference problems, including model selection, construction of confidence intervals, and assessment of model adequacy. We describe and illustrate these topics in the following subsections.

2.5.1 Model Selection

In our model-based approach to statistical inference, the classical problem of testing a scientific hypothesis is equivalent to the problem of selecting between two *nested* models of the data. By nested, we mean that the parameters of one model are a restricted set of the parameters of the other model. To illustrate, suppose we are interested in selecting between two linear regression models, one which simply contains an intercept parameter α and another which contains both intercept and slope parameters, α and β, respectively. By defining the parameter vector $\theta = (\alpha, \beta)$, we can specify the restricted model as $\theta = (\alpha, 0)$ and the full model as

```
--------------data-------------------
> y=c(15,11,12,5,12)
> lambda=50
> a=1
> b=1
--------------arguments for R2WinBUGS--------------
> data = list(n=length(y), y=y, lambda=lambda, a=a, b=b)
> params = list('phi','N')
>
> inits = function() {
+  phi = rbeta(1,a,b)
+  N = rpois(length(y),lambda)
+  list(phi=phi, N=N)
+   }
>
--------------native WinBUGS code------------
> sink('MarginalDensity.txt')
> cat('
 model {
  phi ~ dbeta(a,b)
  for (i in 1:n) {
      N[i] ~ dpois(lambda)
      y[i] ~ dbin(phi, N[i])
  }
  }
   ', fill=TRUE)
> sink()
>
--------------call bugs() to fit model-----------------
> library(R2WinBUGS)
> fit = bugs(data, inits, params,
  model.file='MarginalDensity.txt',
  debug=F, n.chains=1, n.iter=100000, n.burnin=50000, n.thin=5)
>
> phi = fit$sims.matrix[,'phi']
>
> summary(phi)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.1198  0.2033  0.2227  0.2242  0.2435  0.3541
```

Panel 2.8. **R** and **WinBUGS** code for sampling the posterior distribution of survival probability ϕ.

$\theta = (\alpha, \beta)$. Therefore, a decision to select the full model over the restricted model is equivalent to rejecting the null hypothesis that $\beta = 0$ in favor of the alternative hypothesis that $\beta \neq 0$, given that α is present in both models.

The connection between model selection and hypothesis testing can be specified quite generally. Let H_0 and H_1 denote two complementary hypotheses which represent the *null* and *alternative* hypotheses of a testing problem. In addition, assume that both of these hypotheses can be specified in terms of a (possibly vector-valued) parameter θ, which lies in the parameter space Θ. Given these definitions, the null and alternative hypotheses can be represented as follows:

$$H_0 : \theta \in \Theta_0$$
$$H_1 : \theta \in \Theta_0^c,$$

where Θ_0 is a subset of the parameter space Θ and Θ_0^c is the complement of Θ_0. The model selection problem is basically equivalent to a two-sided hypothesis test wherein

$$H_0 : \theta = \theta_0$$
$$H_1 : \theta \neq \theta_0.$$

Thus, if we reject H_0, we accept the more complex model for which $\theta \neq \theta_0$; otherwise, we accept the simpler model for which $\theta = \theta_0$.

In classical statistics the decision to accept or reject H_0 is based on the asymptotic distribution of a test statistic. Although a variety of test statistics have been developed, we describe only two, the Wald statistic and the likelihood-ratio statistic, because they are commonly used in likelihood-based inference.

2.5.1.1 Wald test

The Wald test statistic is derived from the asymptotic normality of MLEs. Recall from Eq. (2.3.8) that the distribution of the discrepancy $\hat{\theta} - \theta$ is asymptotically normal

$$(\hat{\theta} - \theta) \mid \theta \sim \mathrm{N}(0, [I(\hat{\theta})]^{-1}).$$

Suppose we have a single-parameter model and want to test $H_0 : \theta = \theta_0$. Under the assumptions of the null model, the asymptotic normality of $\hat{\theta}$ implies

$$[I(\hat{\theta})]^{1/2}(\hat{\theta} - \theta_0) \mid \theta_0 \sim \mathrm{N}(0, 1)$$

or, equivalently,

$$\frac{\hat{\theta} - \theta_0}{\mathrm{SE}(\hat{\theta})} \sim \mathrm{N}(0, 1), \tag{2.5.1}$$

where $\text{SE}(\hat{\theta}) = [I(\hat{\theta})]^{-1/2}$ is the asymptotic standard error of $\hat{\theta}$ computed by fitting the alternative model H_1. The left-hand side of Eq. (2.5.1) is called the Wald test statistic and is often denoted by z. Based on Eq. (2.5.1), we reject H_0 when $|z| > z_{1-\alpha/2}$, where $z_{1-\alpha/2}$ denotes the $(1-\alpha/2)$ quantile of a standard normal distribution.

The distribution of the square of a standard normal random variable is a chi-squared distribution with 1 degree of freedom, which we denote by $\chi^2(1)$; therefore, an equivalent test of H_0 is to compare z^2 to the $(1-\alpha)$ quantile of a $\chi^2(1)$ distribution. We mention this only to provide a conceptual link to tests of hypotheses that involve multiple parameters. For example, if several parameters are held fixed under the assumptions of H_0, the multi-parameter version of the Wald test statistic and its asymptotic distribution are

$$(\hat{\theta} - \theta_0)' I(\hat{\theta})(\hat{\theta} - \theta_0) \sim \chi^2(\nu),$$

where ν denotes the rank of $I(\hat{\theta})$ (i.e., the number of parameters to be estimated under the alternative model H_1).

2.5.1.2 Likelihood ratio test

The likelihood ratio test is rooted in the notion that the likelihood function $L(\theta|\boldsymbol{y})$ provides a measure of *relative* support for different values of the parameter θ. Therefore, in the model selection problem the ratio

$$\Lambda = \frac{L(\hat{\theta}_0|\boldsymbol{y})}{L(\hat{\theta}|\boldsymbol{y})} \tag{2.5.2}$$

provides the ratio of likelihoods obtained by computing the MLE of θ_0 (the parameters of the model associated with H_0) and the MLE of θ (the parameters of the model associated with H_1). Because θ_0 is a restricted version of θ, $L(\hat{\theta}|\boldsymbol{y}) > L(\hat{\theta}_0|\boldsymbol{y})$ (by definition), the likelihood ratio must be a fraction (i.e., $0 < \Lambda < 1$). Thus, lower values of Λ lend greater support to H_1.

Under the assumptions of the null model H_0, the asymptotic distribution of the statistic $-2\log\Lambda$ is chi-squared with ν degrees of freedom

$$-2\log(\Lambda) \sim \chi^2(\nu), \tag{2.5.3}$$

where ν equals the difference in the number of free parameters to be estimated under H_0 and H_1. The left-hand side of Eq. (2.5.3), which is called the *likelihood ratio statistic*, is strictly positive. In practice, we reject H_0 for values of $-2\log\Lambda$ that exceed $\chi^2_{1-\alpha}(\nu)$, the $(1-\alpha)$ quantile of a chi-squared distribution with ν degrees of freedom.

The likelihood ratio test can be used to evaluate the *goodness of fit* of a model of counts provided the sample is sufficiently large. In this context H_1 corresponds to a 'saturated' model in which the number of parameters equals the sample size n. We cannot learn anything new from a saturated model because its parameters essentially amount to a one-to-one transformation of the counts $\boldsymbol{y}$; however, a likelihood ratio comparison between the saturated model and an approximating model H_0 *can* be used to assess the goodness of fit of H_0. For example, suppose the approximating model contains k free parameters to be estimated; then, the value of the likelihood ratio statistic $-2\log\Lambda$ can be compared to $\chi^2_{1-\alpha}(n-k)$ to determine whether H_0 is accepted at the α significance level. If H_0 is accepted, we may conclude that the approximating model provides a satisfactory fit to the data. In the context of this test the likelihood ratio statistic provides a measure of discrepancy between the counts in $\boldsymbol{y}$ and the approximating model's estimate of $\boldsymbol{y}$; consequently, $-2\log\Lambda$ is often called the *deviance* test statistic, or simply the deviance, in this setting. We will see many applications of the deviance test statistic in later chapters.

2.5.1.3 Example: Mortality of moths exposed to cypermethrin

We illustrate model selection and hypothesis testing using data observed in a dose-response experiment involving adults of the tobacco budworm (*Heliothis virescens*), a moth species whose larvae are responsible for damage to cotton crops in the United States and Central and South America (Collett, 1991, Example 3.7). In the experiment, batches of 20 moths of each sex were exposed to a pesticide called cypermethrin for a period of 72 hours, beginning two days after the adults had emerged from pupation. Both sexes were exposed to the same range of pesticide doses: 1, 2, 4, 8, 16, and 32 μg cypermethrin. At the end of the experiment the number of moths in each batch that were either knocked down (movement of moth was uncoordinated) or dead (moth was unable to move and was unresponsive to a poke from a blunt instrument) was recorded.

The experiment was designed to test whether males and females suffered the same mortality when exposed to identical doses of cypermethrin. The results are shown in Figure 2.6 where the empirical logit of the proportion of moths that died in each batch is plotted against $\log_2(\text{dose})$, which linearizes the exponential range of doses. The *empirical logit*, which is defined as follows

$$\log\left(\frac{y_i + 0.5}{N - y_i + 0.5}\right)$$

(wherein y_i $(i = 1, \ldots, 12)$ denotes the number of deaths observed in the ith batch of $N = 20$ moths per batch), is the least biased estimator of the true logit of the proportion of deaths per batch (Agresti, 2002). We use the empirical logit because

it allows outcomes of 100 percent mortality ($y = N$) or no mortality ($y = 0$) to be plotted on the logit scale along with the other outcomes of the experiment.

The empirical logits of mortality appear to increase linearly with $\log_2(\text{dose})$ for both sexes (Figure 2.6); therefore, we consider logistic regression models as a reasonable set of candidates for finding an approximating model of the data. Let x_i denote the $\log_2(\text{dose})$ of cypermethrin administered to the ith batch of moths that contained either males ($z_i = 1$) or females ($z_i = 0$). A logistic-regression model containing 3 parameters is

$$\begin{aligned} y_i|N, p_i &\sim \text{Bin}(N, p_i) \\ \text{logit}(p_i) &= \alpha + \beta x_i + \gamma z_i, \end{aligned}$$

where α is the intercept, β is the effect of cypermethrin and γ is the effect of sex.

Let $\theta = (\alpha, \beta, \gamma)$ denote a vector of parameters. The relevant hypotheses to be examined in the experiment are

$$\begin{aligned} H_0 &: \theta = (\alpha, \beta, 0) \\ H_1 &: \theta = (\alpha, \beta, \gamma), \end{aligned}$$

where $\gamma \neq 0$ in H_1. In other words, the test of H_0 amounts to selecting between two models of the data: the null model, wherein only α and β are estimated, and the alternative model, wherein all 3 parameters are estimated.

To conduct a Wald test of H_0, we need only fit the alternative model, which yields the MLE $\hat{\theta} = (-3.47, 1.06, 1.10)$. We obtain $\text{SE}(\hat{\gamma}) = 0.356$ from the inverse of the observed information matrix, and we compute a Wald test statistic of $z = (\hat{\gamma} - 0)/\text{SE}(\hat{\gamma}) = 3.093$. Because $|z| > 1.96$, we reject H_0 at the 0.05 significance level and select the alternative model of the data in favor of the null model.

To conduct a likelihood ratio test of H_0, we must compute MLEs for the parameters of both null and alternative models. Using these estimates, we obtain $\log L(\hat{\alpha}, \hat{\beta}|\boldsymbol{y}) = -23.547$ for the null model and $\log L(\hat{\alpha}, \hat{\beta}, \hat{\gamma}|\boldsymbol{y}) = -18.434$ for the alternative model. Therefore, the likelihood ratio statistic is $-2\log(\Lambda) = -2\{\log L(\hat{\alpha}, \hat{\beta}|\boldsymbol{y}) - \log L(\hat{\alpha}, \hat{\beta}, \hat{\gamma}|\boldsymbol{y})\} = 10.227$. The number of parameters estimated under the null and alternative models differ by $\nu = 3 - 2 = 1$; therefore, to test H_0 we compare the value of the likelihood ratio statistic to $\chi^2_{0.95}(1) = 3.84$. Since $10.227 > 3.84$, we reject H_0 at the 0.05 significance level and select the alternative model of the data in favor of the null model.

The null hypothesis is rejected regardless of whether we use the Wald test or the likelihood ratio test; therefore, we may conclude that the difference in mortality of male and female moths exposed to the same dose of cypermethrin in the experiment is statistically significant, a result which certainly appears to be supported by the data in Figure 2.6.

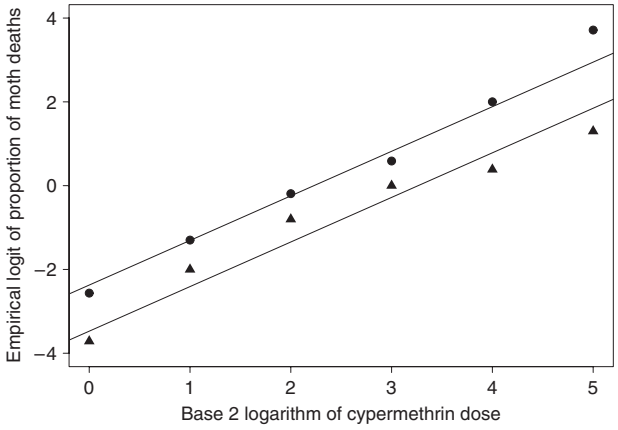

Figure 2.6. Mortality of male (circle) and female (triangle) moths exposed to various doses of cypermethrin. Lines indicate the fit of a logistic regression model with dose ($\log_2$ scale) and sex as predictors.

We may also use a likelihood ratio test to assess the goodness of fit of the model that we have selected as a basis for inference. In this test the parameter estimates of the alternative ('saturated') model H_1 correspond to the observed proportions of moths that died in the 12 experimental batches, i.e., $\hat{\theta} = (\hat{p}_1, \hat{p_2}, \ldots, \hat{p}_{12}) = (y_1/N, y_2/N, \ldots, y_{12}/N)$. For this model $\log L(\hat{p}_1, \ldots, \hat{p}_{12}|\boldsymbol{y}) = -15.055$; therefore the likelihood ratio statistic for testing goodness of fit is $-2\{\log L(\hat{\alpha}, \hat{\beta}, \hat{\gamma}|\boldsymbol{y}) - \log L(\hat{p}_1, \ldots, \hat{p}_{12}|\boldsymbol{y}) = 6.757\}$. To test H_0, we compare this value to $\chi^2_{0.95}(9) = 16.92$. Since $6.757 \leq 16.92$, we accept H_0 and conclude that the model with parameters (α, β, γ) cannot be rejected for lack of fit at the 0.05 significance level.

2.5.2 Inverting Tests to Estimate Confidence Intervals

In many studies a formal test of the statistical significance of an effect (e.g., a treatment effect in a designed experiment) is less important scientifically than an estimate of the magnitude of the effect. This is particularly true in observational studies where the estimated level of association between one or more observable outcomes is of primary scientific interest. In these cases the main inference problem is to estimate a parameter and to provide a probabilistic description of the uncertainty in the estimate.

In classical statistics the solution to this problem involves the construction of a confidence interval for the parameter. However, a variety of procedures have been developed for constructing confidence intervals. For example, in Section 2.3.2

we described how the asymptotic normality of MLEs provides a $(1-\alpha)$ percent confidence interval of the form

$$\hat{\theta} \pm z_{1-\alpha/2}\,\mathrm{SE}(\hat{\theta}) \tag{2.5.4}$$

for a scalar-valued parameter θ wherein $\mathrm{SE}(\hat{\theta}) = [I(\hat{\theta})]^{-1/2}$ (cf. Eq. (2.3.9)). It turns out that this confidence interval may be constructed by inverting the Wald test described in Section 2.5.1. To see this, note that the null hypothesis $H_0 : \theta = \theta_0$ is accepted if

$$\left| \frac{\hat{\theta} - \theta_0}{\mathrm{SE}(\hat{\theta})} \right| \leq z_{1-\alpha/2}.$$

Simple algebra can be used to prove that the range of θ_0 values that satisfy this inequality are bounded by the confidence limits given in Eq. (2.5.4); consequently, there is a direct correspondence between the acceptance region of the null hypothesis (that is, the values of θ for which $H_0 : \theta = \theta_0$ is accepted) and the $(1-\alpha)$ percent confidence interval for θ.

Confidence intervals also may be constructed by inverting the likelihood ratio test described in Section 2.5.1. To see this, note that the null hypothesis $H_0 : \theta = \theta_0$ is accepted if

$$-2\{\log L(\hat{\theta}_0|\boldsymbol{y}) - \log L(\hat{\theta}|\boldsymbol{y})\} \leq \chi^2_{1-\alpha}(\nu). \tag{2.5.5}$$

Therefore, the range of the fixed parameters in $\hat{\theta}_0$ that satisfy this inequality provide a $(1-\alpha)\%$ confidence region for those parameters. Such confidence regions are often more difficult to calculate than those based on inverting the Wald test because the free parameters in θ_0 must be estimated by maximizing $L(\theta_0|\boldsymbol{y})$ for each value of the parameters in θ_0 that are fixed. For this reason $L(\theta_0|\boldsymbol{y})$ is called the *profile likelihood* function of θ_0. If a confidence interval is required for only a single parameter, a numerical root-finding procedure may be used to calculate the values of that parameter that satisfy Eq. (2.5.5).

In sufficiently large samples, confidence intervals computed by inverting the Wald test or the likelihood ratio test will be nearly identical. An obvious advantage of intervals based on the Wald test is that they are easy to compute. However, in small samples, such intervals can produce undesirable results, as we observed in Table 2.3, where an interval for the probability of occurrence ψ includes negative values. Intervals based on inverting the likelihood ratio test can be more difficult to calculate, but an advantage of these intervals is that they are *invariant to transformation* of a model's parameters. Therefore, regardless of whether we parameterize the probability of occurrence in terms of ψ or $\mathrm{logit}(\psi)$, the confidence intervals for ψ will be the same. Table 2.4 illustrates the small-sample benefits of computing intervals for ψ by inverting the likelihood ratio test.

Table 2.4. Comparison of 95 percent confidence intervals for ψ based on inverting a Wald test and a likelihood ratio test. All intervals have the same MLE ($\hat{\psi} = 0.2$), but sample size n differs.

n	Wald test	Likelihood ratio test
5	[-0.15, 0.55]	[0.01, 0.63]
50	[0.09, 0.31]	[0.11, 0.32]

2.5.2.1 Example: Mortality of moths exposed to cypermethrin

As an additional comparison of procedures for constructing confidence intervals, we compute 95 percent confidence intervals for γ, the parameter of the logistic regression model that denotes the effect of sex on mortality of moths exposed to cypermethrin (see Section 2.5.1.3). Recall that the MLE for the logit-scale effect of sex is $\hat{\gamma} = 1.10$, and an estimate of its uncertainty is $\text{SE}(\hat{\gamma}) = 0.356$. Therefore, the 95 percent confidence interval for γ based on inverting the Wald test is $[0.40, 1.80]$ $(=1.10 \pm 1.96 * 0.356)$.

To compute a 95 percent confidence interval for γ based on inverting the likelihood ratio test, we must find the values of γ_0 that satisfy the following inequality

$$-2\{\log L(\hat{\alpha}, \hat{\beta}, \gamma_0 | \boldsymbol{y}) - \log L(\hat{\alpha}, \hat{\beta}, \hat{\gamma} | \boldsymbol{y})\} \leq 3.84. \tag{2.5.6}$$

Using a numerical root-finding procedure, we find that $[0.42, 1.82]$ is the 95 percent confidence interval for γ. In Figure 2.7 we plot the left-hand side of Eq. (2.5.6) against fixed values of γ_0 to show that we have calculated this interval correctly.

2.5.3 A Bayesian Approach to Model Selection

In this section we develop a Bayesian approach to the problem of selecting between two nested models, which was described in Section 2.5.1. Recall that H_0 and H_1 denote two complementary hypotheses that can be specified in terms of a (possibly vector-valued) parameter θ:

$$H_0 : \theta \in \Theta_0$$
$$H_1 : \theta \in \Theta_0^c,$$

where Θ_0 is a subset of the parameter space Θ and Θ_0^c is the complement of Θ_0. The problem is to select either the null model represented by H_0 or the alternative (more complex) model represented by H_1.

A Bayesian analysis of the problem requires a prior for each model's parameters. Let $\pi(\theta|H_0)$ and $\pi(\theta|H_1)$ denote prior density functions for the parameters of null

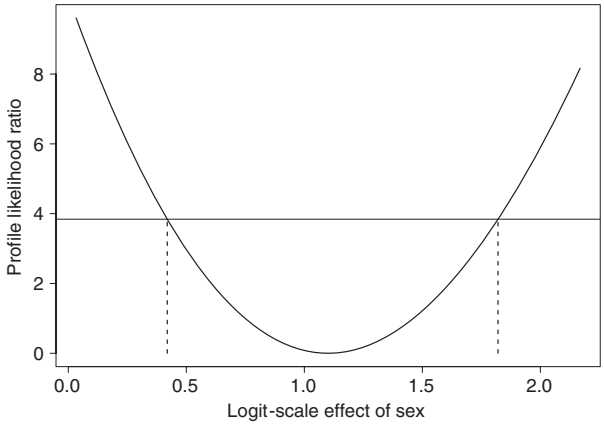

Figure 2.7. Profile likelihood ratio for the logit-scale effect of sex on mortality of moths exposed to cypermethrin. Dashed vertical lines indicate 95% confidence limits. Horizontal line is drawn at $\chi^2_{0.95}(1)$.

and alternative models, respectively. Given these priors, the posterior densities are defined in the usual way (using Bayes' rule) and have normalizing constants

$$m(\boldsymbol{y}|H_0) = \int_{\Theta_0} f(\boldsymbol{y}|\theta)\pi(\theta|H_0)\,\mathrm{d}\theta$$

and

$$m(\boldsymbol{y}|H_1) = \int_{\Theta_0^c} f(\boldsymbol{y}|\theta)\pi(\theta|H_1)\,\mathrm{d}\theta,$$

where $f(\boldsymbol{y}|\theta)$ specifies a model-specific likelihood of the observed data $\boldsymbol{y}$. An additional requirement for Bayesian analysis is that we assign prior probabilities to H_0 and H_1. Let $q_0 = \Pr(H_0) \equiv \Pr(\theta \in \Theta_0)$ quantify our prior opinion of whether the null model should be accepted. Since only two models are involved in the comparison, $\Pr(H_1) \equiv \Pr(\theta \in \Theta_0^c) = 1 - q_0$. As an example, we might use $q_0 = 0.5$ to specify an impartial prior for the two models.

A Bayesian solution of the model selection problem is obtained by computing the *posterior* probability of H_0 given $\boldsymbol{y}$

$$\Pr(H_0|\boldsymbol{y}) = \frac{\Pr(H_0)\, m(\boldsymbol{y}|H_0)}{\Pr(H_0)\, m(\boldsymbol{y}|H_0) + \Pr(H_1)\, m(\boldsymbol{y}|H_1)} \tag{2.5.7}$$

$$= \frac{q_0\, m(\boldsymbol{y}|H_0)}{q_0\, m(\boldsymbol{y}|H_0) + (1-q_0)\, m(\boldsymbol{y}|H_1)} \tag{2.5.8}$$

which follows from a single application of Bayes' rule. Because there are only two models involved, the posterior probability of H_1 must be $\Pr(H_1|\boldsymbol{y}) = 1 - \Pr(H_0|\boldsymbol{y})$. It's worth mentioning that Eq. (2.5.8) could have been derived by specifying the prior density of θ as a finite mixture of model-specific priors as follows: $\pi(\theta) = q_0\pi(\theta|H_0) + (1 - q_0)\pi(\theta|H_1)$.

How should a Bayesian use posterior model probabilities in model selection? One possibility is to reject the null model (and accept the alternative) if $\Pr(H_0|\boldsymbol{y}) < \Pr(H_1|\boldsymbol{y})$ or, equivalently, if $\Pr(H_0|\boldsymbol{y}) < 0.5$. On the other hand, if a Bayesian wants to guard against falsely rejecting the null model, he may wish to reject H_0 under a more conservative (lower) threshold (e.g., $\Pr(H_0|\boldsymbol{y}) < 0.1$).

The Bayesian approach to model selection possesses some appealing features. Unlike the classical approach, the analyst does not have to choose a particular statistic for testing. Furthermore, decisions to reject or accept H_0 do not require samples to be large enough for the application of asymptotic distributions. That said, the Bayesian approach to model selection is not without its problems. For one thing, the posterior model probabilities can be sensitive to the priors assumed for each model's parameters. In addition, the model-specific normalizing constants needed for computing posterior model probabilities can be difficult to calculate when models contain many parameters. Recall that this difficulty impeded routine use of Bayesian statistics for many years and motivated the development of MCMC sampling algorithms.

In Section 2.5.4 we will examine some alternative Bayesian methods of model selection. For now, however, we illustrate a clever approach developed by Kuo and Mallick (1998), which avoids the calculation of normalizing constants but is equivalent conceptually to the approach described earlier in this section. Kuo and Mallick introduce a latent, binary *inclusion parameter*, say w, that is used as a multiplier on those parameters which are present in the alternative model and absent from the null model. By assuming $w \sim \text{Bern}(\psi)$, the prior probability of the alternative model is specified implicitly because

$$\begin{aligned}\Pr(H_1) \equiv \Pr(\theta \in \Theta_0^c) &= \frac{\Pr(w = 1)}{\Pr(w = 0) + \Pr(w = 1)} \\ &= \frac{\psi}{1 - \psi + \psi} \\ &= \psi.\end{aligned}$$

Therefore, we have the identity $q_0 = 1 - \psi$. To specify prior impartiality in the selection of null and alternative models, we simply assume $\psi = 0.5$.

Kuo and Mallick showed that Gibbs sampling may be used to fit this elaboration of the alternative model. The Gibbs output yields a sample from the marginal posterior distribution of $w|\boldsymbol{y}$, which provides a direct solution to the model selection

problem because the posterior probability of the null model $\Pr(H_0|\boldsymbol{y}) = \Pr(w = 0|\boldsymbol{y})$ is easy to estimate from the sample.

The model-selection methodology developed by Kuo and Mallick (1998) applies equally well to much more difficult problems. For example, suppose selection is required among regression models that may include as many as p distinct predictors. In this problem a $p \times 1$ vector $\boldsymbol{w}$ of inclusion parameters may be used to select among the 2^p possible regression models by computing posterior probabilities for the 2^p values of $\boldsymbol{w}$. Alternatively, these posterior probabilities can be used to compute some quantity of scientific interest, by averaging over model uncertainty. Thus, model-averaging can be done at practically no additional computational expense once the posterior is calculated. We conclude this section with an illustration of this approach based on a comparison of two logistic regression models described in a previous example (see Section 2.5.1.3).

2.5.3.1 Example: Mortality of moths exposed to cypermethrin

Recall that the logistic-regression model of moth mortalities contains 3 parameters, an intercept α, the effect of cypermethrin β, and the effect of sex γ. In the test of H_0 we wish to compare a null model for which $\gamma = 0$ against an alternative for which $\gamma \neq 0$. To apply the approach of Kuo and Mallick, we elaborate the alternative model using the binary inclusion parameter w as follows:

$$\begin{aligned} y_i|N, p_i &\sim \text{Bin}(N, p_i) \\ \text{logit}(p_i) &= \alpha + \beta x_i + w\,\gamma z_i \\ w &\sim \text{Bern}(0.5) \\ \alpha &\sim \text{N}(0, \sigma^2) \\ \beta &\sim \text{N}(0, \sigma^2) \\ \gamma &\sim \text{N}(0, \sigma^2), \end{aligned}$$

where σ is chosen to be sufficiently large to specify vague priors for the regression parameters.

We fit this model using **WinBUGS** and estimated the posterior probability of the null model $\Pr(w = 0|\boldsymbol{y})$ to be 0.13. Since $0.13 < 0.5$, we reject H_0 and accept the alternative model. We can estimate the parameters of the alternative model by computing summary statistics from the portion of the Gibbs output for which $w = 1$. This yields the following posterior means and standard errors (in parentheses) of the regression parameters: $\hat{\alpha} = -3.53\,(0.47)$, $\hat{\beta} = 1.08\,(0.13)$, and $\hat{\gamma} = 1.11\,(0.36)$. Given the sample size and our choice of priors for this model's parameters, it is not surprising that the posterior means are similar in magnitude to the MLEs that we reported earlier.

2.5.4 Assessment and Comparison of Models

In this chapter we have described both classical and Bayesian approaches for conducting model-based inference. In applications of either approach, the role of the approximating model of the data is paramount, as noted in Section 2.2. The approximating model provides an unambiguous specification of the data-gathering process and of one or more ecological processes that are the targets of scientific interest.

Statistical theory assures us that inferences computed using either classical or Bayesian approaches are valid, *provided* the model correctly specifies the processes that have generated the data (i.e., nature and sampling). However, since we never know all of the processes that might have influenced the data, we are never really able to determine the accuracy of an approximating model. What we *can* do is define, in precise terms, the operating characteristics of an 'acceptable' model. This definition of acceptability is necessarily subjective, but it provides an honest basis for assessing a candidate model's adequacy. This approach is also useful in the comparison and selection of alternative approximating models. For example, if we are asked questions such as, "Is model A better than model B?" or "Should model A be selected in favor of model B?", a clear definition is needed for what 'better than' means in the context of the scientific problem.

In defining the operating characteristics of an acceptable model, one clearly must consider how the model is to be used. Will the model be used simply to compute inferences for a single data set, or will the model be used to make decisions that depend on the model's predictions? In either case, decision theory may be used to define acceptability formally in terms of a utility function that specifies the benefits and costs of accepting a particular model (Draper, 1996). Similarly, utility functions may be used to specify the benefits and costs of selecting a particular model in favor of others. Such functions provide a basis for deciding whether a model should be simplified, expanded, or left alone (that is, assuming the model is acceptable as is) (Spiegelhalter, 1995; Key et al., 1999). In ecology a popular method of model selection involves the comparison of a scalar-valued criterion, such as Akaike information criteria (AIC) (Burnham and Anderson, 2002) or deviance information criteria (DIC) (Spiegelhalter et al., 2002). This approach is an application of decision theory, though the utility function on which the criterion is based is not always stated explicitly.

The need for a unified theory or set of procedures for assessing and comparing statistical models seems to be more important than ever, given the complexity of models that can be fitted with today's computing algorithms (e.g., MCMC samplers) and software. Unfortunately, no clear solution or consensus seems to have emerged regarding this problem. There does seem to be a growing appreciation among many statisticians, that while inference and prediction are best conducted using Bayes' theorem, the evaluation and comparison of alternative models are best

accomplished from a frequentist perspective (Box, 1980; Rubin, 1984; Draper, 1996; Gelman et al., 1996; Little, 2006). The idea here is that if a model's inferences and predictions are to be well-calibrated, they should have good operating characteristics in a sequence of hypothetical repeated samples.

In the absence of a unified theory for assessing and comparing statistical models, a variety of approaches are currently practiced. We do not attempt to provide an exhaustive catalog of these approaches because recent reviews (Claeskens and Hjort, 2003; Kadane and Lazar, 2004) and books (Zellner et al., 2001; Burnham and Anderson, 2002; Miller, 2002) on the subject are available. Instead we list a few of the more commonly used methods, noting some of their advantages and disadvantages.

In Section 2.5.1 we described the likelihood ratio test for comparing nested models and for assessing the goodness of fit of models of counts. These are frequentist procedures based on the calculation of MLEs and on their asymptotic distributions in repeated samples. We also described a Bayesian procedure for model selection that is based on the calculation of posterior model probabilities. An advantage of the Bayesian procedure is that it can be used to select among *non-nested* models; however, the posterior model probabilities can be *extremely sensitive* to the form of priors assumed for model parameters when such priors are intended to convey little or no information (Kass and Raftery, 1995; Kadane and Lazar, 2004). Several remedies have been proposed for this deficiency (e.g., intrinsic Bayes factors, fractional Bayes factors, etc.), but none is widely used or is without criticism, as noted by Kadane and Lazar (2004).

Some approaches to model selection involve the comparison of an omnibus criterion, which typically values a model's goodness of fit and penalizes a model's complexity in the interest of achieving parsimony. Examples of such criteria include the Akaike, the Bayesian, and the deviance information criteria (i.e., AIC, BIC, and DIC) (Burnham and Anderson, 2002; Spiegelhalter et al., 2002), but there are many others (Claeskens and Hjort, 2003).

Alternative approaches to model selection recognize that some components of a model's specification may not be adequately summarized in a single omnibus criterion. For example, scientific interest may be focused on a particular quantity that requires predictions of the model. In this instance, the operating characteristics of a model's predictions of the scientifically relevant estimand should be used to define a basis for comparing models. Examples of this approach include the use of posterior-predictive checks (Laud and Ibrahim, 1995; Gelman et al., 1996; Gelfand and Ghosh, 1998; Chen et al., 2004) and the focused information criterion (Claeskens and Hjort, 2003).

At this point, the reader may realize that the list of approaches for assessing and comparing statistical models is long. In fact, the published literature on this subject is vast. That the field of Statistics has not produced a unified theory or set of

procedures for assessing and comparing models may seem disconcerting; however, given the wide variety of problems to which models are applied, the absence of unified procedures is perhaps understandable. At present, the construction and evaluation of statistical models necessarily include elements of subjectivity, and perhaps that is as it should be. One cannot automate clear thinking and the subjective inputs required for a principled analysis of data.

2.6 HIERARCHICAL MODELS

In Chapter 1 we argued that hierarchical modeling provides a framework for building models of ecological processes to solve a variety of inference problems. We defined the term *hierarchical model* only conceptually without using mathematics. Here, we define hierarchical models in more concrete terms and provide a number of examples as illustrations. We also illustrate the inferential implications of fitting these models with classical or Bayesian methods.

2.6.1 Modeling Observations and Processes

As with any approximating model of data, hierarchical models must account for the observational (or data-gathering) process and for one or more underlying ecological processes. An interesting and useful feature of hierarchical models is that these processes are specified separately using constituent models of observable and unobservable quantities.

To illustrate, let's consider a typical hierarchical model of 3 components. The first component corresponds to the data y, which are assumed to have been generated by an observation process $f(y|z, \theta_y)$ that depends on some state variable z and parameter(s) θ_y. The state variable z is often the thing we would like to know about (and compute inferences for), especially in situations where z is the outcome of an ecological process. Unfortunately, z is unobserved (or only partially so), and inferences about z can only be achieved by including it as a second component in the hierarchy. This requires a model $g(z|\theta_z)$ whose parameters θ_z are used to specify the variation in z. In many situations g is a model of an underlying ecological process. It's worth noting that if the state variable z had been observed, the model g could have been fitted directly (i.e., ignoring y and the observation process) using statistical methods appropriate for the particular form of g. The third component of the hierarchy corresponds to the parameters θ_z. We typically make a distinction between parameters that appear in the observation model but not the state model – nuisance parameters – and those appearing in the state model that are often of some intrinsic interest.

By providing an explicit representation of the model into its constituents (*observations*, *state variables*, and ecological *processes*), a complex modeling problem often can be rendered into smaller, and simpler, pieces. For example, if we were to fit the model using Bayesian methods, we would need to specify the joint probability density $[y, z, \theta_y, \theta_z]$. The hierarchical formulation of the model provides the following *hierarchical factorization* of the joint density:

$$[y, z, \theta_y, \theta_z] = f(y|z, \theta_y)\, g(z|\theta_z)\, [\theta_y, \theta_z],$$

where $[\theta_y, \theta_z]$ denotes a prior density for the parameters. In practice, it can be very difficult to conceptualize or describe $[y, z, \theta_y, \theta_z]$ directly, but each of the components (mainly $f(y|z, \theta_y)$ and $g(z|\theta_z)$) are often relatively simple to construct. Thus, hierarchical modeling allows an analyst to focus on more manageable components of the problem.

Hierarchical models also provide conceptual clarity by allowing factors that influence observations to be decoupled from those that influence the ecological process. For example, although we cannot observe the state variable z directly, hierarchical models allow us to proceed with model construction as if we *could* have observed it. Thus, by pretending as if we had observed the state variable, the formulation of model components f and g is greatly simplified.

2.6.1.1 Example: estimating survival of larval fishes

We have already encountered an example of a hierarchical model in Section 2.4.4.2, though we did not identify it as such. In this example the number of surviving larval fishes y_i in the ith replicate container is the observation. In the first stage of the model, y_i is assumed to depend on the survival probability ϕ (the parameter of scientific interest) and on the unknown number of fishes N_i initially added to the ith container. In this model N_i corresponds to the state variable, but it is not of any scientific interest (i.e., N_i is just a nuisance parameter in the observation model). The second stage of the hierarchical model expresses variation in N_i among replicate containers in terms of a known parameter λ, the average number of fishes added to each container.

2.6.2 Versatility of Hierarchical Models

The general approach to hierarchical modeling that we described above can be used to solve many inference problems. In fact, hierarchical models appear almost everywhere in contemporary statistics (Hobert, 2000). Such models are commonly used to specify overdispersion and to account for correlated outcomes that arise by

design, as with 'repeated measurements' taken on the same individual or at the same location. In modeling, overdispersion hierarchical models sometimes give rise to standard, 'named' distributions. For example, the beta-binomial and negative-binomial distributions, which are often used to model overdispersed counts, can be generated by marginalizing the binomial-beta and Poisson-gamma mixtures, respectively. Both mixtures are examples of hierarchical models.

Although various synonyms have been used for hierarchical models (e.g., *mixed-effects models* and *multi-level models*), conceptually there is no difference between these models. For historical reasons, statisticians sometimes refer to mid-level parameters as '*random effects*' and to upper-level parameters as '*fixed effects*'. However, in this book we do not focus on the labels that have been attached to different kinds of model parameters or on differences in terminology.

We are more concerned with specifying hierarchical models in terms of quantities that have a recognizable, scientific interpretation — what we have referred to as explicit process models in Chapter 1. In any given problem a scientifically relevant quantity may correspond to an unobserved state variable, a mid-level parameter, an upper-level parameter, a prediction, or perhaps some function of these different model components. That said, we also are interested in selecting an inference method (classical or Bayesian) that allows us to make conclusions (about quantities of interest) that are useful in the context of the problem. The choice of inference method can be especially important in small samples and in problems where predictions are needed. In subsequent chapters of this book we provide several examples where the choice of inference method can be crucial.

For now, we would like to illustrate a few of the differences associated with fitting hierarchical models by different inference methods. In this illustration we use an example from the class of *linear mixed-effects models* (Demidenko, 2004), which are widely applied in biology, agriculture, and other disciplines. We also use this example to introduce some terms that are often used in association with hierarchical modeling.

2.6.2.1 Example: a normal-normal mixture model

Let $\boldsymbol{y}_i = (y_{i1}, y_{i2}, \ldots, y_{im})$ denote a set of replicate measurements made on each of $i = 1, 2, \ldots, n$ subjects. We consider a simple observation model in which y_{ij} is assumed to be normally distributed with a mean that depends on a subject-specific parameter α_i as follows:

$$y_{ij} \sim \mathrm{N}(\alpha_i, \sigma^2). \tag{2.6.1}$$

In practice, m might be relatively small, in which case this model contains many parameters relative to the number of observations in the sample ($=nm$). A common modification is to impose additional structure on α_i, and in many cases this

additional structure makes sense in the context of the problem. For example, suppose the subjects in the sample are selected to be representative of a larger population of individuals. Then we might reasonably assume α_i itself to be a realization of a random variable, say,

$$\alpha_i \sim \mathrm{N}(\mu, \sigma_\alpha^2). \tag{2.6.2}$$

The model implied by combining the assumptions in Eqs. (2.6.1) and (2.6.2) is quite common. In classical statistics the subject-level parameters $\{\alpha_i\}$ are often called *random effects*, whereas the parameters, μ, σ^2, and σ_α^2, that are assumed to be fixed among all subjects in the sample (and population) are called *fixed effects*. Since the model contains both kinds of parameters it qualifies as a *mixed-effects model*.

The issue we examine here is "how does a Bayesian analysis of this model compare to a non-Bayesian analysis?" The answer to that question depends, in part, on the objective of the analysis. Suppose interest is focused primarily on computing inferences for the random effect α_i. The classical (non-Bayesian) solution to this problem requires two steps. In the first step the fixed effects are estimated by removing the random effects from the likelihood; then in the second step the random effects are estimated conditional on the estimated values of the fixed effects. To be more specific, in the first step the MLEs of the fixed effects are computed by maximizing a *marginal* or *integrated likelihood* function that does not involve the random effects,

$$L(\mu, \sigma^2, \sigma_\alpha^2 \mid \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n) = \prod_{i=1}^{n} [\boldsymbol{y}_i \mid \mu, \sigma^2, \sigma_\alpha^2],$$

where

$$[\boldsymbol{y}_i \mid \mu, \sigma^2, \sigma_\alpha^2] = \int_{-\infty}^{\infty} \prod_{j=1}^{m} [y_{ij} \mid \alpha_i, \sigma^2]\, [\alpha_i \mid \mu, \sigma_\alpha^2]\, \mathrm{d}\alpha_i \tag{2.6.3}$$

denotes the marginal probability density of $\boldsymbol{y}_i$ given the fixed effects. This step regards $\{\alpha_i\}$ as a set of nuisance parameters to be eliminated by integration (i.e., marginalization over the joint density of $\boldsymbol{y}_i$ and α_i) (Berger et al., 1999). In the normal-normal mixture the integral in Eq. (2.6.3) can be evaluated in closed form to establish that the vector $\boldsymbol{y}_i$ has a multivariate normal distribution with mean $\mu\mathbf{1}$ and covariance matrix having diagonal elements $\sigma^2 + \sigma_\alpha^2$ and off-diagonal elements σ_α^2; therefore, it is relatively straightforward to compute the MLE of the fixed effects by maximizing the integrated likelihood function.

The second step of the classical solution involves estimating α_i *conditional* on the data and on estimates of the fixed effects. Estimates of α_i are based on the conditional distribution of $\alpha_i|\boldsymbol{y}_i, \mu, \sigma^2, \sigma^2_\alpha$, which is normal with mean

$$\mathrm{E}(\alpha_i|\cdot) = \frac{\sigma^2_\alpha}{\sigma^2_\alpha + \sigma^2/m}\,(\bar{y}_i - \mu) \tag{2.6.4}$$

where $\bar{y}_i = (1/m)\sum_{j=1}^m y_{ij}$ denotes the sample mean of the measurements taken on the ith subject. In classical statistics Eq. (2.6.4) is known as the *best linear unbiased predictor* (BLUP) of α_i (Robinson, 1991; Searle et al., 1992). To compute an estimate of the BLUP (sometimes called an *empirical* BLUP), the MLE computed in the first step is substituted for the fixed effects in Eq. (2.6.4).

The classical solution to the estimation of random effects is also known as *parametric empirical Bayes* (Morris, 1983) because it stops short of a fully Bayesian analysis by not specifying a prior distribution for the fixed effects. However, many Bayesians find this approach to be objectionable, primarily because it uses the data twice (first to compute the MLE as a surrogate for the unspecified prior, and then to compute the posterior of the random effect on which the BLUP is based). Carlin and Louis (2000) provide some colorful historical quotations of some prominent Bayesians (de Finetti, Lindley, and Savage), who were strongly opposed to the empirical Bayes approach.

The main problem with the empirical BLUP (or empirical Bayes estimator) of α_i is that by conditioning on the MLE of the fixed effects, it fails to account for the uncertainty in the MLE. Therefore, estimates of variation in α_i based on $[\alpha_i|\boldsymbol{y}_i, \hat{\mu}, \hat{\sigma}^2, \hat{\sigma}^2_\alpha]$ will generally be negatively biased. Laird and Louis (1987) developed a method for correcting for this bias (at least approximately), but the method requires a considerable amount of additional calculation and the analysis of a large number of parametric bootstrap samples.

So how does a Bayesian compute inferences for the random effect α_i? The answer is simple if one adopts modern methods of Bayesian computation. After specifying a prior distribution for the fixed effects, MCMC methods may be used to compute an arbitrarily large sample from the joint posterior distribution, whose unnormalized density is

$$[\mu, \sigma^2, \sigma^2_\alpha, \alpha_1, \ldots, \alpha_n \mid \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n] \propto [\mu, \sigma^2, \sigma^2_\alpha] \prod_{i=1}^n \left\{ [\alpha_i \mid \mu, \sigma^2_\alpha] \prod_{j=1}^m [y_{ij} \mid \alpha_i, \sigma^2] \right\}.$$

Inferences for the random effect α_i are based on its marginal posterior distribution. A sample from this distribution may be obtained without any additional calculations. We simply ignore the other parameters in the sample of the joint posterior to obtain a sample from $[\alpha_i \mid \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n]$. In doing so, we implicitly

integrate over all parameters (except α_i) of the joint posterior and thereby account for the uncertainty in estimating those parameters. The posterior sample of α_i values may be used to compute an estimate, such as $\mathrm{E}(\alpha_i|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$, a credible interval for α_i, or any other summary deemed to be useful in the context of the problem.

2.6.3 Hierarchical Models in Ecology

Hierarchical models can be used to solve many common inference problems in ecology. The canonical example is probably that of estimating the occurrence or distribution of a species using 'presence/absence' data collected by many different observers in a standardized survey. Variables that describe habitat or landscape typically influence the occurrence of a species, but there may be other variables, such as sampling effort or weather, that primarily influence one's ability to observe species and hence prevent an error-free assessment of the true occurrence state. We address this type of inference problem in detail in Chapter 3. Often, this problem is attacked using logistic regression to build complex models of the mean response, without regard to the distinction between occurrence and its observation. We provide a hierarchical formulation of the problem that provides a good illustration of the sensibility and ease with which hierarchical models may be fashioned from simpler, component models. In this case a hierarchical model for observations contaminated by false-negative errors consists of a compound, logistic regression model – a logistic regression model for the observations conditional on the true occurrence state, and a logistic regression model for the true occurrence state.

Hierarchical models also can be used to estimate abundance from spatially referenced counts of individual animals. For example, in Chapter 1 we described a model wherein the local abundance of harbor seals at each of n spatial sample units was estimated from a set of independent counts at each unit while accounting for the imperfect detectability of the seals. The same modeling strategy can be adopted with other types of counts, which arise when other sampling protocols, such as capture–recapture or double-observer sampling, are used. In this setting hierarchical modeling can be used to produce maps of the spatial distribution of abundance estimates. We describe these types of hierarchical models in Chapter 8.

A final example of hierarchical models that we wish to introduce involves inference about the structure of a biological community. There are competing views as to how to solve this inference problem: the observation-driven view and the process-driven view. Both views are pervasive in ecology. Advocates of the observation-driven view regard the community of species as a classical closed population, wherein species play the role of individuals. Such models may be used to estimate summaries of community structure, such as species richness, but these summaries fail to

preserve species identity and are less likely to meet the needs of conservation and management. Although the observation error is formally accounted for in this view, solving the actual inference problem often requires a second analysis, wherein the estimates of species richness are treated as though they were data. The process-driven view essentially ignores the sampling process and regards the collection of species in the sample as *the* community of interest. Inference about these species proceeds accordingly. Our view is that inference problems associated with community structure are naturally formulated in terms of occurrence of individual species. In a recent series of papers (Dorazio and Royle, 2005a; Dorazio et al., 2006; Kéry and Royle, 2008a,b) we have described a hierarchical, multi-species formulation of occupancy models that accommodates imperfect detection of species and allows species-level models of occurrence to be developed. We demonstrated how such models can be used to estimate many important summaries of community structure. Despite the technical complexity of these models, the multi-species occupancy model stems from a simple hierarchical construction and is relatively easy to implement in **WinBUGS**. We provide a more detailed description of this type of hierarchical model in Chapter 12.