Chapter One

Introduction and background

SUMMARY

This chapter gives a broad overview of the philosophy and techniques of ecological modeling. A small data set on seed removal illustrates the three most common frameworks for statistical modeling in ecology: frequentist, likelihood-based, and Bayesian. The chapter also reviews what you should know to get the most out of the book, discusses the R language, and spells out a step-by-step process for building models of ecological systems.

If you're impatient with philosophical discussion, you can read Section 1.4 and the R supplement at the end of the chapter and move on to Chapter 2.

1.1 INTRODUCTION

This book is about combining models with data to answer ecological questions. Pursuing this worthwhile goal will lead to topics ranging from basic statistics, to the cutting edge of modern statistics, to the nuts and bolts of computer programming, to the philosophy of science. Remember as we go along not to miss the ecological forest for the statistical trees; all of these complexities are in the service of answering ecological questions, and the most important thing is to keep your common sense about you and your focus on the biological questions you set out to answer. "Does this make sense?" and "What does this answer really mean?" are the two questions you should ask constantly. If you cannot answer them, back up to the last point you understood.

If you want to combine models with data, you need to use statistical tools. Ecological statistics has gotten much more complicated in the last few decades. Research papers in ecology now routinely refer to likelihood, Markov chain Monte Carlo, and other arcana. This new complexity arises from the explosion of cheap computing power, which allows us to run complicated tests quickly and easily — or at least more easily than before. But there is still a lot to know about how these tests work, which is what this book is about. The good news is that we can now develop statistical methods that directly answer our ecological questions, adapting statistics to the data rather than *vice versa*. Instead of asking "what is the probability of observing at least this much variability among the arcsine-square-root-transformed counts of seeds in different treatments?", we can ask "is the number of seeds removed consistent with standard foraging theory, and what are the attack rates and handling times of predators? Do the attack rates or handling times increase with mean seed size? With the time that the seeds have been available? Is there evidence for variability among seeds?". By customizing statistical tests we can squeeze more information,

and more relevant information, from expensive data. Building your own statistical tests is not easy, but it is not really harder than using any of the other tools ecologists have picked up in their ongoing effort to extract meaning from the natural world (stable isotope techniques, radiotelemetry, microsatellite population genetics, geographic information systems, otolith analysis, flow cytometry, mist netting . . . you can probably identify several more from your own field). Custom statistical techniques are just another set of tools in the modern ecologist's toolbox; the information this book presents should show you how to use them on your own data, to answer your own questions.

For example, Sandin and Pacala (2005) combined population counts through time with remote underwater video monitoring to analyze how the density of reef fishes in the Caribbean affected their risk of predation. The classic approach to this problem would be to test for a significant correlation between density and mortality rate, or between density and predator activity. A positive correlation between prey population density and the number of observed predator visits or attacks would suggest that prey aggregations attract predators. If predator attacks on the prey population are proportional to population density, then the predation rate per prey *individual* will be independent of density; predator attacks would need to accelerate with increasing population density in order for predators to regulate the prey population. One could test for positive correlations between prey density and *per capita* mortality to see whether this is so.

However, correlation analysis assumes the data are bivariate normally distributed, while linear regression assumes a linear relationship between a predictor variable and a normally distributed response variable. While one can sometimes transform data to satisfy these assumptions, or simply ignore minor violations, Sandin and Pacala took a more powerful approach: they built explicit models to describe how absolute and *per capita* predator visits or mortality depended on prey population density. For example, the absolute mortality probability would be $r_0 + r_1 n$ and the *per capita* mortality probability would be $(r_0 + r_1 n)/n$ if predator visits are proportional to prey density. They also used realistic binomial and Poisson probability distributions to describe the variation in the data, rather than assuming normality (a particularly awkward assumption when there are lots of zeros in the data). By doing so, they were able to choose among a variety of possible models and conclude that predators induce *inverse* density dependence in this system (i.e., that smaller prey populations experience higher *per capita* mortality, because predators are present at relatively constant numbers independent of prey density). Because they fitted models rather than running classical statistical tests on transformed data, they were also able to estimate meaningful parameter values, such as the increase in preda-

tor visits per hour for every additional prey individual present. These values are more useful than p (significance) values, or than regression slopes from transformed data, because they express statistical information in ecological terms.

1.2 WHAT THIS BOOK IS NOT ABOUT

1.2.1 What you should already know

To get the most out of the material presented here you should already have a good grasp of basic statistics, be comfortable with computers (e.g. have used Microsoft Excel to deal with data), and have some rusty calculus. But attitude and aptitude are more important than previous classroom experience. Getting into this material requires some hard work at the outset, but it will become easier as you brush up on basic concepts*.

Statistics

I assume that you've had the equivalent of a one-semester undergraduate statistics course. The phrases *hypothesis test*, *analysis of variance*, *linear regression*, *normal distribution* (maybe even *Central Limit Theorem*) should be familiar to you, even if you don't remember all of the details. The basics of experimental design — the meaning of and need for randomization, control, independence, and replication in setting up experiments, the idea of statistical power, and the concept of pseudoreplication (Hurlbert, 1984; Hargrove and Pickering, 1992; Heffner et al., 1996; Oksanen, 2001) — are essential tools for any working ecologist, but you can learn them from a good introductory statistics class or textbook such as Gotelli and Ellison (2004) or Quinn and Keough (2002)†.

Further reading: If you need to review statistics, try Crawley (2002), Dalgaard (2003), or Gotelli and Ellison (2004). Gonick and Smith's 1993 *Cartoon Guide to Statistics* gives a gentle introduction to some basic concepts, but you will need to go beyond what they cover. Sokal and Rohlf (1995), Zar (1999), and Crawley (2005) cover a broader range of classical statistics. For experimental design, try Underwood (1996), Scheiner and

*After teaching with Hilborn and Mangel's excellent book *The Ecological Detective* I wanted to a write a book that included enough nitty-gritty detail for students to tackle their own problems. If this book feels too hard for you, consider starting with *The Ecological Detective* — but consider reading *ED* in any case.

†Ideally, you would think about how you will analyze your data before you go into the field to collect it. This rarely happens. Fortunately, if your observations are adequately randomized, controlled, independent, and replicated, you will be able to do *something* with your data. If they aren't, no fancy statistical techniques can help you.

Gurevitch (2001), or Quinn and Keough (2002) (the latter two discuss statistical analysis as well).

Computers

This book will teach you how to use computers to understand data. You will be writing a few lines of R code at a time rather than full-blown computer programs, but you will have to go beyond pointing and clicking. You need to be comfortable with computers, and with using spreadsheets like Excel to manipulate data. It will be useful to be familiar with a mainstream statistics package like SPSS or SAS, although you should definitely use R to work through this book instead of falling back on a familiar software package. (If you have used R already you'll have a big head start.) You needn't have done any programming.

Math

Having "rusty" calculus means knowing what a derivative and an integral are. While it would be handy to remember a few of the formulas for derivatives, a feeling for the meanings of logarithms, exponentials, derivatives and integrals is more important than the formulas (you'll find the formulas in the Appendix). In working through this book you will have to *use* algebra, as much as calculus, in a routine way to solve equations and answer questions. Most of the people who have taken my classes were very rusty when they started.

Further reading: Adler (2004) gives a very applied review of basic calculus, differential equations, and probability, while Neuhauser (2003) covers calculus in a more rigorous and traditional way, but still with a biological slant.

Ecology

I have assumed you know some basic ecological concepts, since they are the foundation of ecological data analysis. You should be familiar, for example, with exponential and logistic growth from population ecology; functional responses from predator-prey ecology; and competitive exclusion from community ecology.

Further reading: For a short introduction to ecological theory, try Hastings (1997) or Vandermeer and Goldberg (2004) (the latter is more

general). Gotelli (2001) is more detailed. Begon et al. (1996) gives an extremely thorough introduction to general ecology, including some basic ecological models. Case (1999) provides an illustrated treatment of theory, while Roughgarden (1997) integrates ecological theory with programming examples in MATLAB. Mangel (2006) and Otto and Day (2007), two new books, both give basic introductions to the "theoretical biologist's toolbox".

1.2.2 Other kinds of models

Ecologists sometimes want to "learn how to model" without knowing clearly what questions they hope the models will answer, and without knowing what kind of models might be useful. This is a bit like saying "I want to learn to do experiments", or "I want to learn molecular biology": do you want to analyze microsatellites? Use RNA inactivation to knock out gene function? Sequence genomes? What people usually mean by "I want to learn how to model" is "I have heard that modeling is a powerful tool and I think it could tell me something about my system, but I'm not really sure what it can do".

Ecological modeling has many facets. This book covers only one: statistical modeling, with a bias towards mechanistic descriptions of ecological patterns. The next section briefly reviews a much broader range of modeling frameworks, and gives some starting points in the modeling literature in case you want to learn more about other kinds of ecological models.

1.3 FRAMEWORKS FOR MODELING

This book is primarily about how to combine models with data and how to use them to discover the answers to theoretical or applied questions. To help fit statistical models into the larger picture, Table 1.1 presents a broad range of dichotomies that cover some of the kinds and uses of ecological models. The discussion of these dichotomies starts to draw in some of the statistical, mathematical and ecological concepts I suggested you should know. However, if a few are unfamiliar, don't worry — the next few chapters will review the most important concepts. Part of the challenge of learning the material in this book is a chicken-and-egg problem: in order to know why certain technical details are important, you need to know the big picture, but the big picture itself involves knowing some of those technical details. Iterating, or cycling, is the best way to handle this problem. Most of the material introduced in this chapter will be covered in more detail in later chapters. If you don't completely get it this time around, hang on and see

Scope and approach	
abstract	concrete
strategic	tactical
general	specific
theoretical	applied
qualitative	quantitative
descriptive	predictive
mathematical	statistical
mechanistic	phenomenological
pattern	process
Technical details	
analytical	computational
dynamic	static
continuous	discrete
population-based	individual-based
Eulerian	Lagrangian
deterministic	stochastic
Sophistication	
simple	complex
crude	sophisticated

Table 1.1 Modeling dichotomies. Each column contrasts a different qualitative style of modeling. The loose association of descriptors in each column gets looser as you work downwards.

if it makes more sense the second time.

1.3.1 Scope and approach

The first set of dichotomies in the table subdivides models into two categories, one (theoretical/strategic) that aims for general insight into the workings of ecological processes and one (applied/tactical) that aims to describe and predict how a particular system functions, often with the goal of forecasting or managing its behavior. Theoretical models are often mathematically difficult and ecologically oversimplified, which is the price of generality. Paradoxically, although theoretical models are defined in terms of precise numbers of individuals, because of their simplicity they are usually only used for qualitative predictions. Applied models are often mathematically simpler (although they can require complex computer code), but tend to capture more of the ecological complexity and quirkiness needed to make

detailed predictions about a particular place and time. Because of this complexity their predictions are often less general.

The dichotomy of mathematical *vs.* statistical modeling says more about the culture of modeling and how different disciplines go about thinking about models than about how we should actually model ecological systems. A mathematician is more likely to produce a deterministic, dynamic process model without thinking very much about noise and uncertainty (e.g. the ordinary differential equations that make up the Lotka-Volterra predator prey model). A statistician, on the other hand, is more likely to produce a stochastic but static model, that treats noise and uncertainty carefully but focuses more on static patterns than on the dynamic processes that produce them (e.g. linear regression)*.

The important difference between phenomenological (pattern) and mechanistic (process) models will be with us throughout the book. Phenomenological models concentrate on observed patterns in the data, using functions and distributions that are the right shape and/or sufficiently flexible to match them; mechanistic models are more concerned with the underlying processes, using functions and distributions based on theoretical expectations. As usual, there are shades of gray; the same function could be classified as either phenomenological or mechanistic depending on why it was chosen. For example, you could use the function $f(x) = ax/(b + x)$ (a Holling type II functional response) as a mechanistic model in a predator-prey context because you expected predators to attack prey at a constant rate and be constrained by handling time, or as a phenomenological model of population growth simply because you wanted a function that started at zero, was initially linear, and leveled off as it approached an asymptote (see Chapter 3). All other things being equal, mechanistic models are more powerful since they tell you about the underlying processes driving patterns. They are more likely to work correctly when extrapolating beyond the observed conditions. Finally, by making more assumptions, they allow you to extract more information from your data — with the risk of making the *wrong* assumptions.†

Examples of theoretical models include the Lotka-Volterra or Nicholson-Bailey predator-prey equations (Hastings, 1997); classical metapopulation models for single (Hanski, 1999) and multiple (Levins and Culver, 1971; Tilman, 1994) species; simple food web models (May, 1973; Cohen et al., 1990); and theoretical ecosystem models (Ågren and Bosatta, 1996). Ap-

*Of course, both mathematicians and statisticians are capable of more sophisticated models than the simple examples given here.

†For an alternative, classic approach to the tradeoffs between different kinds of models, see Levins (1966) (criticized by Orzack and Sober (1993); Levins's (1993) defense invokes the fluidity of model-building in ecology).

plied models include forestry and biogeochemical cycling models (Blanco et al., 2005), fisheries stock-recruitment models (Quinn and Deriso, 1999), and population viability analysis (Morris and Doak, 2002; Miller and Lacy, 2005).

Further reading: books on ecological modeling overlap with those on ecological theory listed on p. 6. Other good sources include Nisbet and Gurney (1982) (a well-written but challenging classic) Gurney and Nisbet (1998) (a lighter version) Haefner (1996) (broader, including physiological and ecosystem perspectives) Renshaw (1991) (good coverage of stochastic models), Wilson (2000) (simulation modeling in C), and Ellner and Guckenheimer (2006) (dynamics of biological systems in general).

1.3.2 Technical details

Another set of dichotomies characterizes models according to the methods used to analyze them or according to the decisions they embody about how to represent individuals, time, and space.

An analytical model is made up of equations solved with algebra and calculus. A computational model consists of a computer program which you run for a range of parameter values to see how it behaves.

Most mathematical models and a few statistical models are dynamic; the response variables at a particular time (the state of the system) feed back to affect the response variables in the future. Integrating dynamical and statistical models is challenging (see Chapter 11). Most statistical models are static; the relationship between predictor and response variables is fixed.

One can specify how models represent the passage of time or the structure of space (both can be continuous or discrete); whether they track continuous population densities (or biomass or carbon densities) or discrete individuals; whether they consider individuals within a species to be equivalent or divide them by age, size, genotype, or past experience; and whether they track the properties of individuals (individual-based or Eulerian) or the number of individuals within different categories (population-based or Lagrangian).

Deterministic models represent only the average, expected behavior of a system in the absence of random variation, while stochastic models incorporate noise or randomness in some way. A purely deterministic model allows only for qualitative comparisons with real systems; since the model will never match the data *exactly*, how can you tell if it matches closely

enough? For example, a deterministic food-web model might predict that introducing pike to a lake would cause a trophic cascade, decreasing the density of phytoplankton (because pike prey on sunfish, which eat zooplankton, which in turn consume phytoplankton); it might even predict the expected magnitude of the change. In order to test this prediction with real data, however, you would need some kind of statistical model to estimate the magnitude of the average change in several lakes (and the uncertainty), and to distinguish between observed changes due to pike introduction and those due to other causes (measurement error, seasonal variation, weather, nutrient dynamics, population cycles . . .).

Most ecological models incorporate stochasticity crudely, by simply assuming that there is some kind of (perhaps normally distributed) variation, arising from a combination of unknown factors, and estimating the magnitude of that variation from the variation observed in the field. We will go beyond this approach, specifying different sources of variability and something about their expected distributions. More sophisticated models of variability enjoy some of the advantages of mechanistic models: models that make explicit assumptions about the underlying causes of variability can both provide more information about the ecological processes at work and can get more out of your data.

There are essentially three kinds of random variability:

- *Measurement error* is the variability imposed by our imperfect observation of the world; it is always present, except perhaps when we are counting a small number of easily detected organisms. It is usually modeled by the standard approach of adding normally distributed variability around a mean value.
- *Demographic stochasticity* is the innate variability in outcomes due to random processes even among otherwise identical units. In experimental trials where you flip a coin 20 times you might get 10 heads, or 9, or 11, even though you're flipping the same coin the same way each time. Likewise, the number of tadpoles out of an initial cohort of 20 eaten by predators in a set amount of time will vary between experiments. Even if we controlled everything about the environment and genotype of the predators and prey, we would still see different numbers dying in each run of the experiment.
- *Environmental stochasticity* is variability imposed from "outside" the ecological system, such as climatic, seasonal, or topographic variation. We usually reserve environmental stochasticity for unpredictable variability, as opposed to predictable changes (such as seasonal or latitudinal changes in temperature) which we can incorporate into our

models in a deterministic way.

The latter two categories, demographic and environmental stochasticity, make up *process variability** which unlike measurement error affects the future dynamics of the ecological system. Suppose we expect to find three individuals on an isolated island. If we make a measurement error and measure zero instead of three, we may go back at some time in the future and still find them. If an unexpected predator eats all three individuals (process variability), and no immigrants arrive, any future observations will find no individuals. The conceptual distinction between process and measurement error is most important in dynamic models, where the process error has a chance to feed back on the dynamics.

The distinctions between stochastic and deterministic effects, and between demographic and environmental variability, are really a matter of definition. Until you get down to the quantum level, any "random" variability can in principle be explained and predicted. What determines whether a tossed coin will land heads-up? Its starting orientation and the number of times it turns in the air, which depends on how hard you toss it (Keller, 1986). What determines exactly which and how many seedlings of a cohort die? The amount of energy with which their mother provisions the seeds, their individual light and nutrient environments, and encounters with pathogens and herbivores. Variation that drives mortality in seedlings — e.g. variation in available carbohydrates among individuals because of small-scale variation in light availability — might be treated as a random variable by a forester at the same time that it is treated as a deterministic function of light availability by a physiological ecologist measuring the same plants. Climatic variation is random to an ecologist (at least on short time scales) but might be deterministic, although chaotically unpredictable, to a meteorologist. Similarly, the distinction between demographic variation, internal to the system, and environmental variation, external to the system, varies according to the focus of a study. Is the variation in the number of trees that die every year an internal property of the variability in the population or does it depend on an external climatic variable that is modeled as random noise?

1.3.3 Sophistication

I want to make one final distinction, between simple and complex models and between crude and sophisticated ones. One could quantify simplicity vs. complexity by the length of the description of the analysis, or the number

*Process variability is also called *process noise* or *process error* (Chapter 10).

of lines of computer script or code required to implement a model. Crudity and sophistication are harder to recognize; they represent the conceptual depth, or the amount of *hidden* complexity, involved in a model or statistical approach. For example, a computer model that picks random numbers to determine when individuals give birth and die and keeps track of the total population size, for particular values of the birth and death rates and starting population size, is simple and crude. Even simpler, but far more sophisticated, is the mathematical theory of random walks (Okubo, 1980) which describes the same system but — at the cost of challenging mathematics — predicts its behavior for *any* birth and death rates and any starting population sizes. A statistical model that searches at random for the line that minimizes the sum of squared deviations of the data is crude and simple; the theory of linear models, which involves more mathematics, does the same thing in a more powerful and general way. Computer programs, too, can be either crude or sophisticated. One can pick numbers from a binomial distribution by virtually flipping the right number of coins and seeing how many come up heads, or by using numerical methods that arrive at the same result far more efficiently. A simple R command like `rbinom`, which picks random binomial deviates, hides a lot of complexity.

The value of sophistication is generality, simplicity, and power; its costs are opacity and conceptual and mathematical difficulty. In this book, I will take advantage of many of R's sophisticated tools for optimization and random number generation (since in this context it's more important to have these tools available than to learn the details of how they work), but I will avoid many of its sophisticated statistical tools, so that you can learn from the ground up how statistical models really work and make your models work the way you want them to rather than being constrained by existing frameworks. Having reinvented the wheel, however, we'll briefly revisit some standard statistical frameworks like generalized linear models and see how they can solve some problems more efficiently.

1.4 FRAMEWORKS FOR STATISTICAL INFERENCE

This section will explore three different ways of drawing statistical conclusions from data — frequentist, Bayesian, and likelihood-based. While the differences among these frameworks are sometimes controversial, most modern statisticians know them all and use whatever tools they need to get the job done; this book will teach you the details of those tools, and the distinctions among them.

To illustrate the ideas I'll draw on a seed predation data set from Duncan and Duncan (2000) that quantifies how many times seeds of two different species disappeared (presumably taken by seed predators, although we can't