Statistical Modeling for Naturalists

Statistical Modeling for Naturalists

By

Pedro F. Quintana Ascencio,
Federico López Borghesi
and Eric S. Menges

Cambridge
Scholars
Publishing

Statistical Modeling for Naturalists

By Pedro F. Quintana Ascencio, Federico López Borghesi
and Eric S. Menges

This book first published 2022

Cambridge Scholars Publishing

Lady Stephenson Library, Newcastle upon Tyne, NE6 2PA, UK

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN (10): 1-5275-7587-X
ISBN (13): 978-1-5275-7587-5

CONTENTS

ACKNOWLEDGEMENTS

The main structure of this book took shape while teaching multiple iterations of statistical modeling classes to graduate students in the University of Central Florida, Orlando, Florida, USA, and short courses at El Colegio de La Frontera Sur, Chiapas, México, Universidad Técnica Particular de Loja, Loja, Ecuador, Universidad Regional Amazónica – Ikiam, Tena, Ecuador, and Universidad Rey Juan Carlos, Móstoles, Madrid, Spain. We are profoundly thankful to all participants for challenging our views. We thank James Angelo and Lina María Sánchez-Clavijo for their invaluable contributions. Reed Bowman kindly provided the front cover photo and Haoyu Li most of the black and white pictures. Conversations with Dave Jenkins, Carlos Iván Espinosa, Diego Vélez Mora, David Duncan, Betsey Boughton, Mario González Espinosa, Neptalí Ramírez Marcial, Brad Ochocki, Xavier Picó, Maria Paniw, Jesse Anaya, Aaron David, Will Crampton, Carl Weekley, Jennifer Navarra, Mark Burgman, Margaret Evans, John Fauth, Stephanie Koontz, David Nickerson, Fernando Quintana-Ascencio, José María Iriondo Alegría, Adrián Escudero, José Miguel Olano and Roberto Salguero Gómez were edifying. The support and advice of Eréndira Malinali Quintana Morales, Amarantha Zyanya Quintana Morales, Adrian Edward De Angelis, and Laura Cristina De Angelis have enriched this and many other ventures. Most especially, we are grateful for the unwavering patience and companionship of María Cristina Morales Hernández and Ericka Vanessa Correa Roldán without whom this book would not exist. We thank research assistants and interns at Archbold Biological Station and students at the University of Central Florida for their participation in collecting and organizing much of the data. The staff at Archbold Expeditions offered extensive aid and guidance. They are instrumental for the preservation of the ecosystems described in this book. We are particularly indebted to Luis Cayuela Delgado and Ian Biazzo for their extensive reviews and invaluable suggestions. They caught many errors and considerably improved the content of every chapter. Helen Edwards, our commissioning editor, provided continuous support during the final preparation of this document. We greatly celebrate the contributions of people that historically provided the conceptual basis allowing us to organize information, examine questions and guide our actions in efficient and creative ways. Any imprecisions and mistakes on

the interpretation and implementation of the methods in this book are the responsibility and privilege of the authors.

FOREWORD

Statistical Modeling for Naturalists presents an outstanding introduction to different statistical methods focusing on the application of models for ecological problem-solving. It covers many statistical methods typically used in the field of ecology, such as linear models, generalized linear models and mixed modeling. The book provides access to datasets and code used for the different case studies. Two freely available, open-source software platforms are used in this book, R and Stan. R is one of the most popular programming languages for data analysis, statistical modeling, and visualization. Stan is a platform for statistical modeling and high-performance statistical computation. Both R and Stan run on all major platforms (Linux, Mac, Windows) and can interface with other data analysis languages.

But what makes this book unique compared with other published books featuring words such as “statistical” and “ecology”?

First, it is written by ecologists, and therefore the emphasis is on natural history. This is pretty well illustrated throughout the book, where each chapter develops a case study in population ecology based on decades of research conducted by the authors at Archbold Biological Station in Central Florida, USA. Each case study represents a research question, which is carefully framed by ecological theory about the problem at hand. This puts statistics at the service of ecological problem-solving and makes the book particularly appealing for biology and environmental science students and practitioners. By the end of the book, you will have acquired a pretty good knowledge of the ecological functioning of plant populations in the Florida scrub ecosystem even if, as in my case, you have never set foot in it.

Second, the book is based on a Bayesian approach to statistics; that is, statistical inference is based on the posterior distribution, which expresses all that is known about the parameters of a statistical model, given the data and existing knowledge. There are many advantages of the Bayesian approach, such as the possibility to combine previous with existing information to address the research question or the ease of error propagation, which is making the Bayesian paradigm gain tremendous momentum in statistics and its applications, including ecology. Yet, the

Bayesian approach might turn back many students who have been trained in the use of frequentist methods. But do not panic! The book provides very clear explanations about the way models are constructed and how to interpret the results. So regardless of you having received previous training either as a Bayesian or a frequentist, you will find the book very easy to follow. I must admit I am considering turning Bayesian myself after reading the book!

The way the book is structured makes it easy for readers to find good recipes for framing and analyzing their own research questions based on the presented case studies. But reading the book from beginning to end also gives an added value to the reader, as problems are presented in increasing complexity and quite often based on the results found in previous chapters. Statistical Modeling for Naturalists is worth reading. It offers insightful clues for conducting sound ecological research to students, practitioners, and researchers in different stages of their academic careers.

Enjoy the reading!

Luis Cayuela
Department of Biology and Geology, Physics, and Inorganic Chemistry
Universidad Rey Juan Carlos, Spain

PREFACE

This book aims to introduce the reader to captivating and useful methods for statistical data analysis and basic mathematical modeling. It is directed at people with rudimentary/intermediate statistical and mathematical instruction. It is neither a book for the beginner nor the specialist. For the former, we make our best efforts to explain every topic in the book in terms that are approachable to a novice. For those interested in mathematical demonstrations and programming specifics, we do our best to redirect them to more sophisticated sources. The book is written to appeal to naturalists like us. To this end, we focus on the questions and considerations that guide our modeling choices. We recapitulate three decades of ecological work on the Florida Scrub at Archbold Biological Station in Central Florida, USA. Each chapter follows two threads: one mathematical/statistical and the other ecological. This approach has been successful in retaining the attention of graduate students who are more interested in ecology and evolution than in data analysis and modeling. We are convinced about the prominence of mathematical and statistical thinking in producing robust, reproducible, and relevant science. But we are also persuaded of the need to constantly relate our models to the concepts being evaluated. The approaches in this book require code writing, so we use freely available software – R and Stan. We find these platforms to be among the best communitarian projects in recent decades, transforming the use and progress of mathematical and statistical applications. We maintain an associated website (https://github.com/StatsForNats/Book) where the reader can access all of the data and code used in the book. We must remark, however, that this book is not a guideline on either language. We recommend the reader to use the suggested literature to complement the material presented here. Most important, we employ Bayesian inference. Bayesian models are flexible and powerful but have been overlooked, mostly due to computational limitations and lack of training. Bayesian analyses formalize the process of evaluating information involved in the generation of knowledge, and they are particularly suited for the complex data used in ecological studies. With this book we intend to help make these tools accessible to naturalists and other people interested in understanding ecological mechanisms.

CHAPTER I

MODELS IN ECOLOGY

"All models are wrong, but some are useful"
– George Box

Relevant notions

Models allow us to explore complex phenomena and develop our understanding. They do this by offering a tradeoff – we must focus on a few aspects while excluding many others. Which factors we choose to include, and which to exclude, depends on what it is we are trying to understand. A model airplane might be useful to evaluate aerodynamic properties but it ignores the propulsion system.

The notion of models as simplified representations of complex reality is very much present across science. Such a vague definition of scientific

models is purposeful. First, it allows us to recognize that models are ubiquitous to our way of thinking and can, therefore, take many forms. A scientific model can be anything from a visual model (such as a flowchart) to a computer program combining complex mathematical functions. In this book we deal mostly with statistical models, in which we propose a data-generating process and then estimate its parameters empirically. Regardless of its form, the purpose of any model is to make the phenomena in question easier to understand. The corollary of this statement is that we should not make models more complicated than they need to be. Second, this definition helps us to acknowledge the limitations of any given model. A frequent aphorism in statistics states that "all models are wrong, but some are useful." We prefer to see the glass half full and say that no model is fully correct. Models can provide different insights and help us to identify important information, but we must not feel guilty of letting go one model in favor of another. Finally, recognizing models as simplified descriptions of reality forces us to acknowledge that many assumptions are necessary in their construction. Model assumptions are the topic of Chapter V, but we want to emphasize here the importance of keeping them in mind and validating them whenever possible.

Within ecology, we make inaccurate observations on phenomena which arise from a myriad of interacting variables that change continuously. The size of a plant, for instance, is determined by multiple environmental factors (such as sunlight, water, and nutrients), biotic interactions (including diseases and predation), physical, chemical, and mechanical constraints, and the individual genetic background and ontogeny. The timing and intensity of each of these variables are affected by an increasing number of other factors.

Realistically, our studies can only deal with a small number of those factors, and models are the tools we use to help us focus on those relations. This is not to say that we randomly pick variables from a hat and use models to see whether they are helpful in explaining our observations. Our choices are biased and informed by our motivations, education, prior experiences, and ideas (Larson 2011). A butterfly flapping its wings might affect the initial conditions of a system sufficiently to help bring about a hurricane, but it would be silly for meteorologists to go around recording flying butterflies to forecast hurricanes. Their focus on data collected from weather tracking technology is well justified.

In this book we explore statistical approaches to evaluate ideas using data as arbitrator. When we propose models to formalize hypotheses (suggesting

possible explanations of our observations), we require a way to substantiate them. One possible approach is through classical frequentist statistics. However, we agree with Howson and Urbach (1989) that the Newman-Pearson/Fisher approach fails to reach this goal in many realms. Their aspiration of objectivity, supposedly intended to avoid using degrees of belief, culminates in several paradoxes. These start with the problematic definition of "accepting" and "rejecting" a hypothesis since accepting it does not mean it is true and rejecting it does not imply it is false. The decision is arbitrary on the choice for the null hypothesis and its frequent lack of relevance for the question of interest. In any case, attempting to validate a hypothesis with the falsification of a null hypothesis does not inform how much the single alternative hypothesis compares to other logical explanations (Aho et al. 2014). Since they rely exclusively on the data at hand, ignoring any prior information, they are poorly suited for evaluating complex hypotheses and studies with limited data. There is an unwarranted belief in the certainty of a long-run frequency associated to the model parameters. Importantly, we believe the confidence placed in an arbitrary threshold as a single criterion when evaluating evidence is unjustified, and leads to poor inference (see Nuzzo 2014 and Wasserstein and Lazar 2016).

An alternative approach (the one embraced in this book) relies on the use of Bayesian inference. This approach ranks hypotheses within a spectrum of degrees of certainty in the light of evidence. It uses information available independently of a study alongside the likelihood of the data at hand to generate posterior probability distributions (see Chapter III). The posterior distributions from one study become available as prior information for future studies. Another advantage of Bayesian inference is that it considers model parameters as random variables instead of assuming they are true, fixed quantities (Ellison 2004). We find this long-standing philosophical tradition to be a satisfactory and robust way to evaluate ideas. It is a clear method for estimating model parameters and evaluating the degree of their uncertainty. It offers a probabilistic measure of the relative confidence we can have in competing models.

Touchon and McCoy (2016) documented that, while there has been a clear increase in the use of Bayesian statistics in the mainstream ecological literature, most doctoral programs in the United States still do not cover these approaches. In this book, we introduce Bayesian probabilistic frameworks as tools to create and evaluate models with the hope of reducing this disparity.

In his brilliant contribution, “Remedies for a scientific disease”, Mark Burgman warns against a propensity among scientists to feel overly confident in our predictions. Since most ecological studies do not even attempt to determine if the intensity of monitoring is sufficient, there is a great risk of overlooking effects when the data are too limited. Often, many scientists wrongly interpret a lack of statistical effects as a lack of “real” effect. This is particularly damaging when assessing environmental impacts since the costs of false negatives can lead to real problems being disregarded. Burgman advocates for the use of the precautionary principle, weighting the costs of false positives and false negatives when evaluating the results of our analysis. We concord with him that one of the best safeguards against irrational interpretation is the use of images of the data and models’ predictions in the form of scatterplots and histograms, and illustrations of estimates of associated uncertainty. With these ideas in mind, we try to present extensive images depicting the level of support for our models and encourage the reader to do the same.

Our model system is the Florida scrub and, in particular, an endangered and rare plant species inhabiting this ecosystem, both described in Chapter II. We use these model systems to introduce what we believe are powerful concepts that can help in developing descriptive statistics, inference, and modeling. We build these models as examples of tools to gain an understanding of ecological phenomena in this ecosystem. Using case studies dealing with contemporary ecological theories, we present relevant statistical concepts and procedures used to build models aimed to improve their understanding.

CHAPTER II

FLORIDA SCRUB

"The job of an ecologist is like putting together a giant puzzle, only all the pieces are hidden and the puzzle keeps changing"
– Warren Abrahamson

As natural ecosystems go, Florida scrub does not have a high profile. It is not majestic like a redwood forest, iconic as the Everglades, or as romantic as a midwestern grassland. But still, it is a beautiful, interesting, and at times confounding landscape. It is also the source of data used in this book. Here is a short description of Florida scrub.

Florida scrub is a shrubland, dominated by woody plants with multiple stems. It is not a forest (despite some prior categorizations of it as "sand pine scrub" or "sand pine forest"). The dominants are mainly shrub oaks, dwarf palmettos, and shrubs in the blueberry family. Elsewhere in the landscape, there are natural forests in Florida, which have the abundant

precipitation necessary to support tree growth. But under most circumstances, Florida scrub has only scattered (or no) trees and is not routinely invaded by trees.

Florida scrub has sometimes been characterized as a desert, in part because there are many xeromorphic characteristics in its plants. These include small evergreen leaves that are thick and waxy, and stems that are often spiny. These traits are often assumed to be a function of low water availability but can also evolve in response to nutrient deficiency or herbivory (see Chapter III). Florida has no shortage of rain (although it is strongly seasonal) but the sandy soils supporting Florida scrub are extremely low in nitrogen, phosphorus, and other nutrients. With added nutrients, both planted citrus and introduced grasses can thrive on these soils. But, left undisturbed by humans, Florida scrub is often the result.

Most of Florida is affected by periodic, natural fire. The high productivity during the rainy season, the pronounced dry season, frequent lightning, and flammable vegetation all contribute to fires that can be frequent or intense. Florida scrub, occurring on less productive sites than more mesic vegetation, burns relatively infrequently (say every 10-30 years) and intensely. On the other hand, grass-dominated habitats like wiregrass-dominated flatwoods (see Chapter VI) burn more frequently and less intensely. The interplay between fire and vegetation is complex and occurs at multiple spatial scales (see Chapters XI and XII).

Most plants in Florida scrub survive fires, resprouting from below-ground storage organs such as rhizomes. The resprouting is rapid but varies with the species, prior plant size, and season. Types of Florida scrub (e.g., scrubby flatwoods) dominated by resprouting oaks and palmettos rapidly regain their height and density in a few years or decades. A few plants take advantage of the short-term openings (gaps) in the resprouting vegetation to complete their life cycles.

Other plants use dormant seeds to deal with periodic fire. Post-fire germination and growth can be rapid and allow species of lesser competitive abilities to survive in the landscape. The seeds that provide the new start are dormant in the soil and cued to germinate by increased light or fire cues such as smoke or heat.

One of these obligately seeding plants is Highlands scrub hypericum, *Hypericum cumulicola*, the source of much of the data in this book (see Chapters IV-V and VII-XIII). This small herbaceous plant can produce

seeds as early as its first year of life (see Chapter VIII). It populates the soil with dormant seeds that can germinate in high numbers after fire, even after many years of quiescence below ground. Plants do particularly well in the years after fire, and populations grow rapidly in size during the first decade after burns. However, plants and populations begin an inexorable decline after this first decade, often disappearing above ground until the next fire (see Chapters XI and XII). Increasing competition with shrubs partially explains this decline.

Hypericum cumulicola occurs mainly in a type of scrub termed Rosemary scrub, named after Florida rosemary (*Ceratiola ericoides)*. This shrub is different from oaks and palmettos in that it is killed by fire. Florida rosemary accumulates dormant seeds in the soil, much like *H. cumulicola.* After a fire, its seeds also germinate, but on a delayed schedule relative to herbaceous plants like *H. cumulicola.* This delay gives the smaller plants a head start. But these diminutive individuals are likely to do well only where there are large gaps between the growing Florida rosemary plants. As time marches on, these gaps shrink as Florida rosemary and other shrubs expand their canopies and root systems. With enough time between fires, the gaps become too small to support the less competitive species.

Florida scrub is full of endemic plants and animals, species that are found only in a small area. Many of the rare plants are specialists for rosemary scrub and for gaps (*H. cumulicola* is an example). The biodiversity of this landscape is dependent upon fire and the creation and maintenance of gaps among the dominant shrubs. This relationship is also true for some iconic animals of Florida scrub and other upland ecosystems. Florida scrub jays need bare sand to cache acorns and few pines that harbor predators. Gopher tortoises need open areas to forage on herbaceous plants that are most abundant following fires.

Florida Rosemary scrub occurs as relatively well-defined patches of different sizes within a matrix of wetlands, scrubby flatwoods and flatwoods which have higher moisture availability and are more heavily covered by vegetation (Chapter XI). To colonize other Florida scrub suitable patches, endemic plants and animals in Florida rosemary scrub, need to disperse through large tracts of unsuitable habitat. The spatial arrangement of the scrub patches determines the probability of occurrence of many of these endemics, with large and aggregated patches having the highest probabilities of occupancy (Chapter XII). These endemic species rely on varying compromises between dormancy and dispersal to take advantage of the temporal occurrence of optimal habitats a few years after fire, while also

minimizing the consequences of suboptimal habitat development associated with post-fire recovery. Their chances of persistence in the landscape are influenced by spatial heterogeneity and frequency of disturbance.

Conservationists and land managers are well aware of the key role of fire in promoting biodiversity in Florida scrub (and other Florida ecosystems) and apply prescribed fire routinely. Robust debate ensues among land managers about the frequency, intensity, seasonality, and patchiness of burns. For example, patchy burns are thought to be important in allowing the co-existence of species with different optimal requirements for different aspects of the fire regime. Unburned patches are key refugia for species that are tolerant of dense vegetation but lacking the ability to resprout and do not form seed banks to recover from fire.

Unfortunately for the biota, the expanding human population in Florida has eliminated a great deal of natural habitat, fragmented the remaining pieces, and successfully reduced the occurrence of natural fire across the landscape. Remaining habitat fragments, if unmanaged, become overgrown with few suitable areas for many plant and animal species. Diversity decreases and the narrowly specialized organisms are particularly likely to disappear. Some species, like *H. cumulicola*, can persist along edges such as sand roads and fire lanes. However, these anthropogenic habitats are not the same as less disturbed scrub, with effects on the demography of many species (Chapter XIII).

Although only about 15% of Florida scrub on the Lake Wales Ridge has been spared from development, the remaining landscapes, when managed with fire, are supporting a vast array of bacteria, fungi, algae, plants and animals, and a fascinating mixture of adaptations and interactions that has received a great deal of study and will receive continued attention from scientists, conservationists, and nature-lovers.

CHAPTER III

CONNECTING MODELS WITH DATA: A QUICK GLANCE AT PROBABILITY

"Extraordinary claims require extraordinary evidence."
– Carl Sagan

Probability: A framework for model arbitration

If models are simplified descriptions of reality, multiple models might be proposed to describe the same phenomena. So, a valid and pertinent question arises: how do we decide which models to favor? In the book "The Ecological Detective", the authors Ray Hilborn and Marc Mangel (1997) make a strong case for letting data arbitrate between competing models. Fitting models to data allows us to judge their relative merit. This seemingly straightforward task can be more challenging than anticipated, as we will see throughout this book. It requires us to be mindful – ask pertinent

questions, perform diagnostic tests, keep track of our assumptions, and, most difficult, acknowledge our ignorance.

The idea of allowing data to judge our models is a powerful and appealing one, but every arbitrator needs some guideline or framework to settle disputes. In this book, the underlying framework used to connect data to models is provided by probabilities.

Although there is still disagreement regarding the most useful definition of probability, it is widely accepted that it relates to how likely an event is to occur. It is defined as the expectation of a particular outcome considering all possible outcomes. A major disagreement among competing frameworks is the basis of this expectation: whether it is obtained from a sample representing outcomes of a long-term (essentially infinite) frequency distribution or if it constitutes a degree of belief that must be updated after new data. The probability of an event is represented by a number between 0 and 1 – where 0 signifies the event is impossible and 1 that it always takes place. When taken as a whole, probabilities provide a way to describe uncertainty. A very instructive account on probabilities can be found in Hilborn and Mangel (1997). Below we discuss some relevant concepts.

Probability of finite events

In order to reach a definition of probability that is applicable to the purposes of this book, let's start by looking at a series of simple Venn logical diagrams representing the probability of finite events.

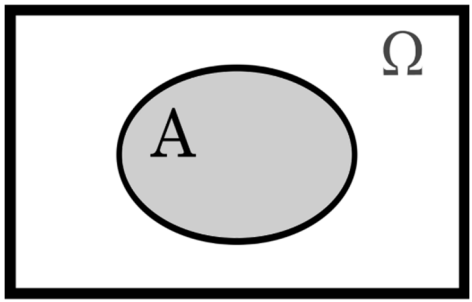

Figure 3.1: Space of event A

If the area of the oval *A* in Figure 3.1 is thought of as representing all instances of event *A*, and the area of the rectangle Ω as representing all

instances of any possible event (including event *A*), the probability of *A* can be described as:

$$P(A) = \frac{area\ of\ A}{area\ of\ \Omega} \qquad \text{(Equation 3.1)}$$

We could think, for example, of Ω as all instances of *Hypericum cumulicola* plants and *A* as all instances of adult *H. cumulicola* defined as a plant older than a year. The remaining blank space in rectangle Ω represents all instances of finding recently recruited plants (< one year old). These two events are mutually exclusive. Now that we have a simple definition of probability, we need to turn our attention to another important concept: conditional probability. To do so, we need to look at two sets of events that might occur simultaneously (not mutually exclusive).

In Figure 3.2, there are two overlapping ovals (*A* and *B*), which represent different events that can potentially occur at the same time. At Archbold Biological Station chances of damage by mammals (mostly rabbits and deer) in *H. cumulicola* are very variable (10-90%), decreasing with time-since-fire and increasing with the number of co-occurring conspecifics in the immediate area (Brudvig and Quintana-Ascencio 2003). Imagine that Ω represents all *H. cumulicola* plants in a Florida Rosemary scrub patch. Oval *A* represents all instances of plants with evidence of consumption by mammals in this patch, and oval B all instances of dead plants. In this case, there are 4 possible outcomes: instances of damaged *H. cumulicola* live plants (the non-overlapping section of *A*); instances of intact *H. cumulicola* plants that died (the non-overlapping section of *B*); instances of damaged dead plants (the overlap between both ovals); and instances of intact live plants (the remaining blank space).

Before turning our attention to conditional probabilities, we need to define joint probabilities, which pertain to the interaction between variables. This probability can be visualized as the area of the overlapping section divided by the total area of the rectangle Ω (Figure 3.2).

$P(A \cap B)$ = the probability of *A* and *B* happening simultaneously

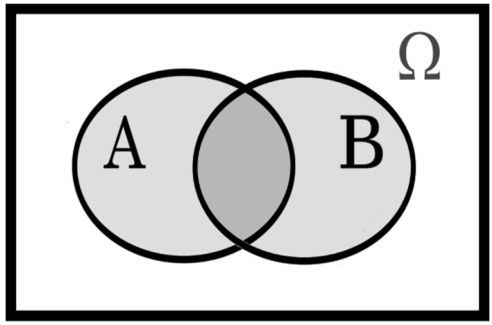

Figure 3.2: Overlapping events

If the probability of a plant dying was independent of damage (i.e., consumed plants have the same mortality as intact plants) we would call these two sets independent events, and $P(A \cap B)$ would simply be the multiplication of *P(A)* and *P(B)*.

You might suspect, however, that the probability of a plant dying is higher for damaged plants. If that were the case, we would not be able to simply multiply *P(A)* and *P(B),* because *P(B)* – the probability of any plant dying – combines information from both damaged and intact plants. So, how do we know if two events are independent? For this we need to look at conditional probabilities.

Let's say event *A* has already taken place (we found a damaged *H. cumulicola*), we want to know the probability that this plant is also dead. The notation for this is *P(B|A),* or the probability of *B* given *A*. The first thing we need to do is recognize that our event space has changed. Any area representing intact plants (both dead and alive) is no longer considered. In other words, we would crop Figure 3.2 and consider only oval *A* and the overlapping section between ovals *B* and *A* (Figure 3.3).

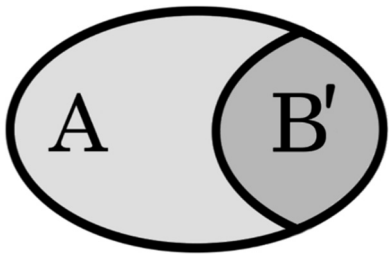

Figure 3.3: Space of *B* given *A*

The probability of *B* given *A*, then, can be visualized as the area of the segment of *B* called *B*' divided by the area of *A*.

$$P(B|A) = \frac{area\ of\ B\prime}{area\ of\ \mathrm{A}} \quad \text{(Equation 3.2)}$$

From looking at the sequence of logic diagrams, we can see that, the area of *A* is equivalent to *P(A)* in the original space **Ω**, and the area of *B*' is equivalent to *P(A ∩ B)*. Therefore, we can rewrite the probability of *B* given *A* as:

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \quad \text{(Equation 3.3)}$$

Following the same logic, we can define the probability of *A* given *B* (having found it dead, what is the probability that the plant was damaged?). The equation is simply:

$$P(A|B) = \frac{P(B \cap A)}{P(B)} \quad \text{(Equation 3.4)}$$

Bayes' theorem

As we saw earlier, calculating joint probabilities for non-independent variables is not trivial. So, calculating a conditional probability would still be difficult. We could instead define one conditional probability in terms of the other one. From equation 3.4, then, we can solve the joint probability $P(B \cap A)$

$$P(B \cap A) = P(A|B) \times P(B) \quad \text{(Equation 3.5)}$$

Since the overlap between *A* and *B* is the same as that between *B* and *A*, *P(B ∩ A)* is the same as *P(A ∩ B)*. So, replacing *P(A ∩ B)* in equation 3.3, we obtain:

$$P(B|A) = \frac{P(A|B) \times P(B)}{P(A)} \quad \text{(Equation 3.6)}$$

This last equation is called Bayes' theorem. As we will see in a subsequent section, this is a powerful tool that provides a way to revise our beliefs (usually called updating probabilities) given new evidence. Returning to our previous example, calculating the mortality of damaged plants might be

difficult because it is hard to estimate how many damaged plants beat the consequences of the herbivory and survived. We could readily get the overall mortality – *P(B)* – from the literature, and conduct a survey to estimate the probability of a plant being damaged – *P(A)* – as well as the probability of a dead plant being damaged *P(A|B)*. Using Bayes' theorem, we could now calculate *P(B|A),* or the mortality of damaged plants.

Expanding the scope

So far, we have dealt with binary events as realizations of variables that could take one of two values (present or absent, infected or not infected). Random variables, however, may also take more complex forms in which events can have multiple realizations, or even infinite, each with its own probability. In these cases, it may not be practical or even possible to specify each probability individually. To describe these probabilities, then, we can make use of probability distribution functions. The last section of this chapter will present several useful probability distributions and examples of how they can be used to model real data.

Probability distributions are often divided as either discrete or continuous depending on how the variable we want to model is measured. Anything that is counted as integers (such as the number of plants in a population) is considered as discrete, while anything that is measured using units that can be divided into infinitely smaller parts (such as weight) is considered as continuous. It is important to remark that whether a variable is discrete or continuous is contingent on how we perform the measurement rather than the nature of the trait. For instance, we could represent the foliage of a tree as the number of leaves (discrete) or as the total area of the leaves (continuous). The key point to bear in mind is that probability distribution functions are themselves models used to describe a more complex reality and, therefore, are not set in stone.

In describing discrete random variables, we can use *probability mass functions* (PMF) to compute the probability that the variable takes a particular value (x_i), so that:

$$f(x_i) = P(X = x_i) \qquad \text{(Equation 3.7)}$$

The actual form of the function $f(x_i)$ will change depending on the discrete distribution being described, but all PMF share a few characteristics. First, the values of the PMF for each potential x_i must be between 0 and 1, as they

represent probabilities. Second, the sum of the probabilities for all realizations of x_i must be equal to 1.

Another way to describe the distribution of a discrete variable is through a *cumulative distribution function* (CDiF) which is the summation of the PMF and computes the probability of a variable being equal to or less than a particular value (x_i), so that:

$$f(x_i) = P(x \leq x_i), \qquad \text{for any } x_i \in R \qquad \text{(Equation 3.8)}$$

To provide a simple example of how the PMF and CDiF relate to each other, let's consider the typical case of a fair six-sided die. The assumed probability of any number in a single roll is given by the PMF:

$$f(x_i) = 1/6 \qquad \text{(Equation 3.9)}$$

which essentially means that *P(1) = P(2) = P(3) = P(4) = P(5) = P(6) =* 1/6. The corresponding CDiF for this example would be:

$$f(x_i) = \sum_{i=1}^{6} \big(i * (1/6)\big) \qquad \text{(Equation 3.10)}$$

which means that the function will result in the sequence {1/6, 2/6, 3/6, 4/6, 5/6, 6/6}. Individual values of the CDiF cannot be less than 0 or more than 1, but they always remain the same or increase.

When referring to continuous random variables, the probability for a single value is not defined since its mass is infinitely small, but the probability of an interval can be obtained with the integration of a *probability density function* (PDeF).

$$f(x_{interval}) = \int \frac{dF}{dx} \qquad \text{(Equation 3.11)}$$

A *cumulative density function* (CDeF) can also be used to describe continuous distributions of probabilities. They are defined in similar ways as for discrete variables and retain the same properties. The main characteristic of a PDeF is that the total integral must be 1. Just like the CDiF for a discrete variable is obtained through the summation of a PMF, in the case of continuous variables, the CDeF can be obtained through the summation of a PDeF.

$$f(x_i) = P(x \leq x_i), \qquad \text{for any } x_i \in R \qquad \text{(Equation 3.12)}$$

Bayesian inference

In this book we approach statistical inference through a Bayesian framework. In order to describe what this entails, and provide reasons for our choice, we will briefly discuss the approach used in classical hypothesis testing: the frequentist framework.

In frequentist inference, the parameters of a model are treated as fixed, even if unknown, and no probability can be associated with them. Any hypothesized value of a parameter cannot be assigned a probability, so statistical significance is determined by looking at how likely the data are under a null hypothesis (usually stated as the probability of the parameter taking the value of zero).

This notion of likelihood of the data given a parameter value is central to an approach called *maximum likelihood estimation* (MLE), in which we estimate the parameters of a model by finding the values that maximize the likelihood of the data. In other words, if we consider a range of potential values for the parameters of a model, which ones are more likely to produce the data in our sample?

While it is true that MLE is a powerful approach rooted in probability theory, it still fails to assign probability distributions to parameters. In other words, the likelihood function of a parameter is obtained by conditioning the data to different potential values of the parameter and, therefore, it is not a probability distribution function (i.e., it does not integrate to 1). In practical terms, this means that we cannot describe the probability of the hypothesized model parameter value given the data. Yet, this is our stated goal in ecological modeling.

So, how can we go from calculating the probability of the data given a parameter – $P(x|\delta)$ – to calculating the probability of a parameter given the data – $P(\delta|x)$? Luckily for us, Bayes' theorem provides an elegant way to connect these two conditional probabilities, so that:

$$P(\delta|x) = \frac{P(x|\delta) \times P(\delta)}{P(x)} \qquad \text{(Equation 3.13)}$$

This application of Bayes' theorem is known as Bayesian inference, and it presents many advantages. We can start identifying some of those advantages by looking at the different parts of this equation. The term $P(\delta|x)$ is commonly known as the *posterior* and, as stated, represents the

probability distribution of the parameter given the data. The term $P(x|\delta)$ is simply known as the *likelihood.* The new term that appears, P(σ), is known as the *prior*, and it represents any knowledge that we possess *a priori* regarding the probability distribution of the parameter. The use of prior information represents one of the main advantages of Bayesian inference and will be treated more carefully in Chapter V. The last term – $P(x)$ – is a normalizing factor that warrants that the posterior is in the range between 0 and 1.

Finding the posterior analytically can only be done in simple cases, for instance by calculating the mean when the shape of the probability distribution of variance can be clearly identified. For most parameters which cannot be solved analytically, there are several ways to estimate their posterior probability distributions. Some rely on assumptions regarding the shape of the posteriors. More flexible procedures that do not require any assumption of distribution usually take much longer and require computation. The estimation of the posterior distribution with these approaches is accomplished using a stochastic process known as Markov Chain Monte Carlo (MCMC) estimation (Hobbs and Hooten 2015). The development of tools to obtain reliable and efficient MCMC estimates is progressing at a fast rate. In the analyses we develop in this book, we depend on the available software STAN (mc-stan.org) to estimate the posterior distributions of our parameters. Discussion on how it works is beyond the scope of this book and we encourage the readers to better inform themselves. There are several excellent sources that explain how these procedures work (for example, Hobbs and Hooten 2015, McElreath 2016).

Modeling randomness: probability distributions

There are many probability distributions that can be used to describe data. Selecting which one to use in our models is a process that requires careful consideration of many questions regarding the nature of the phenomenon and how it is measured. For instance, we should consider what range and types of values the data can take. As mentioned earlier, one of the main ways to classify distributions is by looking at whether they are discrete or continuous. Hilborn and Mangel (1997), Matthiopoulos (2011) and Dietze (2017) provide excellent introductions to these distributions. Here, we briefly present a non-exhaustive list of distributions that we have found to be useful and then show an example of how we might select among them.

Discrete distributions

► Binomial

The binomial distribution is used to represent the incidence of an outcome out of a total number of independent trials (called Bernoulli trials), where only two outcomes are possible. We might use it, for instance, to describe the proportion of reproductive plants or the survival probability from one season to the next. The PMF for a binomial distribution is:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \qquad \text{(Equation 3.14)}$$

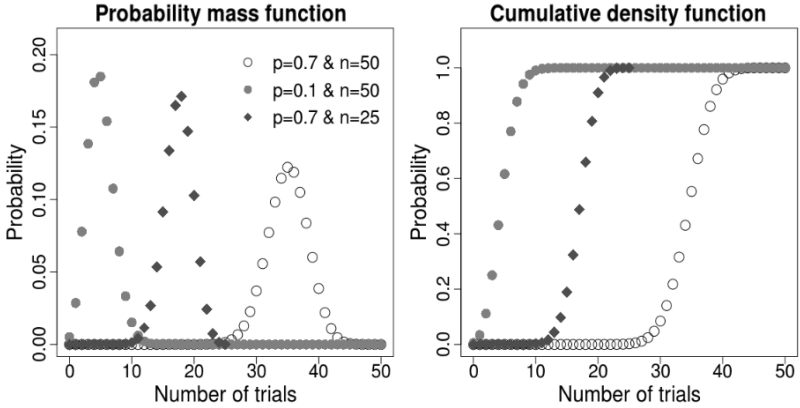

Figure 3.4: Probability mass functions (left) and cumulative distribution functions (right) for binomial distributions with varying levels of *p* and *n*

where *k* is the number of times the outcome of interest occurs, *n-k* the number of times that the alternative outcomes occur, *n* the total number of outcomes and *p* is the probability of the focal outcome. The CDF would simply be the iterative sum of values between 0 and n, so that:

$$P(X \leq k) = \sum_{i=0}^{n} \binom{n}{i} p^i (1-p)^{n-i} \qquad \text{(Equation 3.15)}$$

The mean of the binomial distribution is *np* while its variance is *np(1-p)*. The shape and the location of the binomial functions on the abscissa (*x* axis), then, will depend on the values of *p* and *n* (Figure 3.4).