3 Principles of Probability

3.1 Why Bother with First Principles?

Progress in science is made by comparing predictions of models with observations. All models make imperfect predictions of the operation of nature. A crosscutting theme in this book is that statistics help us solve the essential problem of science: gaining insight about phenomena of high dimension by the judicious reduction of their dimensions. Models reduce manifold influences on ecological processes to a few that can be understood. This means that models are inherently, deliberately approximate by virtue of the dimensions omitted.

To make a proper comparison between a model and observations, we need to understand the approximation inherent in models in terms of uncertainty. The output of the deterministic models we discussed in section 2.2 is a scalar or a vector; that is, for any given set of parameter values and input the model returns exactly the same result. In contrast, the output of a stochastic model is one or more probability distributions that reflect the uncertainty inherent in our model's predictions of a state and the way we observe it.

We use the term *stochastic* to refer to things that are uncertain.[1] Stochasticity arises in models in different ways, with different implications for sampling, experimental design, and forecasting. We mentioned these different sources of uncertainty earlier and will bring them up again and again.

[1] Stochastic models are used in the theoretical modeling tradition to represent random variation occurring over time and space not described by the deterministic core of the model. These models are often analyzed without reference to any data. We will not deal with these kinds of models, but we acknowledge that they are stochastic.

The main sources are the following:

1. **Process variance:** Process variance includes the uncertainty that results because our model fails to represent all the forces causing variation in the ecological quanties we seek to understand. Good models have low process variance because they capture most of the variation in the state they predict. Poor models have high process variance because they can't explain that variation. We often use the process variance to evaluate the fidelity of our model to the processes it represents, which means that it is critical to separate process variance from other sources of uncertainty. Failing to do so may lead to false conclusions about the model—we may have a great model and a poor system for observing the quantities it predicts. Lumping uncertainty about the process with uncertainty about the observations into the same "error term" makes it difficult to evaluate models and can lead to erroneous conclusions about the operation of ecological processes (Dennis et al., 2006; Hefley et al., 2013; Ahrestani et al., 2013).[2] The only way to reduce process variance is to improve our model. Expending more effort to observe the state of interest, improving our instruments, increasing our replications, and expanding the area we sample will do nothing to change process variance.
2. **Observation variance**: We rarely observe perfectly what we seek to understand. Observation variance quantifies these imperfections. There are usually two causes of observation variance. We seek to understand the true state of large areas or many individuals or many points in time, but we are forced by practicality to observe only a sample of them. Our sample is never a perfect reflection of the true state. In addition, we may need to correct our observing system for bias, and that correction is itself uncertain. In both cases, we can reduce uncertainty by taking more observations. Sampling variance asymptotically approaches zero as the number of observations increases. Models correcting for bias in our observations also become more certain with more observations.
3. **Variation among individuals:** Individual organisms differ because of their genetics and their individual histories—characteristics that may be hidden to us. These individual differences create uncertainty when we seek to understand responses of individuals to treatments or to

[2]Sometimes, particularly when designs are unreplicated, this separation is not possible, and lumping process and observation variance may be the best we can do. Section 6.3 covers this in detail.

environmental variation. The same idea can apply to spatial locations, which also have unique attributes.

4. **Model selection uncertainty:** The inferences we make based on a model depend on the model we choose from among many possible alternatives. The uncertainty that arises from our particular choice is called *model selection uncertainty*. We believe that the scientific objectives for the model trump formal model selection procedures, so our view is that sometimes we can ignore model selection uncertainty altogether, at other times we must quantify the uncertainty associated with using one model over another.

Dealing with uncertainty requires the proper tools, and primary among them are the rules of probability and an understanding of probability distributions. Equipped with these, ecologists can analyze the particular research problem at hand regardless of its idiosyncrasies. These analyses extend logically from first principles rather than from a particular statistical recipe. In the sections that follow, we describe these principles. Our approach is to start with the definition of probability and develop a logical progression of concepts extending from it to a fully specified and implemented Bayesian analysis appropriate for a broad range of research problems in ecology.

3.2 Rules of Probability

Ecological research requires learning about quantities that are unobserved from quantities that are observed. Any quantity that we fail to observe, including quantities that are observed imperfectly, involves uncertainty. The Bayesian approach treats all unobserved quantities as random variables to capture that uncertainty. A *random variable* is a quantity that can take on values due to chance—it does not have a single value but instead can take on a range of values. The chance of each value is governed by a probability distribution. We cannot underestimate the importance of this concept. Bayesian analysis is the only branch of statistics that treats all unobserved quantities as random variables. We will return to this idea in chapter 5.

All random variables have probability distributions even though these distributions may be unknown to us. The rules of probability determine how we gain insight about random variables from the distributions that govern their behavior. Understanding these rules lays a foundation for the remainder of the book. This material is not exactly gripping, but we urge you not to skip this section or rush through it unless you are already well grounded in formal principles of probability. Understanding these principles will serve you well.

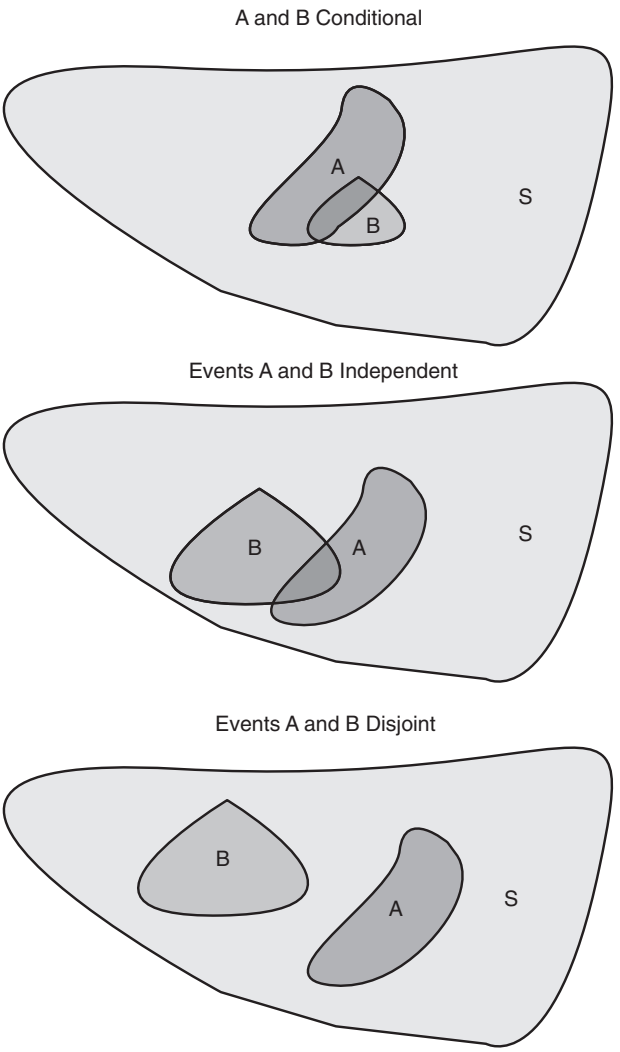


Figure 3.2.1. Illustration of conditional, independent, and disjoint probabilities. The area S defines a sample space including all the possible outcomes of a sample or an experiment. There are two sets of realized outcomes, A and B. The area of each event is proportional to the size of the set. The probability of $A =$ area of A/area of S and the probability of $B =$ area of B/area of S. Knowledge that event A has occurred influences our assessment of the probability of B when the intersection of the two events gives us new information about the probability of B. In this case, the probability of B is conditional on A, and vice versa (upper panel). In other cases the events intersect, but there is no new information. In this case the probability of B given A is

We start with the idea of a *sample space, S,* consisting of a set of all possible outcomes of an experiment or a sample, shown graphically as a polygon with a specific area (fig. 3.2.1). One of the possible outcomes of the experiment or sample is the random variable, "event A," a set of outcomes, which we also depict as a polygon (fig. 3.2.1). The area of A is less than the area of S because it does not include all possible outcomes. The area of A is proportional to the size of the set of outcomes it *does* include. It follows that the probability of A is simply the area of A divided by the area of S.

We now introduce a second event B to illustrate the concept of conditional, independent, and disjoint probabilities—concepts critical to understanding and applying Bayes' theorem to ecological models (chapters 5 and 6) Consider the case when we know that the polygon defining event B intersects the A polygon (fig. 3.2.1 upper panel) and, moreover, we know that event A has occurred. We ask, What is the probability of the new event B given our knowledge of the occurrence of A? The knowledge that A has occurred does two things: It shrinks the sample space from all of S to the area of A—if we know A has occurred, we know that everything outside of A has *not* occurred, so in essence we have a new, smaller space for defining the probability of A. Knowing that A has happened also affects what we know about B—we know that everything within B outside of A has not occurred (fig. 3.2.1). This means that

$$\Pr(B|A) = \frac{\text{area shared by } A \text{ and } B}{\text{area of } A}. \tag{3.2.1}$$

Dividing the numerator and denominator of the right-hand side by S we turn the areas into probabilities:[3]

$$\Pr(B|A) = \frac{\Pr(A \cap B)}{\Pr(A)} = \frac{\Pr(A, B)}{\Pr(A)}. \tag{3.2.2}$$

[3]The set operator $\cap$ means "intersection of."

Figure 3.2.1 (*Continued*).
the same as the probability of B, because area of B shared with A/area of A = area of B/area of S. In this case, we say that A and B are independent (middle panel, areas drawn approximately, but you really can't tell if one event is conditional on another by simply looking at the diagram unless you can see proportionality perfectly! If there is no intersection, then the events are disjoint. Knowing that A has occurred means that we know that B has not occurred (bottom panel). Thus, disjoint probabilities are a special case of conditional probability where $\Pr(A|B) = 0$.

Using the same logic, we obtain

$$\Pr(A|B) = \frac{\text{area shared by } A \text{ and } B}{\text{area of } B} = \frac{\Pr(A \cap B)}{\Pr(B)} = \frac{\Pr(A, B)}{\Pr(B)}. \quad (3.2.3)$$

The expression $\Pr(A|B)$ reads, "the probability of A conditional on knowing B has occurred." The bar symbol (i.e., |) reads "conditional on" or "given," expressing the dependence of event A on event B; if we know B, our knowledge changes what we know about A. It is important to note that $\Pr(A|B) \neq \Pr(B|A)$. The expression $\Pr(A, B)$ reads, "the *joint* probability of A and B" and is interpreted as the probability that both events occur. We will make important use of the algebraic rearrangement of equations 3.2.2 and 3.2.3 to expand their joint probability,

$$\Pr(A, B) = \Pr(B|A)\Pr(A) \quad (3.2.4)$$

$$= \Pr(A|B)\Pr(B).$$

In some cases the area defining the two events overlaps, but no new information results from knowing that either event has occurred (fig. 3.2.1 middle panel). In this case the events are *independent*. Events A and B are independent if and only if

$$\Pr(A|B) = \frac{\text{area of } A \text{ shared by } A \text{ and } B}{\text{area of } B} = \frac{\text{area of } A}{\text{area of } S} = \Pr(A), \quad (3.2.5)$$

or equivalently,

$$\Pr(B|A) = \Pr(B). \quad (3.2.6)$$

Using equations 3.2.1 and 3.2.3, we can substitute for the conditional expressions in equations 3.2.5 and 3.2.6. A little rearrangement gives us the joint probability of independent events:

$$\Pr(A, B) = \Pr(A|B)\Pr(B) = \Pr(A)\Pr(B). \quad (3.2.7)$$

It is important to throughly understand the difference between the joint probability of events that are independent (eq. 3.2.7) and those that are not (eq. 3.2.4). When events are disjoint, there is no overlap between them (fig. 3.2.1 lower panel). In this case, knowing that one event has occurred means that we know the other event has *not* occurred. Thus, events that

are disjoint are a special case of conditional probability: knowledge of one event gives us complete knowledge of the other event.

We may also be interested in the probability that one event or the other occurs (fig. 3.2.1), which is the total area of A and B not including the area they share[4], that is,

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A, B). \tag{3.2.8}$$

When A is independent of B,

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A)\Pr(B), \tag{3.2.9}$$

but if the events are conditional,

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A|B)\Pr(B), \tag{3.2.10}$$

or equivalently,

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(B|A)\Pr(A). \tag{3.2.11}$$

If A and B are disjoint, then

$$\Pr(A \cup B) = \Pr(A) + \Pr(B), \tag{3.2.12}$$

which is simply a special case of equation 3.2.8 where $\Pr(A, B) = 0$.

The final probability rule we consider applies when we can partition the sample space into several nonoverlapping events (fig. 3.2.2). This rule is important because we will use it later to understand the components of Bayes' theorem (chapter 5). We define a set of events $\{B_n : n = 1, 2, 3, \ldots\}$, which taken together, cover the entire sample space, $\sum_n B_n = S$. We are interested in the event A that overlaps one or more of the B_n. The probability of A is

$$\Pr(A) = \sum_n \Pr(A \mid B_n)\Pr(B_n). \tag{3.2.13}$$

[4]The $\cup$ operator means "union of."

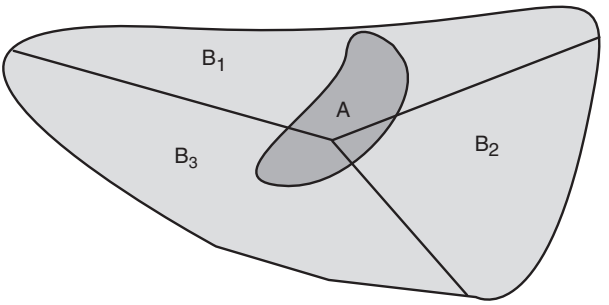


Figure 3.2.2. Illustration of the law of total probability.

Equation 3.2.13 is called the *law of total probability*. As the number of events approaches infinity and the areas of events become infinitesimally small, equation 3.2.13 becomes[5]

$$\Pr(A) = \int [A|B]\,[B]\,dB. \tag{3.2.14}$$

3.3 Factoring Joint Probabilities

It is hard to avoid a modicum of tedium in describing the rules of probability, but there is a very practical reason for understanding them: they allow us to deal with complexity. These rules permit us to take complicated joint distributions of random variables and break them down into manageable chunks that can be analyzed one at a time as if all the other random variables were known and constant. The importance of this idea and its implementation will be developed throughout the book, particularly in chapters 6 and 7. Here, we establish its graphical and mathematical foundation. What you learn here is critical to the model specification step in the general modeling process we described in the preface (fig. 0.0.1B). The material will become clear as we apply it to modeling problems.

[5]A bit about integral notation is needed here. Ecologists with some training in calculus likely recognize the definite integral $\int_L^U [A|B]\,[B]dB$ as the area under the curve $[A|B][B]$ from L to U. When no interval is specified (the $\int$ is "naked"), we denote integration over the domain of the function $[A|B][B]$. Thus, the notation $\int[A|B][B]dB$ means the definite integral of $[A|B]\,[B]$ over *all* possible values of B, the prevailing convention in statistics. It is important not to confuse this convention with that often used in mathematics where $\int[A|B][B]dB$ denotes the indefinite integral, that is, the antiderivative of $[A|B][B]$.

Consider the networks shown in figure 3.3.1. A Bayesian network (also called a *directed acyclic graph*) depicts dependencies among random variables. The random variables in the network are called *nodes*. The nodes at the head of the arrows are charmingly called *children*, and the tails, *parents*. Bayesian networks show how we factor the joint probability distribution of random variables into a series of conditional distributions and thereby represent an application of equation 3.2.4 to multiple variables (fig. 3.3.1). We use factoring to simplify problems that would otherwise be intractably complex.

Bayesian networks are great tools for thinking about relationships in ecology and for communicating them (e.g., fig. 1.2.1). They are useful because they allow us to visualize a complex set of relationships, thus encouraging careful consideration of how knowledge of one random variable informs us about the behavior of others. They lay plain our assumptions about dependence and independence. A properly constructed Bayesian network provides a detailed blueprint for writing a joint distribution as a series of conditional distributions: nodes at the heads of arrows are on the left-hand side of conditioning symbols, those at the tails of arrows are on on the right-hand side of conditioning symbols, and any node at the tail of an arrow without an arrow leading into it must be expressed unconditionally, for example, $\Pr(A)$. The network provides a graphical description of relationships that is perhaps easier to understand than the corresponding mathematical description, facilitating communication of ecological ideas underlying the network.[6]

The mathematics allowing factoring of joint distributions extends directly from the rules of probability we have already developed. Given the vector of jointly distributed random variables $\mathbf{z} = (z_1, \ldots, z_n)'$, the joint probability of the variables satisfies

$$\Pr(z_1, \ldots, z_n | p_1, \ \ldots, \ p_n) = \prod_{i=1}^{n} \Pr(z_i | \{p_i\}), \tag{3.3.1}$$

where $\{p_i\}$ is the set of parents of node z_i, and all the terms in the product are independent.[7] Independence of the terms in equation 3.3.1 is assured if the equation has been properly constructed from a Bayesian network, and the network shows relationships that are conditional and independent.

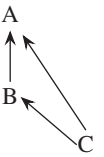
[6] At least Hobbs thinks so. Hooten prefers the equations.

[7] The operator $\prod_{i=1}^{n}$ says take the product of everything with the subscript i over $i = 1, \ldots, n$. It is the multiplicative equivalent of the summation operator, $\sum_{i=1}^{n}$, which may be more familiar to ecologists.

I

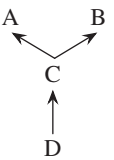
$$\Pr(A, B) = \Pr(A|B)\Pr(B)$$

II



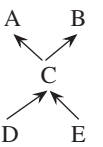
$$\Pr(A, B, C) = \Pr(A|B, C) \times \Pr(B|C)\Pr(C)$$

III



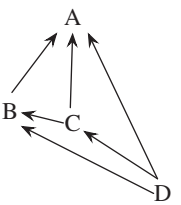
$$\Pr(A, B, C, D) = \Pr(A|C) \times \Pr(B|C)\Pr(C|D)\Pr(D)$$

IV



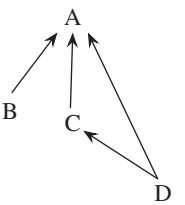
$$\Pr(A, B, C, D, E) = \Pr(A|C) \times \Pr(B|C)\Pr(C|D, E) \times \Pr(D)\Pr(E)$$

V



$$\Pr(A, B, C, D) = \Pr(A|B, C, D) \times \Pr(B|C, D) \times \Pr(C|D)\Pr(D)$$

VI



$$\Pr(A, B, C, D) = \Pr(A|B, C, D) \times \Pr(C|D) \times \Pr(B)Pr(D)$$

Figure 3.3.1. Bayesian networks specify how joint distributions are factored into conditional distributions using nodes to represent random variables and arrows to represent dependencies among them. Nodes at the heads of arrows must be on the left-hand side of conditioning symbols (|); nodes at the tails of arrows are on on the right-hand sides of conditioning symbols. Any node at the tail of an arrow without an arrow leading into it must be expressed unconditionally, for example, Pr(*A*). Some of the examples also indicate independence. The random variables *A* and *B* are independent after accounting for their mutual dependence on *C* in graphs III and IV; *D* and *E* are independent in IV, and *B* is independent of *C* and *D* in VI.

A somewhat more formal way to say the same thing is to generalize the conditioning rule of probability for two random variables (eq. 3.2.3) to factor the joint distribution of any number of random variables using

$$\Pr(z_1, z_2, \ldots, z_n) = \Pr(z_n|z_{n-1}, \ldots, z_1) \ldots \Pr(z_3|z_2, z_1) \Pr(z_2|z_1) \Pr(z_1), \tag{3.3.2}$$

where the components z_i may be scalars or subvectors of $\mathbf{z}$, and the sequence of the conditioning is arbitrary.[8] It is important to see the pattern of conditioning in equation 3.3.2. We can use the independence rule of probability (eq. 3.2.5) to simplify conditional expressions in equation 3.3.2 for random variables known to be independent. For example, if z_1 is independent of z_2, then $\Pr(z_1|z_2)$ simplifies to $\Pr(z_1)$. If z_1 and z_2 depend on z_3 but not on each other—which is to say they are conditionally independent—then

$$\Pr(z_1, z_2, z_3) = \Pr(z_1|z_2, z_3) \Pr(z_2|z_3) \Pr(z_3) \tag{3.3.3}$$

simplifies to

$$\Pr(z_1, z_2, z_3) = \Pr(z_1|z_3) \Pr(z_2|z_3) \Pr(z_3). \tag{3.3.4}$$

Another example of this kind of simplification is shown graphically and algebraically in figure 3.3.1 V and VI. Don't let the formalism in this paragraph put you off. It is simply a compact way to say what we have already shown graphically using Bayesian networks, which for many ecologists will be more transparent.

3.4 Probability Distributions

3.4.1 Mathematical Foundation

The Bayesian approach to learning from data using models makes a fundamental simplifying assumption: we can divide the world into things that are observed and things that are unobserved. Distinguishing between the observable and unobservable is the starting point for all analyses.

[8] We say the sequence is arbitrary to communicate the idea that the ordering of the specific z_i is not required for equation 3.3.2 to be true. In other words, z_n doesn't need to come first. However, the word *arbitrary* should not be taken to mean capricious. As we will learn, it is our understanding of the *biology* that determines what is conditional on what and ultimately governs the sequence of conditioning.

We treat all unobserved quantities as random variables governed by probability distributions, because the things we cannot observe have inherent uncertainty. It follows that understanding probability distributions forms a critical link between models and data in ecology. Becoming familiar with these distributions is the key to developing a flexible approach to the analysis of ecological models and data. Equipped with a toolbox of deterministic models (Otto and Day, 2007, chap. 2; Bolker, 2008) and the probability models described here, you will be able to thoughtfully develop a coherent approach to analyzing virtually any problem in research, regardless of its nuances. At the very least, you will be able to compose an approach that can be discussed productively with your statistical colleagues.

In this section we provide a compact description of distributions commonly used in developing models for ecological data. We first describe the features shared by all probability distributions and then outline features of specific distributions that we will use frequently in later chapters. We organize this section using the two types of random variables, discrete and continuous. Discrete random variables are those that take on discrete values, usually integers. It is possible, however, for discrete random variables to take on non-integer values. For example, a random variable might have support $\{0, \frac{1}{2}, 1\}$. In this case it is discrete but not integer-valued. All discrete random variables in this book will be integers, usually counts of things or membership in categories. In contrast, continuous random variables can take on an infinite number of values on any interval to represent length, mass, time, and energy, for example.

Probability mass functions and probability density functions are the fundamental link between models of ecological processes and observations of those processes. We first explain probability mass functions, then turn to probability density functions.

3.4.1.1 Probability Mass Functions

Assume we are interested in a discrete random variable z. That random variable might be the number of zooplankton in a 1L sample from a lake. The quantity $z = 304$ is a specific value that might be observed for that sample. The random variable could also be the number of individual plants in four categories: native annuals, native perennials, exotic annuals, and exotic perennials, for example, $\mathbf{z} = (4, 6, 18, 3)'$.[9] These examples illustrate an important point that we will return to more than once. Before we observe a quantity like the number of zooplankton in a sample, the quantity is a random variable whose value is governed by a probability distribution.

[9]We will use notation $\mathbf{u} = (a, b, c)'$ to indicate the vector $\mathbf{u}$ with elements a, b and c. The $(')$ following the () indicates that the variables within the parentheses are elements of a column vector.

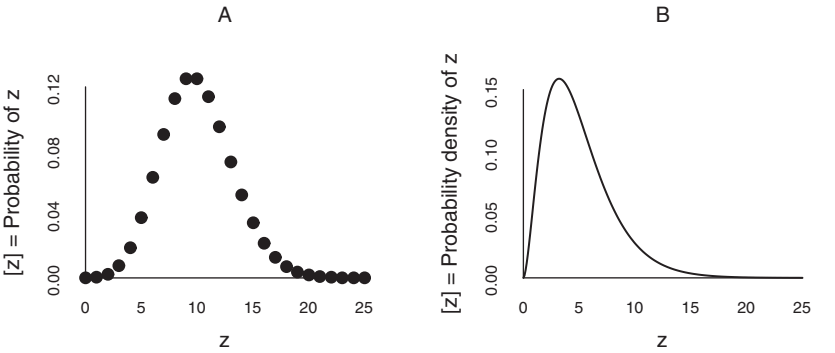


Figure 3.4.1. (**A**) A probability mass function computes the probability that a discrete random variable takes on a single value. Shown here is the distribution of a Poisson random variable with mean = 10. (**B**) A probability density function computes the probability density of a continuous random variable at a point. The integral of the function between two points gives the probability that the random variable falls in the interval between the points. Shown here is the distribution of a gamma random variable with mean = 5 and variance = 9.

After it is observed, the quantity becomes a known, specific value. A *probability mass function*,[10] $[z]$, for the random variable is

$$[z] = \Pr(z), \tag{3.4.1}$$

which simply says that given the argument z, the function $[z]$ returns the probability that the random variable will take on the particular value z (fig. 3.4.1, box 3.4).

All probability mass functions share two properties:

$$0 \leq [z] \leq 1, \tag{3.4.2}$$

$$\sum_{z \in S} [z] = 1, \tag{3.4.3}$$

where S is the *support* of the random variable z, that is, the set of all values of z for which $[z] > 0$. Support is a vital concept because it defines the domain of the function $[z]$; values of z outside that domain have zero probability of occurrence and in some cases are not defined. Equation 3.4.2 says that the value of $[z]$ must be between 0 and 1, which of course makes sense if $[z]$ is a probability. Equation 3.4.3 says that the sum of $[z]$ over all its possible values must equal 1, which, is again, sensible for a probability mass function.

[10]Some statisticians use *probability function* synonymously with *probability mass function*.

Box 3.4 Notation for Probability Distributions

Statisticians read and write notation every day, allowing them to become accustomed to differences in notational styles that frequently occur in the literature. However, these differences can be confusing for ecologists who don't use notation as frequently as statisticians do. Here we explain the notation we will use to represent probability distributions and show how it relates to other widely used notations that mean the same thing.

Our Notation

First introduced in the seminal paper of Gelfand and Smith (1990), brackets have become a preferred notation for ecologists and statisticians using Bayesian methods, because they allow complex, multidimensional models to be written in compact form. We will use the notation $[z]$ to mean the probability of the random variable z if z is discrete and the probability density of z if z is continuous. Thus, $[z]$ denotes "z is distributed as." We will not include additional arguments within brackets when we refer to probability distributions broadly defined. If we are writing about specific distributions and want to refer to parameters, we will use $[z|\alpha, \beta]$ to denote the probability or probability density of z conditional on α and β.

We will often name the specific distribution; for example, we will write gamma $(z|\alpha, \beta)$ to denote that the random variable z follows a gamma distribution with parameters α and β. We will use $z \sim \text{gamma}(\alpha, \beta)$ to mean the same thing. When we refer specifically to the probability of z (excluding probability density) we will use the notation $\Pr(z)$.

We will unapologetically use somewhat unconventional notation to achieve clarity when we want to specifically delineate the deterministic and stochastic components of models. So, for example, we might be interested in a deterministic model $\mu_i \equiv g(\boldsymbol{\theta}, x_i)$ that predicts the central tendency of the distribution of the random variable y_i, that is, $[y_i|\mu_i, \sigma^2]$, where σ^2 is a parameter that controls the dispersion of the distribution. (See chapter 1 for examples.) Equivalently, we will often use $\left[y_i|g(\boldsymbol{\theta}, x_i), \sigma^2\right]$. Of course, we realize that not all distributions have means and variances as parameters, so this notation runs some risk of being confused with the normal distribution when we intend it to refer to the full family of probability distributions. However, we have found in our teaching that it can be very helpful to call specific attention to the deterministic part and the stochastic

(*continued*)

(Box 3.4 *continued*)

part of models. We will reemphasize this point to prevent confusion with the normal distribution throughout the book. Moreover, we will teach how quantities representing central tendency and dispersion can be properly matched to specific parameters of distributions in section 3.4.4.

Notation Used by Others

Bracket notation for distributions is synonymous with the following. Often, mathematical statistics texts use the notation $P(Z = z) = f(z)$ or $P(Z = z) = f_Z(z)$ to mean the probability that random variable Z takes on a specific value z is given by the probability mass function $f(z)$. So, $P(Z = z) = f(z)$ is the same as $[z]$. Sometimes, authors reserve $p(Z = z)$ to refer to probability density, and $\Pr(Z = z)$ to refer to probability. Similarly, the notation $P(z|\alpha, \beta)$ means the same thing as $[z|\alpha, \beta]$, which is identical with $f(z|\alpha, \beta)$ and $f(\alpha, \beta)$ when f has been defined as a probability mass function or probability density function for z. We prefer the bracket notation because it is simpler.

Statisticians often write probability mass functions and probability density functions with only two arguments, the parameters, without specifying the random variable. For example they might write the distribution of the random variable z as gamma (α, β). To err on the side of clarity, we will include the random variable, that is, gamma$(z|\alpha, \beta)$.

3.4.1.2 Probability Density Functions

Probability density functions apply to random variables that are continuous, taking on values that are real numbers instead of integers. Given a continuous random variable z, a probability density function $[z]$ has the characteristics

$$[z] \geq 0, \tag{3.4.4}$$

$$\Pr(a \leq z \leq b) = \int_a^b [z]\, dz, \tag{3.4.5}$$

$$\int_{-\infty}^{\infty} [z]\, dz = 1\,, \tag{3.4.6}$$

which says that the probability density of z is nonnegative, that the probability of z falling between a and b is the integral of the density function from a to b, and that the integral of the density function over all real numbers equals 1.

It is important to understand that probability density functions do not return a probability, as probability mass functions do. Instead, they return a *probability density*. For continuous random variables, probability is defined only for a range of values, that is, $\Pr(a \leq z \leq b)$. The support for a continuous random variable includes all the values of z for which the probability density exceeds 0; that is, $[z] > 0$.

Some intuition for probability density can be gained by thinking about how we would approximate an integral of a probability density function $[z]$ over some range $\Delta z = b - a$ using a rectangular column with height $[(a+b)/2]$ and width Δz, remembering that the brackets indicate a function that returns the probability density of whatever is enclosed within them (fig. 3.4.2). Thus, we can think of $[(a + b)/2]$ as the average height of a bar over the interval from a to b. The probability of z over the interval a to b is $\Pr(a \leq z \leq b) \approx \Delta z[(a + b)/2]$. Thus, for very small Δz, the probability density of z is $[z] \approx \Pr(a \leq z \leq b)/\Delta z$.

We have found in our teaching that a potential point of confusion for many ecologists is that the y-axis for plots of probability mass functions always ranges between 0 and 1, while the y-axis for probability density functions can take on any value greater than 0. Students often ask, "How can the area under the probability density curve equal 1 if there are values on the y-axis that are greater than 1? Why is the axis so different for different random variables?" These questions arise from forgetting the definition of a definite integral, an easy thing to do for those of us who may use integrals infrequently. Recall that when we integrate, we are summing the area of bars (their heights $\times$ widths) under a curve where the number of bars approaches infinity, and the width of bars approaches zero. Thinking of integrals as the sum of the areas of many bars is the key to understanding the values on the y-axis of probability density functions (fig. 3.4.2, inset). The area depends on the height of the bars and the *widths*, which means that the scale of the y-axis of a probability density function depends on the scale of the x-axis. Any nonnegative value can appear on the y-axis for probability density, because those densities depend on the values of the random variable on the x-axis. This must be true to ensure that the integral over the entire support of x equals 1.

3.4.1.3 Moments

Important properties of probability distributions can be summarized succinctly using their *moments*. The first moment describes the central ten-

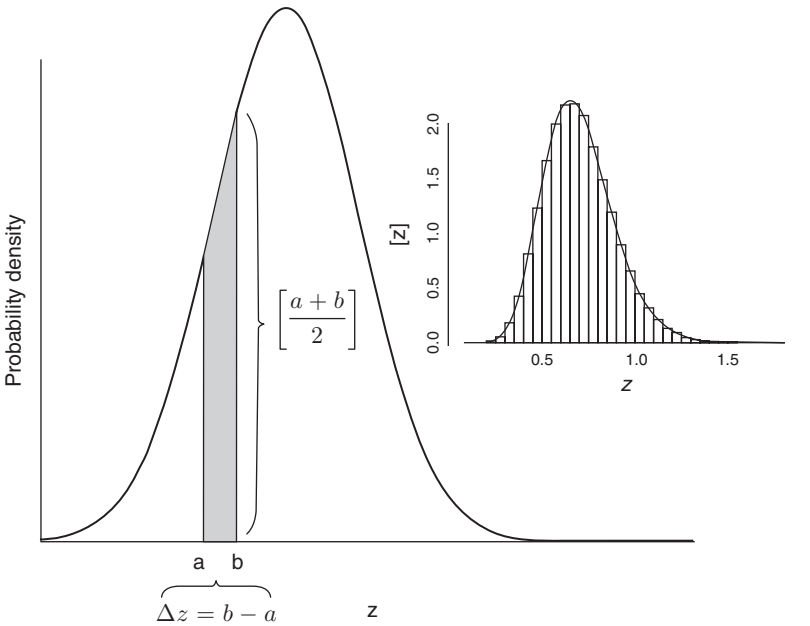


Figure 3.4.2. Illustration of probability and probability density for continuous random variables. When random variables are continuous, probability is defined only for intervals of values of the random variable. For example, the probability that the random variable z is within the interval a to b is the area of the shaded region, $\Pr(a \leq z \leq b) = \int_a^b [z]dz$. Probability density is the height of the shaded area at zero width, that is, $[z] = \Pr(a \leq z \leq b)/\Delta z$ with Δz infinitesimally small. Recall that as Δz approaches zero, Δz becomes dz in $\int_a^b [z]dz$. The inset demonstrates that the probability density, unlike probability can be greater than 1.

dency of the distribution, and the second central moment describes the dispersion or spread of the distribution. For the discrete random variable z, the first moment is the expected value of z, that is, the mean of its distribution:

$$\mu = \mathrm{E}(z) = \sum_{z \in S} z[z]. \tag{3.4.7}$$

Equation 3.4.7 says the expected value of the random variable z is the sum of all possible values of z, each multiplied by its probability. Thus, the expected value is a weighted average of values of the random variable, where the weights are probabilities, $[z]$. The second central moment of the distribution of discrete random variables, the *variance*, is the expected value

of the squared difference between the value of the random variable and the mean of the random variable

$$\sigma^2 = \mathrm{E}\left((z-\mu)^2\right) = \sum_{z\in S}(z-\mu)^2\,[z]\,. \tag{3.4.8}$$

For continuous random variables, we integrate rather than sum to obtain the moments:

$$\mu = \mathrm{E}\,(z) = \int_{-\infty}^{\infty} z[z]dz; \tag{3.4.9}$$

$$\sigma^2 = \mathrm{E}\left((z-\mu)^2\right) = \int_{-\infty}^{\infty}(z-\mu)^2[z]dz. \tag{3.4.10}$$

There are additional moments; skewness is the third, and kurtosis is the fourth, but we will not use them in the material that follows.

It is important to know how we approximate the first and second moments of random variables, continuous or discrete, using a technique called *Monte Carlo integration*. If we make many random draws from the distribution $[z]$, then its mean is approximately

$$\mu = \mathrm{E}(z) \approx \frac{1}{n}\sum_{i=1}^{n} z_i, \tag{3.4.11}$$

where n is the number of draws, and z_i is the ith value of the draw[11] of random variable z. In a similar way we can approximate the variance as[12]

$$\sigma^2 = \mathrm{E}\left((z-\mu)^2\right) \approx \frac{1}{n}\sum_{i=1}^{n}(z_i-\mu)^2, \tag{3.4.12}$$

where μ is estimated using equation 3.4.11. It is important to understand these approximations, because random draws from a distribution form the fundamental basis for learning about distributions of parameters and latent

[11]The idea of making draws from a distribution to approximate its mean or other moments will be very important in subsequent chapters, particularly 7 and 8. Be sure you understand this concept.

[12]Don't confuse this formula with the variance of a small sample in frequentist statistics, which you may remember as $\sigma^2 \approx \frac{1}{n-1}\sum_{i=1}^{n}(z_i-\bar{z})^2$, using $1/(n-1)$ instead of $1/n$. The version with n in the denominator provides the maximum likelihood estimate of σ^2, an estimate that is biased. An unbiased estimate for a small sample is obtained using $1/(n-1)$. However, in sampling from a distribution, we make so many draws (n) that using $1/(n-1)$ versus $1/n$ has no practical effect on the estimate of the variance.

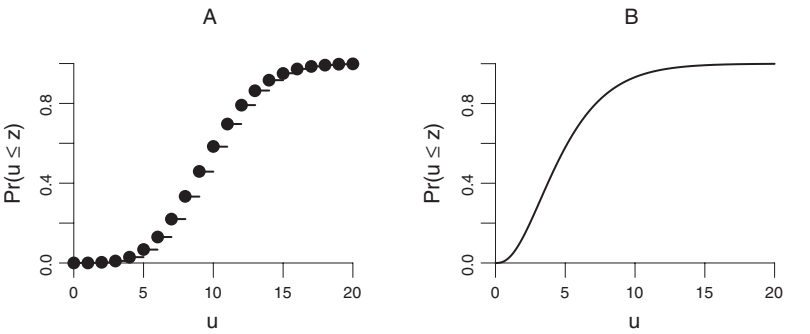


Figure 3.4.3. Cumulative distribution functions calculate the probability that a random variable takes on values less than or equal to a threshold, $F(z) = \Pr(u \leq z)$. Shown here are cumulative distribution functions for a Poisson random variable (**A**) with mean = 10 and a gamma random variable (**B**) with mean = 5 and variance = 9. Both functions asymptotically approach 1 as u approaches infinity. Corresponding probability mass and probability density functions are shown in figure 3.4.1.

quantities as well as quantities derived from them using Markov chain Monte Carlo, which we will treat in detail in chapters 7 and 8.

3.4.1.4 Distribution and Quantile Functions

We often wish to know the probability that a random variable takes on values above or below some threshold, for example, the probability that a population falls below 1000 individuals or that the proportion of plots containing a rare species exceeds 0.10. We do this by using the cumulative distribution function $F(z)$ for a discrete random variable[13] (fig. 3.4.3 A), which is defined as the probability that the value of the random variable is at most z:

$$F(z) = \sum_{u \leq z} [u]. \tag{3.4.13}$$

Given an argument z, the cumulative distribution function returns the probability that the random variable u is less than or equal to z, where $\Sigma_{u \leq z}\,[u]$ is the sum of all the probabilities of u for values of u less than or equal to z. For continuous random variables, the cumulative distribution

[13] Also called the *distribution function*.

function (fig. 3.4.3 B) is

$$F(z) = \int_{-\infty}^{z} [u]\, du. \tag{3.4.14}$$

The cumulative distribution function for a continuous random variable is the integral of the probability density function. It follows from the fundamental theorem of calculus that the density function is the derivative of the distribution function.

Finally, we introduce the *quantile function*, $F^{-1}(p)$, which we will often use for interval estimation of unobserved quantities. For a discrete random variable the quantile function returns the largest value of z for which $F(z) \leq p$, where p is the quantile of interest. For a continuous random variable, the quantile function is the inverse of the cumulative distribution function, $F^{-1}(p) = z$.

3.4.2 Marginal Distributions

Marginal distributions of random variables arise in many contexts in Bayesian modeling, so it is important to understand what they are. We start with the simplest case, two discrete random variables, x and y, that are jointly distributed:

$$[x, y] \equiv \Pr(x, y), \tag{3.4.15}$$

which simply means that $[x, y]$ is defined as ($\equiv$) the probability of the joint occurrence of x and y. The marginal distributions of x and y are most easily understood by example. Imagine that we are interested in a species for which births occur in pulses. We observe 100 females and record the age of each individual (as an integer) and also record the number of offspring she produced. We divide the frequency of each observed age and offspring combination by 100 to obtain the joint probabilities (see table 3.1). Cells in the table give the joint probability of age and number of offspring. The bottom row gives the marginal distribution of the number of offspring. The rightmost column gives the marginal distribution of age. Thus, y is "marginalized out" by summing to obtain the marginal distribution of y. The same idea applies to summing over x.

For a joint distribution of two random variables, we can focus on the probability of occurrence of one of them by summing over the probabilities of the other, effectively turning a bivariate distribution into a univariate one. If we are interested in the probability distribution of the number of offspring irrespective of the age of the mother, we ignore age by summing

TABLE 3.1
Example of Joint and Marginal Distributions for Age (x) and Number of Offspring (y)

	y = Number of Offspring			
x = Age	1	2	3	$\sum_y [x, y]$
1	0.1	0	0	0.1
2	0.13	0.12	0.02	0.27
3	0.23	0.36	0.04	0.63
$\sum_x [x, y]$	0.46	0.48	0.06	

down the columns in table 3.1 to obtain the *marginal distribution* of number of offspring. It is easy to see why this distribution is called marginal—it is based on the sums listed in the margins of the table. Thus, the marginal distribution of x is

$$[x] = \sum_y [x, y]$$
$$= \sum_y [x|y][y],$$

and the marginal distribution of y

$$[y] = \sum_x [x, y]$$
$$= \sum_x [y|x][x],$$

results that follow directly from the law of total probability (eq. 3.2.14). It is important to note that these are true probability distributions: $\sum_x [x] = 1$, and $\sum_y [y] = 1$.

Now, imagine that we add a third dimension to the data, sex of offspring, z, so that the joint distribution is now $[x, y, z]$, and the matrix of data in table 3.1 becomes a $3 \times 3 \times 2$ array. If we were interested in the probability distribution of male versus female offspring, we would sum over the probabilities of number of offspring and age, $[z] = \sum_x \sum_y [x, y, z]$. We could add any number of dimensions to the joint distribution and would follow the same procedure to focus on one of them. We sum over the probabilities of all the random variables except the one we are interested in. We obtain the distribution of the variables of interest by "marginalizing" over the distribution of the variables being discarded. The variables we leave out are said to have been "marginalized out."

We now consider random variables that are continuous rather than discrete (fig. 3.4.4). For example, we might be interested in the joint distribution of the mass of the mother, x, and the total mass of its offspring, y. When the joint random variables are continuous,

$$\begin{aligned}[x] &= \int [x, y] dy \\ &= \int [x|y][y] dy, \end{aligned} \tag{3.4.16}$$

where $\int [x|y][y]dy$ is the integral of $[x|y][y]$ over the support of y. Similarly,

$$\begin{aligned}[y] &= \int [x, y] dx \\ &= \int [y|x][x] dx. \end{aligned} \tag{3.4.17}$$

It may help you understand what this means by imagining subdividing the rows and columns in table 3.1 into increasingly smaller divisions and summing as before, except that now the numbers of rows and columns are infinite, requiring integration (fig. 3.4.4). But the concept illustrated above is exactly the same. Extending to multiple random variables, $[z_1, z_2, \ldots, z_n]$, we integrate over all the random variables except the one we seek to marginalize, an operation sometimes referred to as "integrating out." We accomplish the same thing as we did in the discrete case: we convert a joint distribution into a univariate distribution.

Integrals like those in equations 3.4.16 and 3.4.17 will appear frequently later in the book as a way to focus on the univariate distribution of unknown quantities that are parts of joint distributions that might contain many parameters and latent quantities. Thus, they are a vital tool for simplification. We urge you to be sure you understand what these integrals mean based on the simple example here before you proceed.

3.4.3 Useful Distributions and Their Properties

We now describe probability distributions that we have found to be most useful for modeling random variables in ecology. We describe key features of probability mass functions and probability density functions here, summarizing other aspects in appendix A. Most ecologists routinely use functions in software like R (R Core Team, 2013) to compute these functions. However, it is important to be familiar with the mathematical

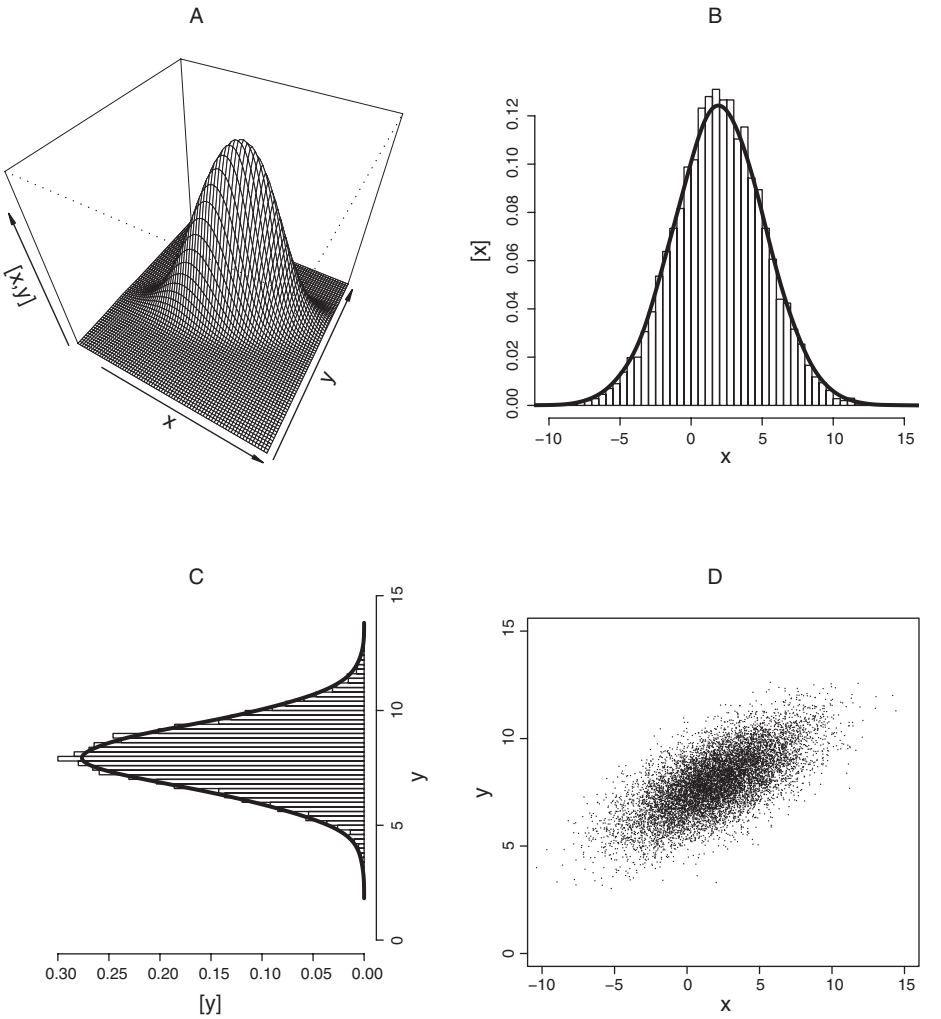


Figure 3.4.4. Examples of marginal distributions. Panel **A** shows a bivariate normal distribution for the correlated random variables x and y. The mean of x is 2, the mean of y is 8, and their covariance matrix is $\mathbf{\Sigma} = \begin{pmatrix} 10 & 3 \\ 3 & 2 \end{pmatrix}$. (If you are unfamiliar with covariance matrices, skip ahead momentarily to see box 3.4). Panel **D** shows 10,000 draws from the joint distribution of x and y. (**B**) Imagine that we "binned" all the observations of x to create a normalized histogram, ignoring the values of y. If the bins are infinitely small, we are "integrating out" y. (**C**) We similarly obtain the marginal density of y by integrating out x.

formulas that stand behind the computational versions of these functions because we will occasionally use these formulas later in this book, because mathematical expressions for distributions often appear in the literature, and because the expressions can be usefully manipulated, as will be illustrated with the exponential distribution later in this section.

One other concept is needed here. The material in section 3.4.1 treated probability distributions as if they had a single argument, z. This simplification made it easier to concentrate on the basic properties of probability mass functions and probability density functions undistracted by other arguments to the functions. However, specific distributions require *parameters* as arguments in addition to the random variable. *Parameters* are arguments to functions that give probability distributions a particular shape, that is, a central tendency, dispersion, and skew. Many ecologists are most familiar with the normal distribution for which the parameters and the first and second moments (i.e., the mean and variance) are the same. This is also true for the Poisson distribution, which has a single parameter, the mean, which is equal to the variance. For all other distributions, the moments and the parameters are not the same. Instead, the moments are *functions* of the parameters, and, hence, the parameters are functions of the moments. We will use these functional relationships between moments and parameters in a powerful way in section 3.4.4. For now, we highlight differences between random variables and parameters by using a consistent notation. We will denote random variables as z,[14] and we will use Greek letters to symbolize parameters.

3.4.3.1 Probability Mass Functions for Discrete Random Variables

Poisson. The Poisson distribution describes the probability of a number of events (z) occurring in a given unit of time or space assuming that the occurrence of one event has no influence on the probability of occurrence of the subsequent event (fig. 3.4.5). The distribution has a single parameter, λ, the average number of occurrences, also called the *intensity*. Note that λ is a positive real number, whereas z must be a nonnegative integer.[15]

[14]We typically use lowercase letters to denote univariate random variables throughout. Keep in mind that it is also common to see uppercase letters for random variables and lowercase letters for "realizations" of random variables (i.e., their numerical values). We use uppercase letters for matrices, so to avoid inconsistencies, we let the context dictate whether a lowercase letter denotes a random variable or a value. We find that our notation does the job most of the time and leads to substantially less confusion than the somewhat more conventional notation.

[15]Negative values for z are not defined because $z!$ is not defined for $z < 0$. However, functions in software, notably R (R Core Team, 2013), return 0 for negative arguments to the Poisson probability mass function.

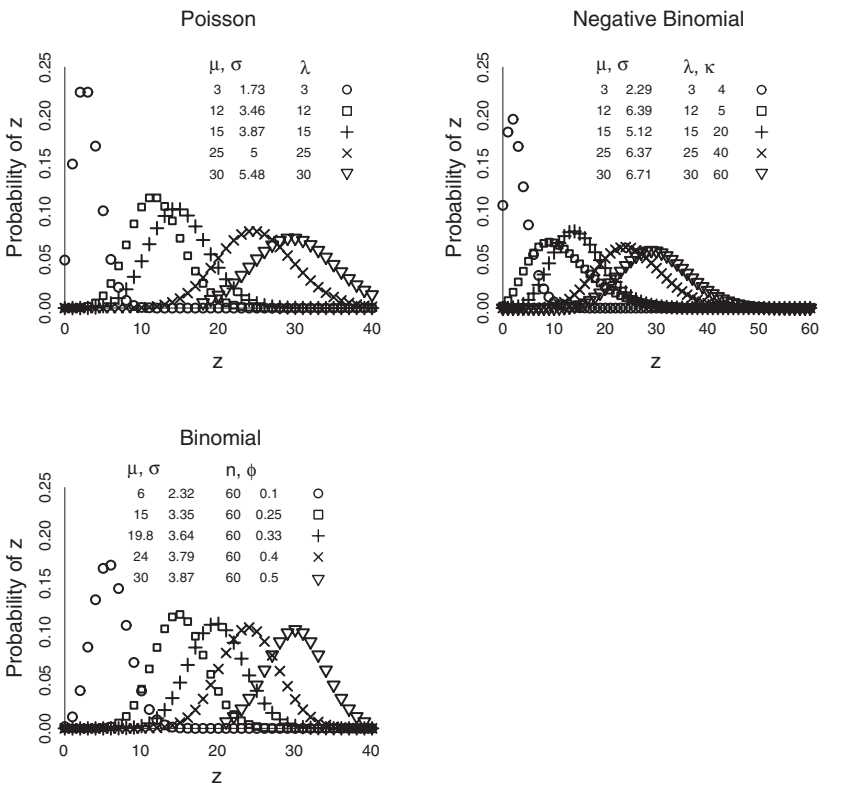


Figure 3.4.5. Example probability mass functions for discrete random variables with means (μ), standard deviations (σ), and parameters (λ, κ, n, ϕ).

The Poisson distribution applies to random variables for which the average number of events is the same as the variance in the number of events, that is, the first and second central moments are equal.

The Poisson probability mass function is

$$[z|\lambda] = \text{Poisson}(z|\lambda) = \frac{\lambda^z}{z!}e^{-\lambda}. \tag{3.4.18}$$

The function returns the probability of occurrence of z events conditional on the value of λ. You may also see the Poisson distribution written as

$$[z|\lambda] = \text{Poisson}(z|\lambda) = \frac{(\gamma\lambda)^z}{z!}e^{-\gamma\lambda}, \tag{3.4.19}$$

where γ specifies a specific interval of time, length, area, or volume; and λ is a rate, the number of occurrences per unit of time, length, area, or volume. Thus, the parameter γ sets the scale of the process. The quantity γ is entered into the equation as known (that is, as perfectly observed data)[16] and is called an *offset*. When the Poisson is expressed as equation 3.4.19, the parameter γ has units of time, length, area, or volume, and λ is a rate with units that are the reciprocal of the units of γ. When the Poisson is expressed as equation 3.4.18, then λ is unitless, because it refers to the average number of things. Units are implicit and are defined by the area, time interval, and so forth, over which counts are made.

Negative Binomial. Using the Poisson distribution requires the restrictive assumption that the mean of the distribution equals the variance. Sometimes we will wish to model the intensity of a number of events where the variance in the number of events exceeds the mean. In this case, the negative binomial distribution is a logical choice (fig. 3.4.5). It applies to the same type of count data as the Poisson but contains a second parameter, κ, controlling the dispersion of the distribution:

$$[z|\lambda, \kappa] = \text{negative binomial}(z|\lambda, \kappa) = \frac{\Gamma(z+\kappa)}{\Gamma(\kappa)\, z!}\left(\frac{\kappa}{\kappa+\lambda}\right)^{\kappa}\left(\frac{\lambda}{\kappa+\lambda}\right)^{z}. \tag{3.4.20}$$

The gamma function, $\Gamma(\ )$, may not be familiar to all readers. It is a function that interpolates a smooth curve connecting the points (x, y) given by $y = (x-1)!$ at the positive integer values for x. The mean of the negative binomial distribution is λ, and the variance is $\lambda + \lambda^2/\kappa$.

A second version of the negative binomial, less commonly used in ecology, gives the probability of a number (z) of failures that occur in a sequence of Bernoulli trials[17] before a target number of successes (k) is obtained:

$$[z|k, \phi] = \text{negative binomial}(z|k, \phi) = \frac{\Gamma(z+k)}{\Gamma(k) z!}\phi^{k}(1-\phi)^{z}. \tag{3.4.21}$$

The parameter ϕ is the probability of a success on a single trial. The parameter k is usually referred to as the *size* in this parameterization.

Software sometimes uses equation 3.4.21 in functions for the negative binomial, so you need to be careful using them if your intention is to use

[16] This is why it does not appear in the distribution $[z|\lambda]$. More about this later.

[17] A *Bernoulli trial* is an experiment with two possible outcomes. Coin flipping is the classical example of a Bernoulli trial.

3.4.20. A little algebra allows you to modify the arguments to functions implemented in software. To modify equation 3.4.20 so that the parameter k represents dispersion (as in eq. 3.4.20), substitute $\kappa/(\lambda+\kappa)$ for the function argument ϕ.

Binomial. The binomial distribution portrays counts that can be assigned to one of two possible categories, for example, alive or dead, present or absent, male or female, exotic or native, diseased or healthy (fig. 3.4.5). The distribution describes the probability of the number of "successes" out of n trials conditional on the parameter ϕ, the probability of a success on any single trial. "Successes" arbitrarily refers to one of the two categories such that successes + failures = number of trials. Trials most often represent observations of a specific number of individual organisms, locations, or sampling plots in the two categories. The probability mass function for a binomial random variable[18] is

$$[z|n,\phi] = \text{binomial}(z|n,\phi) = \binom{n}{z}\phi^z(1-\phi)^{n-z}. \qquad (3.4.22)$$

The random variable z and parameter n must be integers; the parameter ϕ is a real number, $0 \leq \phi \leq 1$. The function returns the probability of z successes conditional on n and ϕ. The mean of the binomial distribution is $n\phi$, and the variance is $n\phi(1-\phi)$.

Bernoulli. The Bernoulli is a special case of the binomial where the number of trials = 1. Its probability mass function is

$$[z|\phi] = \text{Bernoulli}(z|\phi) = \phi^z(1-\phi)^{1-z} \quad \text{for } z \in \{0,1\}. \qquad (3.4.23)$$

The parameter ϕ is the probability of success ($z = 1$) on a single trial. The function computes $z = 1$ with probability ϕ and $z = 0$ with probability $1-\phi$. The Bernoulli distribution has particularly important application in modeling presence or absence data and in occupancy modeling, where it is used to estimate the probability that a particular state is detected, allowing us to separate cases where a state of interest is "unoccupied" from the state "occupied but not observed." Examples of these models will be treated in sections 6.2.3 and 12.3.

[18] The term $\binom{n}{z} = n!/z!(n-z)!$.

Multinomial. The multinomial distribution is used to model random variables that fall into more than two categories on η trials:

$$[\mathbf{z}|\boldsymbol{\phi}, \eta] = \text{multinomial}(\mathbf{z}|\boldsymbol{\phi}, \eta) = \eta! \prod_{i=1}^{k} \frac{\phi_i^{z_i}}{z_i!}. \tag{3.4.24}$$

The symbol $\prod_{i=1}^{k}$ might be unfamiliar to some readers. It means, take a product over all the k quantities indexed by i. So, it is the multiplicative equivalent of the more familiar summation $\sum_{i=1}^{k}$.

The multinomial is the first multivariate distribution we have encountered. It is multivariate because it returns the probability of a *vector* of random variables ($\mathbf{z}$), conditional on a vector of parameters ($\boldsymbol{\phi}$) specifying the probability of occurrence in each category. The parameter η is the total number of occurrences, $\eta = \sum_{i=1}^{k} z_i$, where k is the number of categories. The mean of each category is $\eta\phi_i$, and the variance is $\eta\phi_i\,(1 - \phi_i)$.

The multinomial has many applications in ecological modeling because it is used to represent the number of individuals in a set of mutually exclusive states—a common output of ecological models. It is applied in capture-recapture analysis; in modeling movement of individuals among discrete locations; and in discrete-time matrix modeling of populations, communities, and ecosystems.

3.4.3.2 Probability Density Functions for Continuous Random Variables

Normal. The univariate normal distribution (also known as the *Gaussian distribution*) applies to continuous random variables that can take on values across the entire number line, $-\infty < z < \infty$ (fig. 3.4.6). It is widely used in statistics because it has properties allowing many results to be derived analytically, for example, least-squares estimates of parameters.[19] In addition, the normal distribution is widely used because of the central limit theorem, which states that the sum of a large number of samples from any distribution will be normally distributed. The variance of the normal distribution does not depend in any way on the mean of the distribution. The probability density function for a normally distributed

[19]It is easy to forget that the discipline of statistics developed well before the advent of fast computers. During much of that period of development, analytically tractable distributions were the sole route to useful results in research.

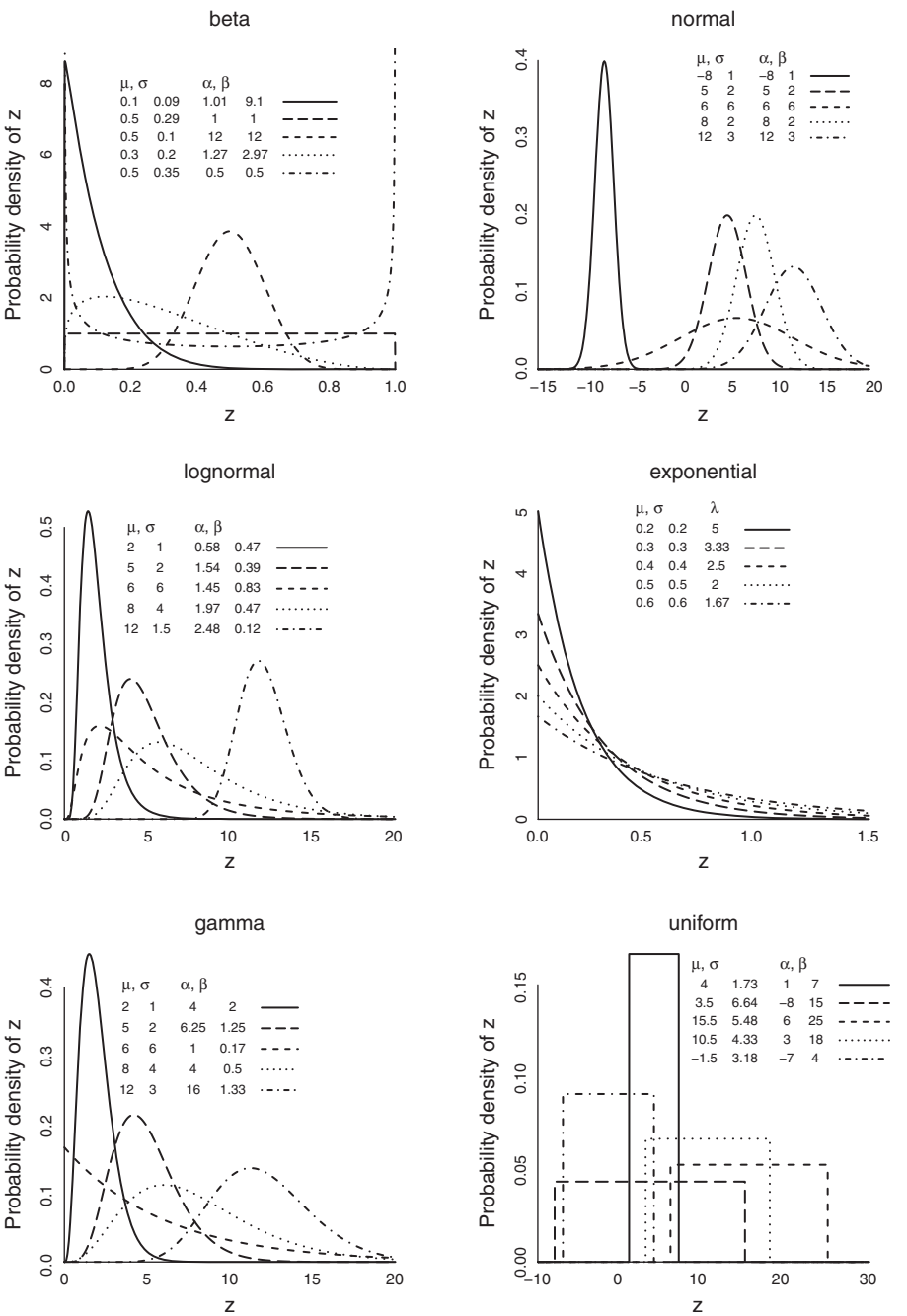


Figure 3.4.6. Example probability density functions for continuous random variables with means (μ), standardard deviations (σ), and parameters (α, β, λ).

random variable is

$$\left[z|\mu,\sigma^2\right] = \text{normal}(z|\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(z-\mu)^2}{2\sigma^2}}. \tag{3.4.25}$$

It is important to note that despite its widespread use in traditional statistics, the normal distribution is often not an appropriate choice for modeling in ecology. The reason is that the preponderance of ecological quantities are nonnegative, which means that the support for these random variables should be zero for negative values (i.e., $\Pr(z \leq 0) = 0$), which of course is not true for the normal. Moreover, the normal is always symmetric and cannot represent the skewness that often arises in distributions of ecological quantities. Finally, the variance of a random variable often increases with the mean, but the normal has a constant variance. These problems are overcome by the lognormal and gamma distributions.

Lognormal. If a random variable z is normally distributed, then e^z is lognormally distributed; if a random variable z is lognormally distributed, then $\log(z)$ is normally distributed. The lognormal distribution (fig. 3.4.6) has properties analogous to those of the central limit theorem for the normal. If we take the product of large numbers of random variables, then the outcome is lognormally distributed regardless of the underlying distributions of the individual random variables. The lognormal distribution is widely used to represent growth of individuals or populations. If we define a growth process as $z_t = \alpha z_{t-1}$, then it follows that the random variable, z_t at time t represents the product of a constant α and the previous state of the random variable, z_{t-1}. The lognormal offers a good choice for modeling the process because it can be represented as a product of states and parameters. The probability density function for the lognormal distribution is

$$[z|\alpha,\beta] = \text{lognormal}(z|\alpha,\beta) = \frac{1}{z\sqrt{2\pi\beta^2}}\, e^{-\frac{(\log(z)-\alpha)^2}{2\beta^2}}, \tag{3.4.26}$$

where α is the mean of $\log(z)$, and β is the standard deviation of $\log(z)$. It would be tempting to think that the mean of the distribution is e^α, which instead gives the median. The mean depends on both parameters, such that

$$\mathrm{E}(z) = \mu = e^{\alpha+\beta^2/2}. \tag{3.4.27}$$

Similarly, the variance of the lognormal distribution also depends on α and β:

$$\mathrm{E}\left((z-\mu)^2\right) = \sigma^2 = (e^{\beta^2} - 1)e^{2\alpha+\beta^2}. \tag{3.4.28}$$

The variance of the lognormal increases in proportion to the square of the mean for any given set of parameters α and β. Ecologists often observe continuous, nonnegative quantities for which the variance increases with the mean, which suggests that the lognormal or gamma can be used to model those quantities.

Gamma. The gamma distribution is broadly useful in ecology for modeling random variables that are nonnegative (fig. 3.4.6). Like the lognormal, the gamma distribution is well suited for representing random variables that are skewed. The gamma distribution was originally derived to model the time required for a specified number of events to occur in a Poisson process, that is, where events occur at average rate λ, and the occurrence of an event has no influence on the occurrence of a subsequent event. The distribution is also used to represent random variability in the mean, λ, of the Poisson distribution and thereby provides the basis for the derivation of the negative binomial.

The probability density function for the gamma distribution can take two forms. We will use

$$[z|\alpha, \beta] = \mathrm{gamma}(z|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} z^{\alpha-1} e^{-\beta z}, \tag{3.4.29}$$

where α is called the *shape*, and β, the *rate.* Both parameters must be positive real numbers. The mean of the gamma distribution is

$$\mathrm{E}(z) = \mu = \frac{\alpha}{\beta}, \tag{3.4.30}$$

and the variance is

$$\mathrm{E}\left((z-\mu)^2\right) = \sigma^2 = \frac{\alpha}{\beta^2}. \tag{3.4.31}$$

An alternative parameterization for the gamma distribution is

$$[z|k, \theta] = \mathrm{gamma}(z|k, \theta) = \frac{1}{\Gamma(k)\theta^k} z^{k-1} e^{-\frac{z}{\theta}}, \tag{3.4.32}$$

where k is called the *shape*, and θ, the *scale*. This form is used to model the waiting time z for k events in a Poisson process where the average waiting time between events is $\theta = \beta^{-1}$. The distribution can also be parameterized with a shape, k, and a mean parameter, $k\beta^{-1}$. The same ideas about "waiting times" can be applied to space by using the gamma distribution to model the distance or area that must be covered before encountering k items that are Poisson distributed over space.

Exponential. The gamma distribution simplifies to the exponential distribution when $\alpha = 1$ or $k = 1$. The exponential distribution (fig. 3.4.6) gives the probability of times or distances between sequential events in a Poisson process,

$$[z|\lambda] = \text{exponential}(z|\lambda) = \lambda e^{-\lambda z}, \tag{3.4.33}$$

where λ is the average number of events per unit of time or space. For example, if prey are captured at an average rate λ, then the number of prey captured follows a Poisson distribution, and the time between captures follows an exponential distribution. The mean of the exponential distribution is λ^{-1}, and the variance is λ^{-2}.

The exponential distribution has direct application to ecological models composed of systems of differential equations in showing the relationship between rates (time^{-1}) and probabilities. Consider the differential equation

$$\frac{dq}{dt} = -kq, \tag{3.4.34}$$

which describes the instantaneous rate of change in a state variable, q. The quantity q can represent anything—number of individuals in a population, grams of nitrogen in the soil, area of landscape in forest. The usual metaphor for q is a "compartment," and equation 3.4.34 describes the rate of movement of particles (individuals, atoms, pixels) out of the compartment. The rate of change per particle is

$$\frac{dq}{dt}\frac{1}{q} = -k. \tag{3.4.35}$$

We can use the exponential distribution to translate the rate k (time^{-1}) into the probability that a particle leaves the compartment during a time interval Δt. We define the event "waiting time of a particle in the compartment" as the random variable z. The cumulative distribution function for

the exponential distribution is

$$\int_{-\infty}^{z} \text{exponential}(u|\lambda)\, du = 1 - e^{-\lambda z}. \tag{3.4.36}$$

If we let $z = \Delta t$ and $\lambda = k$, where k is a continuous time rate constant (with dimensions of time^{-1}, equation 3.4.35), then the probability that a particle has waiting time $z < \Delta t$ is $1 - e^{-k\Delta t}$, which is the probability that it leaves the compartment during Δt. The probability that it remains in the compartment is the complement, $e^{-k\Delta t}$. It follows that the average number of particles that leave the compartment during t to $t + \Delta t$ is $q_t(1 - e^{-k\Delta t})$, and the average that remain is $q_t e^{-k\Delta t}$. For example, if the compartment represents "individuals that are alive," and k is the instantaneous mortality rate, then $e^{-k\Delta t}$ gives the probability that an animal survives from time t to $t + \Delta t$, $1 - e^{-k\Delta t}$ gives the probability that it dies, $q_t\left(1 - e^{-k\Delta t}\right)$ is the average number of deaths, and $q_t e^{-k\Delta t}$ is the average number of survivors. Of course, these population results can also be derived from the solution to the differential equation, $q_{t+\Delta t} = q_t e^{-k\Delta t}$.

Although the approach illustrated here uses constant relative rates (i.e., k), the same equations can be applied to nonlinear rates if Δt is sufficiently small. So, for example, if the rate of conversion of susceptible individuals, S, to infected ones in a population is controlled by the nonlinear rate $\frac{dS}{dt} = -\beta SI$, where I is the number of infected individuals, then per capita rate of infection is $\frac{dS}{dt}\frac{1}{S} = -\beta I$, and the probability that a susceptible becomes infected during a small interval of time Δt is $1 - e^{-\beta I \Delta t}$.

This type of logic forms a fundamental link between traditional, continuous-time state variable models in ecology, discrete-time state variable models, and models that are individual based. However, it is surprisingly absent from texts on individual-based modeling (e.g., Railsback and Grimm, 2012).

Inverse Gamma. The inverse gamma distribution (fig. 3.4.6) models the reciprocal of a gamma-distributed random variable. If $b \sim \text{gamma}(\alpha,\theta)$, and $z = b^{-1}$, then the probability density of z is given by

$$[z|\alpha, \beta] = \frac{\beta^{\alpha}}{\Gamma(\alpha)} z^{-\alpha-1} \exp\left(\frac{-\beta}{z}\right), \tag{3.4.37}$$

where the scale parameter $\beta = \theta^{-1}$. At the risk of getting ahead of ourselves, the inverse gamma is particularly useful in modeling the variance of the normal and lognormal distributions. We will study this application in sections 5.3 and 7.3.2.

The mean of the inverse gamma distribution is

$$\mathrm{E}(z) = \mu = \frac{\beta}{\alpha - 1} \quad \text{for}\, \alpha > 1, \tag{3.4.38}$$

and the variance is

$$\mathrm{E}(z - \mu) = \sigma^2 = \frac{\beta^2}{(\alpha - 1)^2(\alpha - 2)} \quad \text{for}\, \alpha > 2. \tag{3.4.39}$$

Beta. The beta distribution (fig. 3.4.6) models the distribution of random variables that can take on values between 0 and 1.[20] It is often used to model uncertainty in probabilities and proportions, making it an essential part of the toolbox of distributions needed by the ecological modeler. The probability density of a beta-distributed random variable, z, conditional on parameters α and β is

$$[z|\alpha, \beta] = \mathrm{beta}(z|\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} z^{\alpha - 1}(1 - z)^{\beta - 1}. \tag{3.4.40}$$

The mean of the distribution is

$$\mathrm{E}(z) = \mu = \frac{\alpha}{\alpha + \beta}, \tag{3.4.41}$$

and the variance is

$$\mathrm{E}\left((z - \mu)^2\right) = \sigma^2 = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}. \tag{3.4.42}$$

Beta distributions are widely used in analysis of survival, detection probability, and decomposition.

Uniform. The uniform distribution (fig. 3.4.6) returns a single probability density for all values of the random variable for which the probability density is greater than 0:

$$[z|\alpha, \beta] = \begin{cases} \frac{1}{\beta - \alpha} & \text{for}\, \alpha \le z \le \beta, \\ 0 & \text{for}\, z < \alpha \text{ or } z > \beta \end{cases}. \tag{3.4.43}$$

[20] Be careful not to confuse this with the Bernoulli, which models random variables that are 0 *or* 1.

The mean of the uniform distribution is $\mu = (\beta + \alpha)/2$, and the variance is $\sigma^2 = (\beta - \alpha)^2/12$. The uniform is especially useful for defining vague prior distributions in Bayesian analysis.

Multivariate Normal. The multivariate normal distribution is often used to represent a set of correlated real-valued random variables, each of which centers around a mean. It is particularly valuable for representing the probability distribution of data that are correlated over time or space and has important applications in regression. The multivariate normal is a generalization of the normal distribution for a single random variable to the distribution of a random vector of variables. A random vector is k-variate normally distributed if each linear combination of its elements has a univariate normal distribution. The probability density is

$$[\mathbf{z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}] = \text{multivariate normal}(\mathbf{z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{k}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{z}-\boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{z}-\boldsymbol{\mu})}, \tag{3.4.44}$$

where $\mathbf{z}$ is a vector of random variables, $\boldsymbol{\mu}$ is a vector of means, and $\boldsymbol{\Sigma}$ is a variance-covariance matrix (box 3.4). The term $|\boldsymbol{\Sigma}|$ indicates the determinant[21] of $\boldsymbol{\Sigma}$.

Box 3.4 Covariance Matrices

We do not use covariance matrices often in this book, but it is important to understand them, because they are needed when modeling a vector of random variables rather than a single one. Each of these random variables has its own variance, but they also may *covary*. Covariance describes how two or more random variables deviate from their mean–if they covary, then they deviate in similar ways. The covariance of random variables x and y is formally defined as

$$\sigma_{xy} = \mathrm{E}\left((x - \mathrm{E}(x))(y - \mathrm{E}(y))\right). \tag{3.4.45}$$

(*continued*)

[21] Don't confuse $\boldsymbol{\Sigma}$ with the summation sign. Instead, it is uppercase boldface version of the scalar σ used to indicate a matrix. Determinants are quantities calculated on square matrices. The use of determinants is beyond the scope of this book, but we wanted to clarify this notation so that it is not confused with absolute value.

(Box 3.4 *continued*)

As you might expect, covariance is closely related to correlation (ρ_{xy}),

$$\rho_{xy} = \frac{\mathrm{E}\left((x - \mathrm{E}(x))\,(y - \mathrm{E}(y))\right)}{\sigma_x \sigma_y}, \tag{3.4.46}$$

where σ_x and σ_y are the standard deviations of the two random variables. When random variables covary, a scatter plot of their values tends to fall along a line, as in figure 3.4.4 D. When they do not covary, the values form a diffuse cloud.

A covariance matrix is a square matrix with the number of rows and columns equal to the number of elements in the vector being modeled. The diagonal elements are the variances of each random variable in the vector, and the off-diagonal elements i, j are the covariance of random variable i with random variable j. If we assume that all random variables in the vector have the same variance, σ^2, and do not covary, then $\mathbf{\Sigma} = \sigma^2\mathbf{I}$, where $\mathbf{I}$ is the identity matrix, consisting of 1s on the diagonal and 0s elsewhere. In this case, $\mathbf{\Sigma}$ has σ^2 on the diagonal, and 0s elsewhere.

Dirichlet. The beta distribution models a random variable that can take on a value between 0 and 1. The Dirichlet is the multivariate analog of the beta—it models random variables that are vectors of proportions summing to 1. Thus, it can be especially useful in modeling composition of populations, communities, and landscapes as well as the time or energy budgets of individuals. The probability density of a random vector $\mathbf{z}$ conditional on a vector of k parameters $\boldsymbol{\alpha}$ is

$$[\mathbf{z}|\boldsymbol{\alpha}] = \mathrm{Dirichlet}(\mathbf{z}|\boldsymbol{\alpha}) = \frac{\Gamma\left(\sum_{i=1}^{k} \alpha_i\right)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} z_i^{\alpha_i - 1}, \tag{3.4.47}$$

where k is the number of elements in the vector. The mean of the ith element of the random vector $\mathbf{z}$ is $\mathrm{E}(z_i) = \mu_i = \frac{\alpha_i}{\alpha_0}$ with variance $\mathrm{E}((z_i - \mu_i)^2) = \sigma_i^2 = \frac{\alpha_i(\alpha_0 - \alpha_i)}{\alpha_0^2(\alpha_0 + 1)}$, where $\alpha_0 = \sum_{i=1}^{k} \alpha_i$.

3.4.4 Moment Matching

The concept of parameters that differ from moments is unfamiliar to ecologists trained in statistics classes emphasizing methods based on the normal distribution, which is to say, most ecologists. The two parameters of the normal distribution are its first and second central moments, the mean and the variance, motivating students and colleagues to ask us, "Why are parameters necessary? Why not simply use the moments as parameters for distributions?"

The answer is important, if not obvious. In the normal and multivariate normal, the variance does not change for different values of the mean. However, for other distributions we will use—the binomial, multinomial, negative binomial, beta, gamma, lognormal, exponential and Dirichlet—the variance is a function of the mean. Moreover, the parameters of these distributions are functions of both the mean *and* the variance, which allows the relationship between the mean and variance to change as the parameters change. The only time that the moments can be used as parameters is when the mean and the variance are the same, as in the Poisson, or are not related to each other, as is the case for the normal and multivariate normal. This creates a problem for the ecologist who seeks to use the toolbox of distributions that we have described so far, a problem that can easily be seen in the following example.

Assume you want to model the influence of growing season rainfall (x_i) on the mean aboveground standing crop biomass in a grassland at the end of the growing season (μ_i in kg/ha). You might be inclined to reach for the simple linear model $\mu_i = \gamma_0 + \gamma_1 x_i$ to represent this relationship. However, there are structural problems with a linear model, because it predicts values that can be negative, which makes no sense for biomass. Moreover, it predicts that growth increases infinitely with increasing rainfall, which clearly is not correct on biological grounds. So, using your knowledge of deterministic models (chapter 2), you choose

$$\mu_i = \frac{\kappa x_i}{\gamma + x_i}, \tag{3.4.48}$$

thereby deftly assuring that the model's estimate is nonnegative for nonnegative values of x_i and asymptotically approaches a maximum, κ.[22]

[22] If you wanted a model with "linear" components, you could also use $(\kappa \exp(\gamma_0 + \gamma_1 x_i))/(1 + \exp(\gamma_0 + \gamma_1 x_i))$, which also has a maximum at κ.

Equation 3.4.48 is purely deterministic. You would like to represent the uncertainty that arises because the model isn't perfect and because we fail to observe net primary production perfectly.[23] Your first thought about modeling the uncertainty might be to use a normal distribution, normal($y_i|\mu_i, \sigma^2$). So, your model predicts the mean (μ_i) of the distribution of observations of growth (y_i), and the uncertainty surrounding that prediction depends on σ^2. This is the traditional framework for regression. It is convenient because the prediction of the model is the first argument to the distribution. You are probably more familiar with the equivalent formulation, $y_i = \gamma_0 + \gamma_1 x_i + \epsilon_i$, $\epsilon_i \sim$ normal($0, \sigma^2$). We avoid this additive arrangement for representing stochasticity because it cannot be applied to distributions that cannot be centered on zero.

However, informed by the section on continuous distributions, you decide that the normal is a poor choice for your model for two reasons. First, the support is wrong. Biomass cannot be negative, so you need a distribution for data that are continuous and strictly positive. Moreover, a plot of the data shows that the spread of the residuals increases with increasing production, casting doubt on the assumption that the variance is constant. As an alternative, you choose the gamma distribution[24] because it is strictly nonnegative and is parameterized such that the variance increases in proportion to μ^2.

This is entirely sensible, but now you have a problem. How do you get the prediction of your model, the mean prediction of biomass at a given level of rainfall (μ_i), into the gamma probability density function if the function doesn't contain an argument for the mean? How do you represent uncertainty by using the variance, σ^2?

The solution to this problem is *moment matching*. You need equations for the parameters in terms of the moments to allow you to use the gamma distribution to represent the uncertainty in your model. Equations for moments as functions of the parameters can be found in any mathematical statistics text. The converse is not true; it is uncommon to see the parameters expressed as functions of the moments. However, obtaining these functions is easy and useful. You simply solve two equations in two unknowns. On illustration of this solution using the gamma distribution with parameters

[23] In this case, these sources of uncertainty will be inseparable. Later, we will develop models that separate them.

[24] The lognormal would be another logical choice. The gamma or the lognormal would yield virtually identical estimates of parameters if you fit the model.

shape = α and rate = β is

$$\mu = \frac{\alpha}{\beta}, \tag{3.4.49}$$

$$\sigma^2 = \frac{\alpha}{\beta^2}, \tag{3.4.50}$$

so,

$$\alpha = \frac{\mu^2}{\sigma^2}, \tag{3.4.51}$$

$$\beta = \frac{\mu}{\sigma^2}. \tag{3.4.52}$$

You are now equipped to use the gamma distribution to represent the uncertainty in your model of net primary production:

$$\mu_i = \frac{\kappa x_i}{\gamma + x_i},$$

$$\alpha_i = \frac{\mu_i^2}{\sigma^2}, \tag{3.4.53}$$

$$\beta_i = \frac{\mu_i}{\sigma^2}, \tag{3.4.54}$$

$$\left[y_i|\mu_i, \sigma^2\right] = \text{gamma}(y_i|\alpha_i, \beta_i). \tag{3.4.55}$$

As a second example, imagine that you want to model the probability of survival of juvenile birds (μ_i) as a function of population density (x_i). Now, you need a model that makes predictions strictly between 0 and 1, so you might sensibly choose $\mu_i = (\exp(\gamma_0+\gamma_1 x_i))/(1+\exp(\gamma_0+\gamma_1 x_i)), 0 \leq \mu_i \leq 1$. But the gamma distribution is no longer appropriate for representing the uncertainty because it applies to random variables that can exceed 1. The normal is even worse because it includes negative values *and* values that exceed 1. A far better choice is the beta, which models continuous random variables with support on the continuous interval 0 to 1. Solving for the parameters in terms of the moments (using equations 3.4.41 and 3.4.42), you can now make a prediction of survival with your deterministic model

and properly represent uncertainty using the beta distribution:

$$\mu_i = \frac{\exp(\gamma_0 + \gamma_1 x_i)}{1 + \exp(\gamma_0 + \gamma_1 x_i)}, \tag{3.4.56}$$

$$\alpha_i = \frac{\mu_i^2 - \mu_i^3 - \mu_i \sigma^2}{\sigma^2}, \tag{3.4.57}$$

$$\beta_i = \frac{\mu_i - 2\mu_i^2 + \mu_i^3 - \sigma^2 + \mu_i \sigma^2}{\sigma^2}, \tag{3.4.58}$$

$$\left[y_i | \mu_i, \sigma^2\right] = \text{beta}(y_i | \alpha_i, \beta_i). \tag{3.4.59}$$

Equations 3.4.51, 3.4.52 and 3.4.57, 3.4.58 are examples of moment matching. We use the functional relationship between the parameters and the moments to allow us to match the predictions of a model to the arguments of the distribution that is best suited to the model and the data. It is important to see how moment matching allows us to specify characteristics of distributions for which the variance is a function of the mean. These matching relationships are broadly useful for the ecological modeler because they allow use of all the distributions we have already described to represent the stochasticity regardless of the form of the arguments to those distributions. It is easy enough to derive the moment matching relationships yourself, but we saved you the trouble in appendix tables A.1 and A.2.

Up to now, we have matched both mean and variance to parameters. However, sometimes we need to match only the mean[25]. Using the beta distribution as an example, we have $\mu = \alpha/(\alpha + \beta)$, so $\alpha = \frac{\mu\beta}{1-\mu}$, allowing us to use $[y|\mu, \beta] = \text{beta}(y|\frac{\mu\beta}{1-\mu}, \beta)$.

3.4.5 Mixture Distributions

The distributions described in section 3.4.3 provide tremendous flexibility for representing uncertainty in ecological models. Sometimes, however, a single distribution fails to adequately portray the behavior of random variables in a way that is faithful to the process that gives rise to them. In this case, the ecological modeler may need to use mixtures of distributions.

[25] As you will learn soon (chapter 8), you can obtain the variance of the distribution of the mean from the output of a Markov chain Monte Carlo algorithm. If you don't need to know the variance of the distribution of the observations, then you don't need to moment match the observation variance.

The general form of a finite mixture distribution for discrete random variables is as follows. Given a finite set of probability distributions for the random variable z, $[z]_1, \ldots, [z]_n$, and weights $w_1, \ldots, w_n$, $w_i \geq 0$, $\sum_{i=1}^{n} w_i = 1$, the finite mixture distribution of z is

$$[z] = \sum_i w_i [z]_i. \tag{3.4.60}$$

Similarly, the general form of an infinite mixture distribution for the variable z depends on a probability density function $[z|\delta]$ with parameter δ. That is, for each value of z in some set δ, $[z|\delta]$ is a probability density function with respect to z. Given a probability density function $[\delta]$ (meaning that $[\delta]$ is nonnegative and integrates to 1), the function

$$[z] = \int [z|\delta]\,[\delta]\,d\delta \tag{3.4.61}$$

is an infinite mixture distribution for z.

We give two examples here. Suppose you study a species and you want to represent the distribution of the random variable, body mass of an individual, z. The sex of individuals is not easily determined in the field, but there are differences in body mass between sexes. How would you model the distribution of these observations?

Because body mass is strictly positive, a gamma distribution is a logical choice, but a single gamma probability density function is not up to the task of representing the two sources of variation in body mass arising from males and females. Instead, you might use

$$\phi \cdot \text{gamma}(z|\alpha_m, \beta_m) + (1 - \phi) \cdot \text{gamma}(z|\alpha_f, \beta_f),$$
$$\phi \sim \text{beta}(\eta, \rho)$$

where ϕ is the probability that a draw from the population is male. This approach provides a weighted mixture of the male and female body mass distributions where the weighting is controlled by the probability that an individual is a male.

Ecologists often observe random variables that take on zero values more frequently than would be predicted by a single, unmixed distribution, for example a binomial, Poisson, negative binomial, or multinomial. An excessive number of zero-values for a discrete random variable is called *zero inflation*. A second, particularly useful example of a mixture distribution allows the ecological modeler to deal with this overdispersion.

To illustrate, imagine that we sampled many plots along a coastline, counting the number of species of mussels within each plot. In essence there are two sources of zeros. Some zeros arise because the plot was placed in areas that are not mussel habitat, while other zeros occur in plots placed in mussel habitat but that contain no mussels as a result of sampling variation. The Poisson distribution offers a logical choice for modeling the distribution of counts in mussel habitat,[26] but it cannot represent the variation among plots, because it can account only for sampling variation. It cannot portray the zeros that arise because plots were placed in areas where mussels never live.

We can represent these two sources of uncertainty by mixing a Poisson distribution with a Bernoulli distribution. Let z be a random variable representing the number of mussel species in a square meter plot, and w a random variable describing mussel habitat; $w = 1$ if a plot is located in mussel habitat, and $w = 0$ if it is located outside mussel habitat. The distribution of number of mussels in a plot is given by

$$z \sim \begin{cases} 0 & w = 0 \\ \text{Poisson}(\lambda) & w = 1 \end{cases}, \tag{3.4.62}$$

which you will also see written as

$$z \sim \text{Poisson}(z|\lambda w) \cdot \text{Bernoulli}(w|\phi), \tag{3.4.63}$$

$$\phi \sim \text{beta}(\eta, \rho), \tag{3.4.64}$$

where λ is the mean number of mussels per square meter in mussel habitat; ϕ is the probability that a plot contains mussel habitat, and η and ρ are parameters that control the distribution of ϕ. Mixture distributions like this one will be seen again in one of our examples of hierarchical models in sections 6.2.3 and 12.3.

[26]The negative binomial would also work and might be better than the Poisson if the sampling variation is overdispersed.