Chapter Four

Probability distributions

SUMMARY

This chapter continues to review the math you need to fit models to data, moving forward from functions and curves to probability distributions. The first part discusses ecological variability in general terms, then reviews basic probability theory and some important applications, including Bayes' Rule and its application in statistics. The second part reviews how to analyze and understand probability distributions. The third part provides a bestiary of probability distributions, finishing with a short digression on some ways to extend these basic distributions.

4.1 INTRODUCTION: WHY DOES VARIABILITY MATTER?

For many ecologists and statisticians, noise is just a nuisance — it gets in the way of drawing conclusions from the data. The traditional statistical approach to noise in data was to assume that all variation in the data was normally distributed, or transform the data until it was, and then use classical methods based on the normal distribution to draw conclusions. Some scientists turned to nonparametric statistics, which assume only that the shape of the data distribution is the same in all categories and provide tests of differences in the means or "location parameters" among categories. Unfortunately, classical nonparametric approaches make it much harder to draw quantitative conclusions from data (rather than simply rejecting or failing to reject null hypotheses about differences between groups).

In the 1980s, as they acquired better computing tools, ecologists began to use more sophisticated models of variability such as generalized linear models (see Chapter 9). Chapter 3 illustrated a wide range of deterministic functions that correspond to eterministic models of the underlying ecological processes. This chapter will illustrate a wide range of models for the stochastic part of the dynamics. In these models, variability isn't just a nuisance, but actually tells us something about ecological processes. For example, census counts that follow a negative binomial distribution (p. 165) tell us there is some form of environmental variation or aggregative response among individuals that we haven't taken into account (Shaw and Dobson, 1995).

Remember from Chapter 1 that what we treat as "signal" (deterministic) and what we treat as "noise" (stochastic) depends on the question. The same ecological variability, such as spatial variation in light, might be treated as random variation by a forester interested in the net biomass increment of a forest stand and as a deterministic driving factor by an ecophysiologist interested in the photosynthetic response of individual plants.

Noise affects ecological data in two different ways — as *measurement error* and as *process noise* (this will become important in Chapter 11 when we deal with dynamical models). Measurement error is the variability or "noise" in our measurements, which makes it hard to estimate parameters and make inferences about ecological systems. Measurement error leads to large confidence intervals and low statistical power. Even if we can eliminate measurement error, process noise or process error (often so-called even though it isn't technically an "error", but a real part of the system) still exists. Variability affects any ecological system. For example, we can observe thousands of individuals to determine the average mortality rate with great accuracy. The fate of a group of a few individuals, however, depends both on the variability in mortality rates of individuals and on the *demographic stochasticity* that determines whether a particular individual lives or dies ("loses the coin toss"). Even though we know the average mortality rate perfectly, our predictions are still uncertain. *Environmental stochasticity* — spatial and temporal variability in (e.g.) mortality rate caused by variation in the environment rather than by the inherent randomness of individual fates — also affects the dynamics. Finally, even if we can minimize measurement error by careful measurement and minimize process noise by studying a large population in a constant environment (i.e. low levels of demographic and environmental stochasticity), ecological systems can still amplify variability in surprising ways (Bjørnstad and Grenfell, 2001). For example, a tiny bit of demographic stochasticity at the beginning of an epidemic can trigger huge variation in epidemic dynamics (Rand and Wilson, 1991). Variability also feeds back to change the mean behavior of ecological systems. For example, in the damselfish system described in Chapter 2 the number of recruits in any given cohort is the number of settlers surviving density-dependent mortality, but the average number of recruits is *lower* than expected from an average-sized cohort of settlers because large cohorts suffer disproportionately high mortality and contribute relatively little to the average. This widespread phenomenon follows from *Jensen's inequality* (Ruel and Ayres, 1999; Inouye, 2005).

4.2 BASIC PROBABILITY THEORY

In order to understand stochastic terms in ecological models, you'll have to (re)learn some basic probability theory. To define a probability, we first have to identify the *sample space*, the set of all the possible outcomes that could occur. Then the probability of an event A is the frequency with which that event occurs. A few probability rules are all you need to know:

i) If two events are *mutually exclusive* (e.g., "individual is male" and "individual is female") then the probability that either occurs (the prob-

ability of A *or* B, or $\text{Prob}(A \cup B)$) is the sum of their individual probabilities: e.g. Prob(male or female) = Prob(male) + Prob(female).

We use this rule, for example, in finding the probability that an outcome is within a certain numeric range by adding up the probabilities of all the different (mutually exclusive) values in the range: for a discrete variable, for example, $P(3 \leq X \leq 5) = P(X = 3) + P(X = 4) + P(X = 5)$.

ii) If two events A and B are not mutually exclusive — the *joint probability* that they occur together, $\text{Prob}(A \cap B)$, is greater than zero — then we have to correct the rule for combining probabilities to account for double-counting:

$$\text{Prob}(A \cup B) = \text{Prob}(A) + \text{Prob}(B) - \text{Prob}(A \cap B).$$

For example if we are tabulating the color and sex of animals, Prob(blue or male) = Prob(blue) + Prob(male) − Prob(blue male))

iii) The probabilities of all possible outcomes of an observation or experiment add to 1.0. (Prob(male) + Prob(female) = 1.0.)

We will need this rule to understand the form of probability distributions, which often contain a *normalization constant* to make sure that the sum of the probabilities of all possible outcomes is 1.

iv) The *conditional probability* of A *given* B, $\text{Prob}(A|B)$, is the probability that A happens if we know or assume B happens. The conditional probability equals

$$\text{Prob}(A|B) = \text{Prob}(A \cap B)/\text{Prob}(B). \tag{4.2.1}$$

For example:

$$\text{Prob(individual is blue|individual is male)} = \frac{\text{Prob(individual is a blue male)}}{\text{Prob(individual is male)}}. \tag{4.2.2}$$

By contrast, we may also refer to the probability of A when we make no assumptions about B as the *unconditional* probability of A. $\text{Prob}(A) = \text{Prob}(A|B) + \text{Prob}(A|\text{not } B)$.

Conditional probability is central to understanding *Bayes' Rule* (p. 145).

v) If the conditional probability of A given B, $\text{Prob}(A|B)$, equals the unconditional probability of A, then A is *independent* of B. Knowing about B provides no information about the probability of A. Independence implies that

$$\text{Prob}(A \cap B) = \text{Prob}(A)\text{Prob}(B), \tag{4.2.3}$$

which follows from multiplying both sides of (4.2.1) by $\text{Prob}(B)$. The probabilities of combinations of independent events are multiplicative.

Multiplying probabilities of independent events, or adding independent log-probabilities ($\log(\text{Prob}(A \cap B)) = \log(\text{Prob}(A)) + \log(\text{Prob}(B))$ if A and B are independent), is how we find the combined probability of a series of observations.

We can immediately use these rules to think about the distribution of seeds taken in the seed removal experiment (Chapter 2). The most obvious pattern in the data is that there are many zeros, probably corresponding to times when no predators visited the station. The sample space for seed disappearance — is the number of seeds taken, from 0 to N (the number available). Suppose that when a predator *did* visit the station, with probability v, it had an equal probability of taking any of the possible number of seeds (a uniform distribution from 0 to N). Since the probabilities must add to 1, this probability ($\text{Prob}(x \text{ taken}|\text{predator visits})$) is $1/(N+1)$ (0 to N represents $N+1$ different possible events). What is the unconditional probability of x seeds being taken?

If $x > 0$, then there is only one possible type of event — the predator visited and took x seeds — with overall probability $v/(N+1)$ (Figure 4.1, left).

If $x = 0$, then there are two mutually exclusive possibilities. Either the predator didn't visit (probability $1-v$), or it visited (probability v) and took zero seeds (probability $1/(N+1)$), so the overall probability is

$$\underbrace{(1-v)}_{\text{didn't visit}} + \left(\underbrace{v}_{\text{visited}} \times \underbrace{\frac{1}{N+1}}_{\text{took zero seeds}} \right) = 1 - v + \frac{v}{N+1}. \qquad (4.2.4)$$

Now make things a little more complicated and suppose that when a predator visits, it decides independently whether or not to take each seed. If the seeds of a given species are all identical, so that each seed is taken with the same probability p, then this process results in a binomial distribution. Using the rules above, the probability of x seeds being taken when each has probability p is p^x. It's also true that $N - x$ seeds are *not* taken, with probability $(1-p)^{N-x}$. Thus the probability is proportional to $p^x \cdot (1-p)^{N-x}$. To get the probabilities of all possible outcomes to add to 1, though, we have to multiply by a normalization constant $N!/(x!(N-x)!)$*, or $\binom{N}{x}$. (It's too bad we can't just ignore these ugly normalization factors, which are always the least intuitive parts of probability formulas, but we really need them in order to get the right answers. Unless you are doing advanced calculations, however, you can usually just take the formulas for

*$N!$ means $N \cdot (N-1) \cdot \ldots \cdot 2 \cdot 1$, and is referred to as "N factorial".

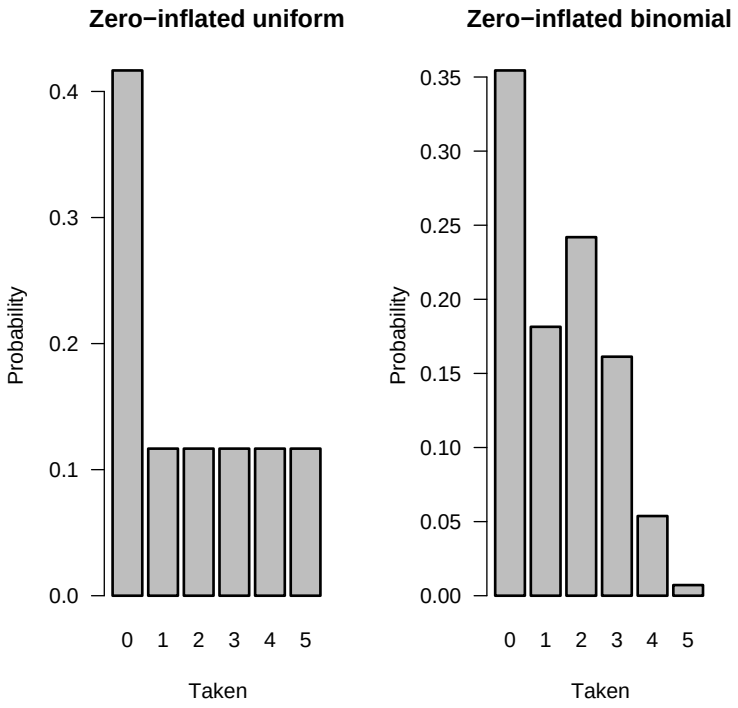

Figure 4.1 Zero-inflated distributions. Left, zero-inflated uniform: right, zero-inflated binomial. Number of seeds $N = 5$, probability of predator visit $v = 0.7$, binomial probability of individual seed predation $p = 0.4$.

the normalization constants for granted, without trying to puzzle out their meaning.)

Now adding the "predator may or may not visit" layer to this formula, we have a probability

$$\underbrace{(1-v)}_{\text{didn't visit}} + \left(\underbrace{v}_{\text{visited}} \cdot \underbrace{\text{Binom}(0,p,N)}_{\text{took zero seeds}} \right) = (1-v) + v(1-p)^N \qquad (4.2.5)$$

if $x = 0$ ($\binom{N}{0} = 1$, so the normalization constant disappears from the second term), or

$$\underbrace{v}_{\text{visited}} \cdot \underbrace{\text{Binom}(x,p,N)}_{\text{took} > 0 \text{ seeds}} = v\binom{N}{x} p^x (1-p)^{N-x} \qquad (4.2.6)$$

if $x > 0$ (Figure 4.1, right).

This distribution is called the *zero-inflated binomial* (Inouye, 1999; Tyre et al., 2003). With only a few simple probability rules, we have derived a potentially useful distribution that might describe the pattern of seed predation better than any of the standard distributions we'll see later in this chapter.

4.3 BAYES' RULE

With the simple probability rules defined above we can also derive, and understand, *Bayes' Rule*. Most of the time we will use Bayes' Rule to go from the likelihood Prob($D|H$), the probability of observing a particular set of data D given that a hypothesis H is true (p. 16), to the information we really want, Prob($H|D$) — the probability of our hypothesis H in light of our data D. Bayes' Rule is just a recipe for turning around a conditional probability:

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)}. \tag{4.3.1}$$

Bayes' Rule is general — H and D can be any events, not just hypothesis and data — but it's easier to understand Bayes' Rule when we have something concrete to tie it to. Deriving Bayes' Rule is almost as easy as remembering it. Rule #4 on p. 142 applied to $P(H|D)$ implies

$$P(D \cap H) = P(H|D)P(D), \tag{4.3.2}$$

while applying it to $P(D|H)$ tells us

$$P(H \cap D) = P(D|H)P(H). \tag{4.3.3}$$

But $P(H \cap D) = P(D \cap H)$ so

$$P(H|D)P(D) = P(D|H)P(H) \tag{4.3.4}$$

and therefore

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)}. \tag{4.3.5}$$

Equation (4.3.5) says that the probability of the hypothesis given (in light of) the data is equal to the probability of the data given the hypothesis (the *likelihood* associated with H), times the probability of the hypothesis, divided by the probability of the data. There are two problems here: we don't know the probability of the hypothesis, $P(H)$ (isn't that what we're trying to figure out in the first place?), and we don't know the unconditional probability of the data, $P(D)$.

Let's think about the second problem first—our ignorance of $P(D)$. We can calculate an unconditional probability for the data if we have a set of *exhaustive*, *mutually exclusive* hypotheses: in other words, we assume that

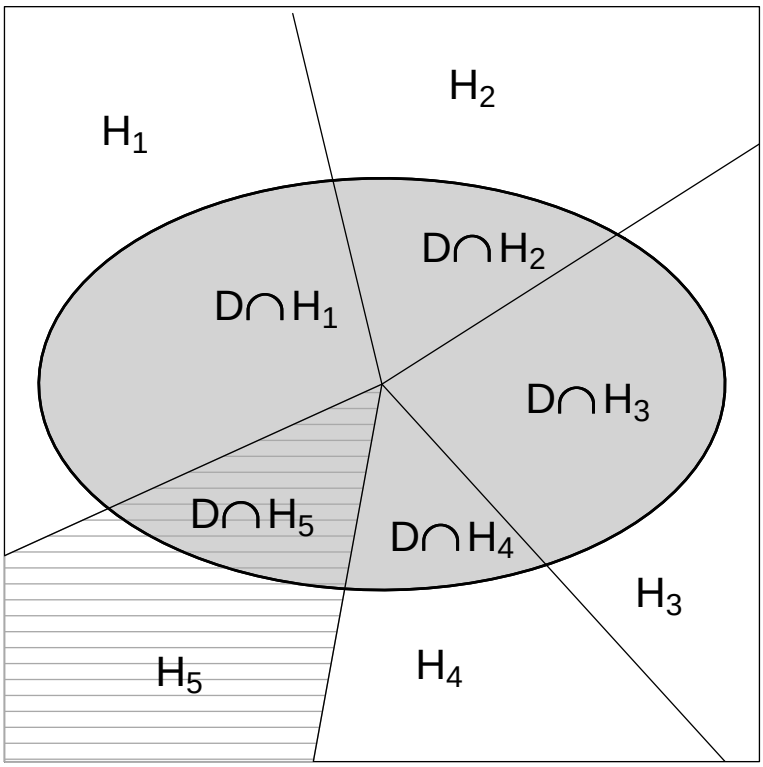

Figure 4.2 Decomposition of the unconditional probability of the observed data (D) into the sum of the probabilities of the intersection of the data with each possible hypothesis ($\sum_{j=1}^{N} D \cap H_j$). The entire gray ellipse in the middle represents D. Each wedge (e.g. the hashed area H_5) represents an alternative hypothesis. The ellipse is divided into "pizza slices" (e.g. $D \cap H_5$, hashed and colored area). The area of each slice corresponds to $D \cap H_j$, the joint probability of the data D (ellipse) *and* the particular hypothesis H_j (wedge).

one, and only one, of our hypotheses is true. Figure 4.2 shows a geometric interpretation of Bayes' Rule. The gray ellipse represents D, the set of all possibilities that could lead to the observed data.

If one of the hypotheses must be true, then the unconditional probability of observing the data is the sum of the probabilities of observing the data under any of the possible hypotheses, For N different hypotheses H_1 to H_N,

$$\begin{aligned} P(D) &= \sum_{j=1}^{N} P(D \cap H_j) \\ &= \sum_{j=1}^{N} P(H_j)P(D|H_j). \end{aligned} \tag{4.3.6}$$

In words, the unconditional probability of the data is the sum of the likelihood of each hypothesis ($P(D|H_j)$) times its unconditional probability ($P(H_j)$). In Figure 4.2, summing the area of overlap of each of the large wedges (the hypotheses H_j) with the gray ellipse ($H_j \cap D$) provides the area of the ellipse (D).

Substituting (4.3.6) into (4.3.5) gives the full form of Bayes' Rule for a particular hypothesis H_i when it is one of a mutually exclusive set of hypotheses $\{H_j\}$. The probability of the truth of H_i in light of the data is

$$P(H_i|D) = \frac{P(D|H_i)P(H_i)}{\sum_j P(H_j)P(D|H_j)} \tag{4.3.7}$$

In Figure 4.2, having observed the data D means we know that reality lies somewhere in the gray ellipse. The probability that hypothesis 5 is true (i.e., that we are somewhere in the hashed area) is equal to the area of the hashed/colored "pizza slice" divided by the area of the ellipse. Bayes' Rule breaks this down further by supposing that we know how to calculate the likelihood of the data for each hypothesis — the ratio of the pizza slice divided by the area of the entire wedge (the area of the pizza slice $[D \cap H_5]$ divided by the hashed wedge $[H_5]$). Then we can recover the area of each slice by multiplying the likelihood by the prior (the area of the wedge) and calculate both $P(D)$ and $P(H_5|D)$.

Dealing with the second problem, our ignorance of the unconditional or *prior* probability of the hypothesis $P(H_i)$, is more difficult. In the next section we will simply assume that we have other information about this probability, and we'll revisit the problem shortly in the context of Bayesian statistics. But first, just to practice with Bayes' Rule, we'll explore two simpler examples that use Bayes' Rule to manipulate conditional probabilities.

4.3.1 False positives in medical testing

Suppose the unconditional probability of a random person sampled from the population being infected (I) with some deadly but rare disease is one in a million: $P(I) = 10^{-6}$. There is a test for this disease that never gives a false negative result: if you have the disease, you will definitely test positive ($P(+|I) = 1$). However, the test does occasionally give a false positive result. One person in 100 who doesn't have the disease (is uninfected, U) will test positive anyway ($P(+|U) = 10^{-2}$). This sounds like a pretty good test. Let's compute the probability that someone who tests positive is actually infected.

Replace H in Bayes' rule with "is infected" (I) and D with "tests positive" ($+$). Then

$$P(I|+) = \frac{P(+|I)P(I)}{P(+)}. \tag{4.3.8}$$

We know $P(+|I) = 1$ and $P(I) = 10^{-6}$, but we don't know $P(+)$, the unconditional probability of testing positive. Since you are either infected (I) or uninfected (U), so these events are mutually exclusive,

$$P(+) = P(+ \cap I) + P(+ \cap U). \tag{4.3.9}$$

Then

$$P(+) = P(+|I)P(I) + P(+|U)P(U) \tag{4.3.10}$$

because $P(I \cap +) = P(+|I)P(I)$ (eq. 4.2.1). We also know that $P(U) = 1 - P(I)$, so

$$\begin{aligned} P(+) &= P(+|I)P(I) + P(+|U)(1 - P(I)) \\ &= 1 \times 10^{-6} + 10^{-2} \times (1 - 10^{-6}) \\ &= 10^{-6} + 10^{-2} + 10^{-8} \\ &\approx 10^{-2}. \end{aligned} \tag{4.3.11}$$

Since 10^{-6} is ten thousand times smaller than 10^{-2}, and 10^{-8} is even tinier, we can neglect them for now.

Now that we've done the hard work of computing the denominator, we can put it together with the numerator:

$$\begin{aligned} P(I|+) &= \frac{P(+|I)P(I)}{P(+)} \\ &\approx \frac{1 \times 10^{-6}}{10^{-2}} \\ &= 10^{-4} \end{aligned} \tag{4.3.12}$$

Even though false positives are unlikely, the chance that you are infected if

you test positive is still only 1 in 10,000! For a sensitive test (one that produces few false negatives) for a rare disease, the probability that a positive test is detecting a true infection is approximately $P(I)/P(\text{false positive})$, which can be surprisingly small.

This false-positive issue also comes up in forensics cases (DNA testing, etc.). Assuming that a positive test is significant is called the *base rate fallacy*. It's important to think carefully about the sample population and the true probability of being guilty (or at least having been present at the crime scene) conditional on having your DNA match DNA found at the crime scene.

4.3.2 Bayes' Rule and liana infestation

A student of mine used Bayes' Rule as part of a simulation model of liana (vine) dynamics in a tropical forest. He wanted to know the probability that a newly emerging sapling would be in a given "liana class" (L_1=liana-free, L_2–L_3=light to moderate infestation, L_4=heavily infested with lianas). This probability depends on the number of trees nearby that are already infested (N). We have measurements of infestation of saplings from the field, and for each one we know the number of nearby infestations. Thus if we calculate the fraction of individuals in liana class L_i with N nearby infested trees, we get an estimate of $\text{Prob}(N|L_i)$. We also know the overall fractions in each liana class, $\text{Prob}(L_i)$. When we add a new tree to the model, we know the neighborhood infestation N from the model. Thus we can figure out what we want to know, $\text{Prob}(L_i|N)$, by using Bayes' Rule to calculate

$$\text{Prob}(L_i|N) = \frac{\text{Prob}(N|L_i)\text{Prob}(L_i)}{\sum_{j=1}^{4}\text{Prob}(N|L_j)\text{Prob}(L_j)}. \tag{4.3.13}$$

For example, suppose we find that a new tree in the model has 3 infested neighbors. Let's say that the probabilities of each liana class (1 to 4) having 3 infested neighbors are $\text{Prob}(N|L_i) = \{0.05, 0.1, 0.3, 0.6\}$ and that the unconditional probabilities of being in each liana class are $L_i = \{0.5, 0.25, 0.2, 0.05\}$. Then the probability that the new tree is heavily infested (i.e. is in class L_4) is

$$\frac{0.6 \times 0.05}{(0.05 \times 0.5) + (0.1 \times 0.25) + (0.3 \times 0.2) + (0.6 \times 0.05)} = 0.21. \tag{4.3.14}$$

We would expect that a new tree with several infested neighbors has a much higher probability of heavy infestation than the overall (unconditional) probability of 0.05. Bayes' Rule allows us to quantify this guess.

4.3.3 Bayes' Rule in Bayesian statistics

So what does Bayes' Rule have to do with Bayesian statistics?

Bayesians translate likelihood into information about parameter values using Bayes' Rule as given above. The problem is that we have the likelihood $\mathcal{L}$(data|hypothesis), the probability of observing the data given the model (parameters): what we want is Prob(hypothesis|data). After all, we already know what the data are!

4.3.3.1 Priors

In the disease testing and the liana examples, we knew the overall, unconditional probability of disease or liana class in the population. When we're doing Bayesian statistics, however, we interpret $P(H_i)$ instead as the *prior probability* of a hypothesis, our belief about the probability of a particular hypothesis *before* we see the data. Bayes' Rule is the formula for updating the prior in order to compute the *posterior probability* of each hypothesis, our belief about the probability of the hypothesis *after* we see the data. Suppose I have two hypotheses A and B and have observed some data D with likelihoods $\mathcal{L}_A = 0.1$ and $\mathcal{L}_B = 0.2$. In other words, the probability of D occurring if hypothesis A is true ($P(D|A)$) is 10%, while the probability of D occurring if hypothesis B is true ($P(D|B)$) is 20%. If I assign the two hypotheses equal prior probabilities (0.5 each), then Bayes' Rule says the posterior probability of A is

$$P(A|D) = \frac{0.1 \times 0.5}{0.1 \times 0.5 + 0.2 \times 0.5} = \frac{0.1}{0.3} = \frac{1}{3} \tag{4.3.15}$$

and the posterior probability of B is 2/3. However, if I had prior information that said A was twice as probable (Prob(A) = 2/3, Prob(B) = 1/3) then the probability of A given the data would be 0.5 (do the calculation). It is in principle possible to get whatever answer you want, by rigging the prior: if you assign B a prior probability of 0, then no data will *ever* convince you that B is true (in which case you probably shouldn't have done the experiment in the first place). Frequentists claim that this possibility makes Bayesian statistics open to cheating (Dennis, 1996): however, every Bayesian analysis must clearly state the prior probabilities it uses. If you have good reason to believe that the prior probabilities are not equal, from previous studies of the same or similar systems, then arguably you should *use* that information rather than starting as frequentists do from the ground up every time. (The frequentist-Bayesian debate is one of the oldest and most virulent controversies in statistics (Ellison, 1996; Dennis, 1996): I can't

possibly do it justice here.)

However, it is a good idea to try so-called *flat* or *weak* or *uninformative* priors — priors that assume you have little information about which hypothesis is true — as a part of your analysis, even if you do have prior information (Edwards, 1996). You may have noticed in the first example above that when we set the prior probabilities equal, the posterior probabilities were just equal to the likelihoods divided by the sum of the likelihoods. Algebraically if all the $P(H_i)$ are equal to the same constant C,

$$P(H_i|D) = \frac{P(D|H_i)C}{\sum_j P(D|H_j)C} = \frac{\mathcal{L}_i}{\sum_j \mathcal{L}_j} \tag{4.3.16}$$

where $\mathcal{L}_i$ is the likelihood of hypothesis i.

You may think that setting all the priors equal would be an easy way to eliminate the subjective nature of Bayesian statistics and make everybody happy. Two examples, however, will demonstrate that it's not that easy to say what it means to be completely "objective" or ignorant of the right hypothesis.

- *partitioning hypotheses*: suppose we find a nest missing eggs that might have been taken by a raccoon, a squirrel, or a snake (only). The three hypotheses "raccoon" (R), "squirrel" (Q), and "snake" (S) are our mutually exclusive and exhaustive set of hypotheses for the identity of the predator. If we have no other information (for example about the local densities or activity levels of different predators), we might choose equal prior probabilities for all three hypotheses. Since there are three mutually exclusive predators, $\mathrm{Prob}(R) = \mathrm{Prob}(Q) = \mathrm{Prob}(S) = 1/3$. Now a friend comes and asks us whether we really believe that mammalian predators are twice as likely to eat the eggs as reptiles ($\mathrm{Prob}(R) + \mathrm{Prob}(Q) = 2\mathrm{Prob}(S)$) (Figure 4.3). What do we do? We might solve this particular problem by setting the probability for snakes (the only reptiles) to 0.5, the probability for mammals ($\mathrm{Prob}(R \cup Q)$) to 0.5, and the probability for raccoons and squirrels equal ($\mathrm{Prob}(R) = \mathrm{Prob}(Q) = 0.25$), but this simple example suggests that such pitfalls are ubiquitous.

- *changing scales*: a similar problem arises with continuous variables. Suppose we believe that the mass of a particular bird species is between 10 and 100 g, and that no particular value is any more likely than other: the prior distribution is uniform, or flat. That is, the probability that the mass is in some range of width Δm is constant: $\mathrm{Prob}(\mathrm{mass} = m) = 1/90\Delta m$ (so that $\int_{10}^{100} \mathrm{Prob}(m)\, dm = 1$: see p. 156 for more on probability densities).

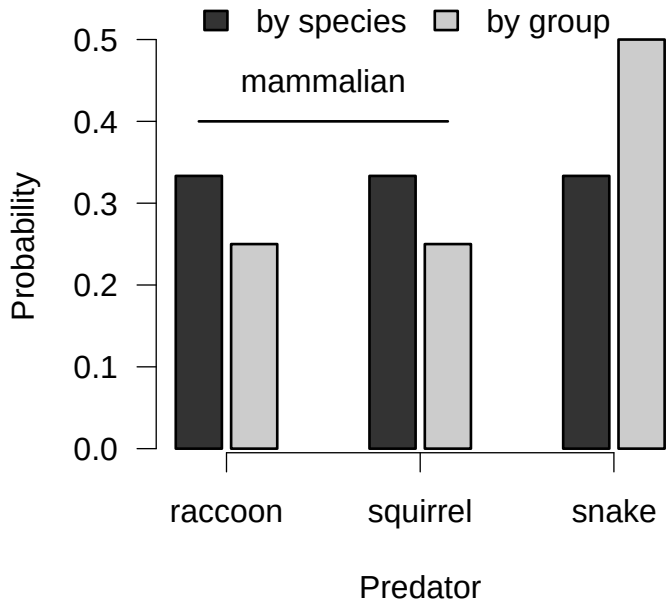

Figure 4.3 The difficulty of defining an uninformative prior for discrete hypotheses. Dark gray bars are priors that assume predation by each species is equally likely; light gray bars divide predation by group first, then by species within group.

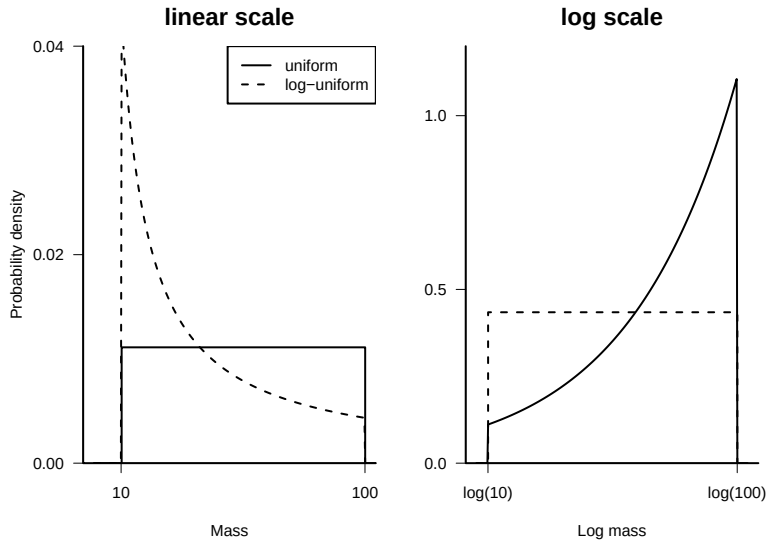

Figure 4.4 The difficulty of defining an uninformative prior on continuous scales. If we assume that the probabilities are uniform on one scale (linear or logarithmic), they must be non-uniform on the other.

But is it sensible to assume that the probability that a species' mass is between 10 and 20 is the same as the probability that it is between 20 and 30, or should it be the same as the probability that it is between 20 and 40 — that is, would it make more sense to think of the mass distribution on a logarithmic scale? If we say that the probability distribution is uniform on a logarithmic scale, then a species is *less* likely to be between 20 and 30 than it is to be between 10 and 20.* Since changing the scale is not really changing anything about the world, just the way we describe it, this change in the prior is another indication that it's harder than we think to say what it means to be ignorant. In any case, many Bayesians think that researchers try too hard to pretend ignorance, and that one really should use what is known about the system. Crome et al. (1996) compare extremely different priors in a conservation context to show that their data really are (or should be) informative to a wide spectrum of stakeholders, regardless of their perspectives.

*If the probability is uniform between a and b on the usual, linear scale ($\text{Prob}(\text{mass} = m) = 1/(b-a)\,dm$), then on the log scale it is $\text{Prob}(\log \text{mass} = M) = 1/(b-a)e^M\,dM$ [if we change variables to log mass M, then $dM = d(\log m) = 1/m\,dm$, so $dm = m\,dM = e^M\,dM$]. Going the other way, a log-uniform assumption gives $\text{Prob}(\text{mass} = m) = 1/(log(b/a)m)dm$ on the linear scale.

4.3.3.2 Integrating the denominator

The other challenge with Bayesian statistics, which is purely technical and does not raise any deep conceptual issues, is the problem of adding up the denominator $\sum_j P(H_j)P(D|H_j)$ in Bayes' rule. If the set of hypotheses (parameters) is continuous, then the denominator is $\int P(h)P(D|h)\,dh$ where h is a particular parameter value.

For example, the binomial distribution says that the likelihood of obtaining 2 heads in 3 (independent, equal-probability) coin flips is $\binom{3}{2}p^2(1-p)$, a function of p. The likelihood for $p = 0.5$ is therefore 0.375, but to get the posterior probability we have to divide by the probability of getting 2 heads in 3 flips for *any* value of p. Assuming a flat prior, the denominator is $\int_0^1 \binom{3}{2}p^2(1-p)\,dp = 0.25$, so the posterior probability density of $p = 0.5$ is $0.375/0.25 = 1.5^*$.

For the binomial case and other simple probability distributions, it's easy to sum or integrate the denominator either analytically or numerically. If we only care about the *relative* probability of different hypotheses, we don't need to integrate the denominator because it has the same constant value for every hypothesis.

Often, however, we do want to know the absolute probability. Calculating the unconditional probability of the data (the denominator for Bayes' Rule) can be extremely difficult for more complicated problems. Much of current research in Bayesian statistics focuses on ways to calculate the denominator. We will revisit this problem in Chapters 6 and 7, first integrating the denominator by brute-force numerical integration, then looking briefly at a sophisticated technique for Bayesian analysis called Markov chain Monte Carlo.

4.3.4 Conjugate priors

Using so-called *conjugate priors* makes it easy to do the math for Bayesian analysis. Imagine that we're flipping coins (or measuring tadpole survival or counting numbers of different morphs in a fixed sample) and that we use the binomial distribution to model the data. For a binomial with a per-trial probability of p and N trials, the probability of x successes is proportional (leaving out the normalization constant) to $p^x(1-p)^{N-x}$. Suppose that instead of describing the probability of x successes with a fixed per-trial

*This value is a probability density, not a probability, so it's OK for it to be greater than 1: probability density will be explained on p. 156.

probability p and number of trials N we wanted to describe the probability of a given per-trial probability p with fixed x and N. We would get Prob(p) proportional to $p^x(1-p)^{N-x}$ — *exactly the same formula*, but with a different proportionality constant and a different interpretation. Instead of a discrete probability distribution over a sample space of all possible numbers of successes (0 to N), now we have a continuous probability distribution over all possible probabilities (all values between 0 and 1). The second distribution, for Prob(p), is called the Beta distribution (p. 176) and it is the conjugate prior for the binomial distribution.

Mathematically, conjugate priors have the same structure as the probability distribution of the data. They lead to a posterior distribution with the same mathematical form as the prior, although with different parameter values. Intuitively, you get a conjugate prior by turning the likelihood around to ask about the probability of a parameter instead of the probability of the data.

We'll come back to conjugate priors and how to use them in Chapters 6 and 7.

4.4 ANALYZING PROBABILITY DISTRIBUTIONS

You need the same kinds of skills and intuitions about the characteristics of probability distributions that we developed in Chapter 3 for mathematical functions.

4.4.1 Definitions

Discrete

A probability distribution is the set of probabilities on a sample space or set of outcomes. Since this book is about modeling quantitative data, we will always be dealing with sample spaces that are numbers — the number or amount observed in some measurement of an ecological system. The simplest distributions to understand are *discrete* distributions whose outcomes are a set of integers: most of the discrete distributions we'll deal with describe counting or sampling processes and have ranges that include some or all of the non-negative integers.

A discrete distribution is most easily described by its distribution function, which is just a formula for the probability that the outcome of an experiment or observation (called a *random variable*) X is equal to a particular

value x ($f(x) = \text{Prob}(X = x)$). A distribution can also be described by its cumulative distribution function $F(x)$ (note the uppercase F), which is the probability that the random variable X is less than or equal to a particular value x ($F(x) = \text{Prob}(X \leq x)$. Cumulative distribution functions are most useful for frequentist calculations of tail probabilities, e.g. the probability of getting n or more heads in a series of coin-tossing experiments with a given trial probability.

Continuous

A probability distribution over a continuous range (such as all real numbers, or the non-negative real numbers) is called a *continuous* distribution. The cumulative distribution function of a continuous distribution ($F(x) = \text{Prob}(X \leq x)$ is easy to define and understand — it's just the probability that the continuous random variable X is smaller than a particular value x in any given observation or experiment — but the probability *density function* (the analogue of the distribution function for a discrete distribution) is more confusing, since the probability of any precise value is zero. You may imagine that a measurement of (say) pH is *exactly* 7.9, but in fact what you have observed is that the pH is between 7.82 and 7.98 — if your meter has a precision of $\pm$ 1%. Thus continuous probability distributions are expressed as probability *densities* rather than probabilities — the probability that random variable X is between x and $x + \Delta x$, divided by Δx ($\text{Prob}(7.82 < X < 7.98)/0.16$, in this case). Dividing by Δx allows the observed probability density to have a well-defined limit as precision increases and Δx shrinks to zero. Unlike probabilities, Probability densities can be larger than 1 (Figure 4.5). For example, if the pH probability distribution is uniform on the interval [7,7.1] but zero everywhere else, its probability density is 10. In practice, we will mostly be concerned with *relative* probabilities or likelihoods, and so the maximum density values and whether they are greater than or less than 1 won't matter much.

4.4.2 Means (expectations)

The first thing you usually want to know about a distribution is its average value, also called its mean or expectation.

In general the expectation operation, denoted by $E[\cdot]$ (or a bar over a variable, such as $\bar{x}$) gives the "expected value" of a set of data, or a probability distribution, which in the simplest case is the same as its (arithmetic) mean value. For a set of N data values written down separately as

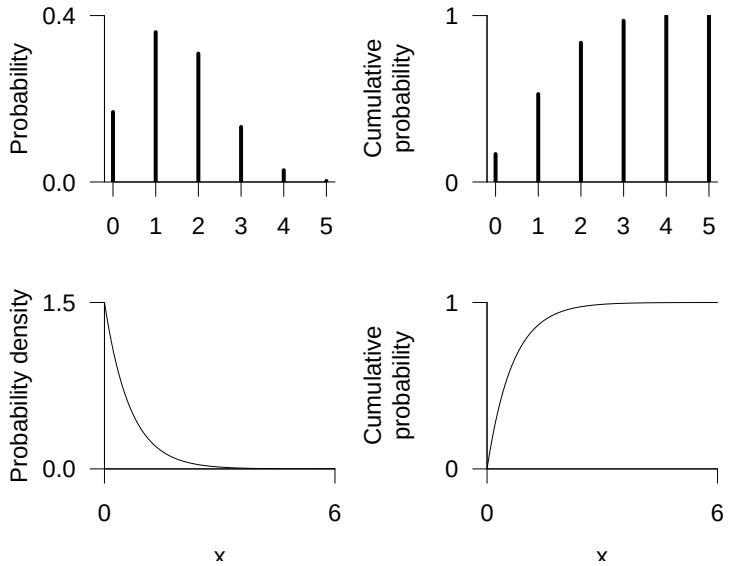

Figure 4.5 Probability, probability density, and cumulative distributions. Top: discrete (binomial: $N = 5$, $p = 0.3$) probability and cumulative probability distributions. Bottom: continuous (exponential: $\lambda = 1.5$) probability density and cumulative probability distributions.

$\{x_1, x_2, x_3, \ldots x_N\}$, the formula for the mean is familiar:

$$E[x] = \frac{\sum_{i=1}^{N} x_i}{N}. \tag{4.4.1}$$

Suppose we have the data tabulated instead, so that for each possible value of x (for a discrete distribution) we have a count of the number of observations (possibly zero, possibly more than 1), which we call $c(x)$. Summing over all of the possible values of x, we have

$$E[x] = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{\sum c(x)x}{N} = \sum \left(\frac{c(x)}{N}\right) x = \sum \text{Prob}(x)x \tag{4.4.2}$$

where $\text{Prob}(x)$ is the discrete probability distribution representing this particular data set. More generally, you can think of $\text{Prob}(x)$ as representing some particular theoretical probability distribution which only approximately matches any actual data set.

We can compute the mean of a continuous distribution as well. First, let's think about grouping (or "binning") the values in a discrete distribution into categories of size Δx. Then if $p(x)$, the density of counts in bin x, is $c(x)/\Delta x$, the formula for the mean becomes $\sum p(x) \cdot x \Delta x$. If we have a continuous distribution with Δx very small, this becomes $\int p(x)x\,dx$.

(This is in fact the definition of an integral.) For example, an exponential distribution $p(x) = \lambda \exp(-\lambda x)$ has an expectation or mean value of $\int \lambda \exp(-\lambda x) x\, dx = 1/\lambda$. (You don't need to know how to do this integral analytically, although the R supplement will show a little bit about numerical integration in R.)

4.4.3 Variances (expectation of X^2)

The mean is the expectation of the random variable X itself, but we can also ask about the expectation of functions of X. The first example is the expectation of X^2. We just fill in the value x^2 for x in all of the formulas above: $E[x^2] = \sum \text{Prob}(x)x^2$ for a discrete distribution, or $\int p(x)x^2\, dx$ for a continuous distribution. (We are *not* asking for $\sum \text{Prob}(x^2)x^2$.) The expectation of x^2 is a component of the variance, which is the expected value of $(x - E[x])^2$ or $(x - \bar{x})^2$, or the expected squared deviation around the mean. (We can also show that

$$E[(x - \bar{x})^2] = E[x^2] - (\bar{x})^2 \tag{4.4.3}$$

by using the rules for expectations that (1) $E[x + y] = E[x] + E[y]$ and (2) if c is a constant, $E[cx] = cE[x]$. The right-hand formula formula is simpler to compute than $E[(x - \bar{x})^2]$, but more subject to roundoff error.)

Variances are easy to work with because they are additive (we will show later that $\text{Var}(a + b) = \text{Var}(a) + \text{Var}(b)$ if a and b are uncorrelated), but harder to compare with means since their units are the units of the mean squared. Thus we often use instead the *standard deviation* of a distribution, $(\sqrt{\text{Var}})$, which has the same units as X.

Two other summaries related to the variance are the *variance-to-mean* ratio and the *coefficient of variation* (CV), which is the ratio of the standard deviation to the mean. The variance-to-mean ratio has units equal to the mean; it is primarily used to characterize discrete sampling distributions and compare them to the Poisson distribution, which has a variance-to-mean ratio of 1. The CV is more common, and is useful when you want to describe variation that is proportional to the mean. For example, if you have a pH meter that is accurate to $\pm 10\%$, so that a true pH value of x will give measured values that are normally distributed with $2\sigma = 0.1x$*, then $\sigma = 0.05x$ and the CV is 0.05.

*Remember that the 95% confidence limits of the normal distribution are approximately $\mu \pm 2\sigma$.

4.4.4 Higher moments

The expectation of $(x - E[x])^3$ tells you the *skewness* of a distribution or a data set, which indicates whether it is asymmetric around its mean. The expectation $E[(x - E[x])^4]$ measures the *kurtosis*, the "pointiness" or "flatness", of a distribution. These are called the third and fourth *central moments* of the distribution. In general, the n^{th} moment is $E[x^n]$, and the n^{th} central moment is $E[(x - \bar{x})^n]$; the mean is the first moment, and the variance is the second central moment. We won't be too concerned with these summaries (of data or distributions), but they do come up sometimes.

4.4.5 Median and mode

The median and mode are two final properties of probability distributions that are not related to moments. The *median* of a distribution is the point which divides the area of the probability density in half, or the point at which the cumulative distribution function is equal to 0.5. It is often useful for describing data, since it is *robust* — outliers change its value less than they change the mean — but for many distributions it's more complicated to compute than the mean. The *mode* is the "most likely value", the maximum of the probability distribution or density function. For symmetric distributions the mean, mode, and median are all equal; for right-skewed distributions, in general mode $<$ median $<$ mean.

4.4.6 The method of moments

Suppose you know the theoretical values of the moments (e.g. mean and variance) of a distribution and have calculated the sample values of the moments (by calculating $\bar{x} = \sum x/N$ and $s^2 = \sum(x-\bar{x})^2/N$: don't worry for the moment about whether the denominator in the sample variance should be N or $N - 1$). Then there is a simple way to estimate the parameters of a distribution, called the *method of moments*: just match the sample values up with the theoretical values. For the normal distribution, where the parameters of the distribution are just the mean and the variance, this is trivially simple: $\mu = \bar{x}$, $\sigma^2 = s^2$. For a distribution like the negative binomial, however (p. 165), it involves a little bit of algebra. The negative binomial has parameters μ (equal to the mean, so that's easy) and k; the theoretical variance is $\sigma^2 = \mu(1 + \mu/k)$. Therefore, setting $\mu = \bar{x}$, $s^2 \approx \mu(1+\mu/k)$, and solving for k, we calculate the method-of-moments estimate

of k:

$$\begin{aligned} \sigma^2 &= \mu(1+\mu/k) \\ s^2 &\approx \bar{x}(1+\bar{x}/k) \\ \frac{s^2}{\bar{x}} - 1 &\approx \frac{\bar{x}}{k} \\ k &\approx \frac{\bar{x}}{s^2/\bar{x} - 1} \end{aligned} \tag{4.4.4}$$

The method of moments is very simple but is biased in many cases; it's a good way to get a first estimate of the parameters of a distribution, but for serious work you should follow it up with a maximum likelihood estimator (Chapter 6).

4.5 BESTIARY OF DISTRIBUTIONS

The rest of the chapter presents brief introductions to a variety of useful probability distributions, including the mechanisms behind them and some of their basic properties. Like the bestiary in Chapter 3, you can skim this bestiary on the first reading. The appendix of Gelman et al. (1996) contains a useful table, more abbreviated than these descriptions but covering a wider range of functions. The book by Evans et al. (2000) is also useful.

4.5.1 Discrete models

4.5.1.1 Binomial

The binomial is probably the easiest distribution to understand. It applies when you have samples with a fixed number of subsamples or "trials" in each one, and each trial can have one of two values (black/white, heads/tails, alive/dead, species A/species B), and the probability of "success" (black, heads, alive, species A) is the same in every trial. If you flip a coin 10 times ($N = 10$) and the probability of a head in each coin flip is $p = 0.7$ then the probability of getting 7 heads ($k = 7$) will will have a binomial distribution with parameters $N = 10$ and $p = 0.7$* Don't confuse the trials (subsamples), and the probability of success in each trial, with the number of samples and the probabilities of the number of successful

*Gelman and Nolan (2002) point out that it is not physically possible to construct a coin that is biased when flipped — although a spinning coin can be biased. Diaconis et al. (2004) even tested a coin made of balsa wood on one side and lead on the other to establish that it was unbiased.

trials in each sample. In the seed predation example, a trial is an individual seed and the trial probability is the probability that an individual seed is taken, while a sample is the observation of a particular station at a particular time and the binomial probabilities are the probabilities that a certain total number of seeds disappears from the station. You can derive the part of the distribution that depends on x, $p^x(1-p)^{N-x}$, by multiplying the probabilities of x independent successes with probability p and $N-x$ independent failures with probability $1-p$. The rest of the distribution function, $\binom{N}{x} = N!/(x!(N-x)!)$, is a *normalization constant* that we can justify either with a combinatorial argument about the number of different ways of sampling x objects out of a set of N (Appendix), or simply by saying that we need a factor in front of the formula to make sure the probabilities add up to 1.

The variance of the binomial is $Np(1-p)$. Like most discrete sampling distributions (e.g. the binomial, Poisson, negative binomial), this variance depends on the number of samples per trial N. When the number of samples per trial increases the variance also increases, but the coefficient of variation ($\sqrt{Np(1-p)}/(Np) = \sqrt{(1-p)/(Np)}$) decreases. The dependence on $p(1-p)$ means the binomial variance is small when p is close to 0 or 1 (and therefore the values are scrunched up near 0 or N), and largest when $p = 0.5$. The coefficient of variation, on the other hand, is largest for small p.

When N is large and p isn't too close to 0 or 1 (i.e. when Np is large), then the binomial distribution is approximately normal (Figure 4.17).

A binomial distribution with only one trial ($N = 1$) is called a *Bernoulli* trial.

You should only use the binomial in fitting data when there is an upper limit to the number of possible successes. When N is large and p is small, so that the probability of getting N successes is small, the binomial approaches the Poisson distribution, which is covered in the next section (Figure 4.17).

Examples: number of surviving individuals/nests out of an initial sample; number of infested/infected animals, fruits, etc. in a sample; number of a particular class (haplotype, subspecies, etc.) in a larger population.

Summary:

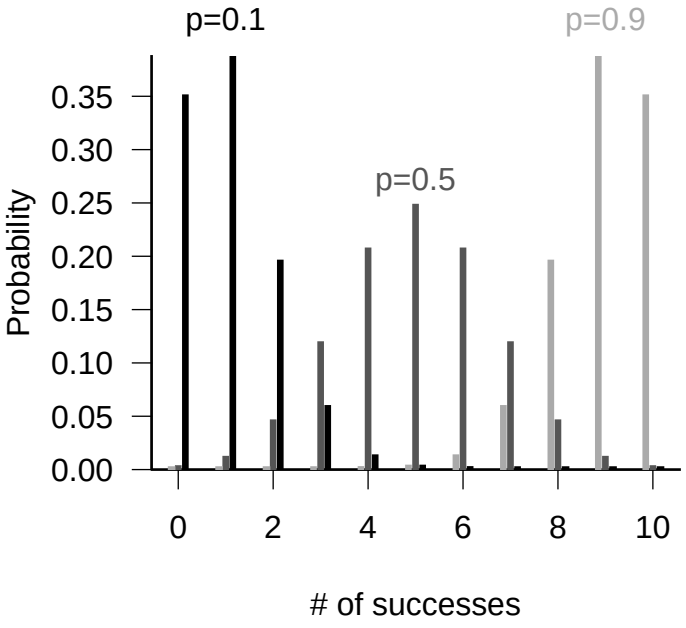

Figure 4.6 Binomial distribution. Number of trials (N) equals 10 for all distributions.

range	discrete, $0 \leq x \leq N$
distribution	$\binom{N}{x} p^x (1-p)^{N-x}$
R	`dbinom`, `pbinom`, `qbinom`, `rbinom`
parameters	p [real, 0–1], probability of success [`prob`]
	N [positive integer], number of trials [`size`]
mean	Np
variance	$Np(1-p)$
CV	$\sqrt{(1-p)/(Np)}$
Conjugate prior	Beta

4.5.1.2 Poisson

The Poisson distribution gives the distribution of the number of individuals, arrivals, events, counts, etc., in a given time/space/unit of counting effort if each event is independent of all the others. The most common definition of the Poisson has only one parameter, the average density or arrival rate, λ, which equals the expected number of counts in a sampling unit. An alternative parameterization gives a density *per unit sampling effort* and then specifies the mean as the product of the density per sampling effort r times the sampling effort t, $\lambda = rt$. This parameterization emphasizes that even when the population density is constant, you can change the Poisson distribution of counts by sampling more extensively — for longer times or over larger quadrats.

The Poisson distribution has no upper limit, although values much larger than the mean value are highly improbable. This characteristic provides a rule for choosing between the binomial and Poisson. If you expect to observe a "ceiling" on the number of counts, you should use the binomial; if you expect the number of counts to be effectively unlimited, even if it is theoretically bounded (e.g. there can't really be an infinite number of plants in your sampling quadrat), use the Poisson.

The variance of the Poisson is equal to its mean. However, the *coefficient of variation* (CV=standard deviation/mean) decreases as the mean increases, so in that sense the Poisson distribution becomes more regular as the expected number of counts increases. The Poisson distribution *only makes sense for count data*. Since the CV is unitless, it should not depend on the units we use to express the data; since the CV of the Poisson is $1/\sqrt{\text{mean}}$, that means that if we used a Poisson distribution to describe data on measured lengths, we could reduce the CV by a factor of 10 by changing from meters to centimeters (which would be silly).

For $\lambda < 1$ the Poisson's mode is at zero. When the expected number of

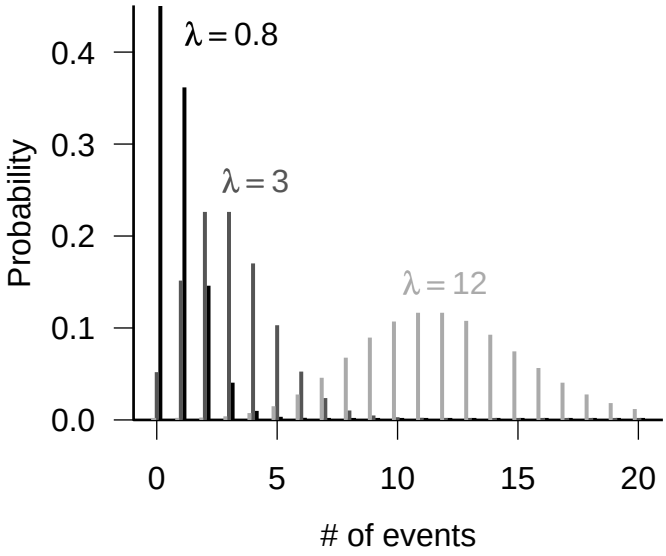

Figure 4.7 Poisson distribution.

counts gets large (e.g. $\lambda > 10$) the Poisson becomes approximately normal (Figure 4.17).

Examples: number of seeds/seedlings falling in a gap; number of offspring produced in a season (although this might be better fit by a binomial if the number of breeding attempts is fixed); number of prey caught per unit time.

Summary:

range	discrete ($0 \leq x$)
distribution	$\frac{e^{-\lambda}\lambda^n}{n!}$ **or** $\frac{e^{-rt}(rt)^n}{n!}$
R	`dpois`, `ppois`, `qpois`, `rpois`
parameters	λ (real, positive), expected number per sample [`lambda`] **or** r (real, positive), expected number per unit effort, area, time, etc. (*arrival ra*
mean	λ (**or** rt)
variance	λ (**or** rt)
CV	$1/\sqrt{\lambda}$ (**or** $1/\sqrt{rt}$)
Conjugate prior	Gamma

4.5.1.3 Negative binomial

Most probability books derive the negative binomial distribution from a series of independent binary (heads/tails, black/white, male/female, yes/no) trials that all have the same probability of success, like the binomial distribution. Rather than count the number of successes obtained in a fixed number of trials, which would result in a binomial distribution, the negative binomial counts the number of *failures* before a predetermined number of successes occurs.

This failure-process parameterization is only occasionally useful in ecological modeling. Ecologists use the negative binomial because it is discrete, like the Poisson, but its variance can be larger than its mean (i.e. it can be *overdispersed*). Thus, it's a good phenomenological description of a patchy or clustered distribution with no intrinsic upper limit that has more variance than the Poisson.

The "ecological" parameterization of the negative binomial replaces the parameters p (probability of success per trial: `prob` in R) and n (number of successes before you stop counting failures: `size` in R) with $\mu = n(1-p)/p$, the mean number of failures expected (or of counts in a sample: `mu` in R), and k, which is typically called an *overdispersion parameter*. Confusingly, k is also called `size` in R, because it is mathematically equivalent to n in the failure-process parameterization.

The overdispersion parameter measures the amount of clustering, or aggregation, or heterogeneity, in the data: a smaller k means more heterogeneity. The variance of the negative binomial distribution is $\mu + \mu^2/k$, and so as k becomes large the variance approaches the mean and the distribution approaches the Poisson distribution. For $k > 10$, the negative binomial is hard to tell from a Poisson distribution, but k is often less than 1 in ecological applications*.

Specifically, you can get a negative binomial distribution as the result of a Poisson sampling process where the rate λ itself varies. If the distribution of λ is a gamma distribution (p. 172) with shape parameter k and mean μ, and x is Poisson-distributed with mean λ, then the distribution of x be a negative binomial distribution with mean μ and overdispersion parameter k (May, 1978; Hilborn and Mangel, 1997). In this case, the negative binomial reflects unmeasured ("random") variability in the population.

*Beware of the word "overdispersion", which is sometimes used with an opposite meaning in spatial statistics, where it can mean "more regular than expected from a random distribution of points". If you took quadrat samples from such an "overdispersed" population, the distribution of counts would have variance less than the mean and be "underdispersed" in the probability distribution sense (Brown and Bolker, 2004) (!)

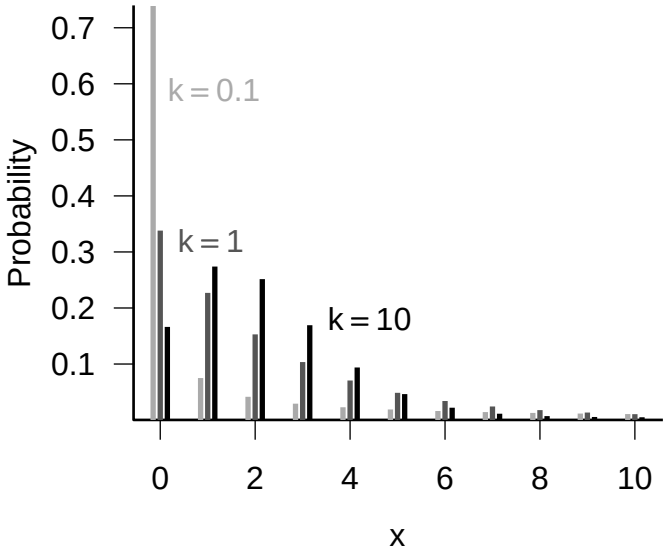

Figure 4.8 Negative binomial distribution. Mean $\mu = 2$ in all cases.

Negative binomial distributions can also result from a homogeneous birth-death process, births and deaths (and immigrations) occurring at random in continuous time. Samples from a population that starts from 0 at time $t = 0$, with immigration rate i, birth rate b, and death rate d will be negative binomially distributed with parameters $\mu = i/(b-d)(e^{(b-d)t} - 1)$ and $k = i/b$ (Bailey, 1964, p. 99).

Several different ecological processes can often generate the same probability distribution. We can usually reason forward from knowledge of probable mechanisms operating in the field to plausible distributions for modeling data, but this many-to-one relationship suggests that it is unsafe to reason backwards from probability distributions to particular mechanisms that generate them.

Examples: essentially the same as the Poisson distribution, but allowing for heterogeneity. Numbers of individuals per patch; distributions of numbers of parasites within individual hosts; number of seedlings in a gap, or per unit area, or per seed trap.

Summary:

range	discrete, $x \geq 0$
distribution	$\frac{(n+x-1)!}{(n-1!)x!}p^n(1-p)^x$
	or $\frac{\Gamma(k+x)}{\Gamma(k)x!}(k/(k+\mu))^k(\mu/(k+\mu))^x$
R	`dnbinom`, `pnbinom`, `qnbinom`, `rnbinom`
parameters	p $(0 < p < 1)$ probability per trial [`prob`]
	or μ (real, positive) expected number of counts [`mu`]
	n (positive integer) number of successes awaited [`size`]
	or k (real, positive), overdispersion parameter [`size`]
	(= shape parameter of underlying heterogeneity)
mean	$\mu = n(1-p)/p$
variance	$\mu + \mu^2/k = n(1-p)/p^2$
CV	$\sqrt{\frac{(1+\mu/k)}{\mu}} = 1/\sqrt{n(1-p)}$
Conjugate prior	No simple conjugate prior (Bradlow et al., 2002)

R's default coin-flipping (n =`size`, p =`prob`) parameterization. In order to use the "ecological" (μ =`mu`, k =`size`) parameterization, you *must* name the `mu` parameter explicitly (e.g. `dnbinom(5,size=0.6,mu=1)`).

4.5.1.4 Geometric

The geometric distribution is the number of trials (with a constant probability of failure) until you get a single failure: it's a special case of the negative binomial, with k or $n = 1$.

Examples: number of successful/survived breeding seasons for a seasonally reproducing organism. Lifespans measured in discrete units.

Summary:

range	discrete, $x \geq 0$
distribution	$p(1-p)^x$
R	`dgeom`, `pgeom`, `qgeom`, `rgeom`
parameters	p $(0 < p < 1)$ probability of "success" (death) [`prob`]
mean	$1/p - 1$
variance	$(1-p)/p^2$
CV	$1/\sqrt{1/(1-p)}$

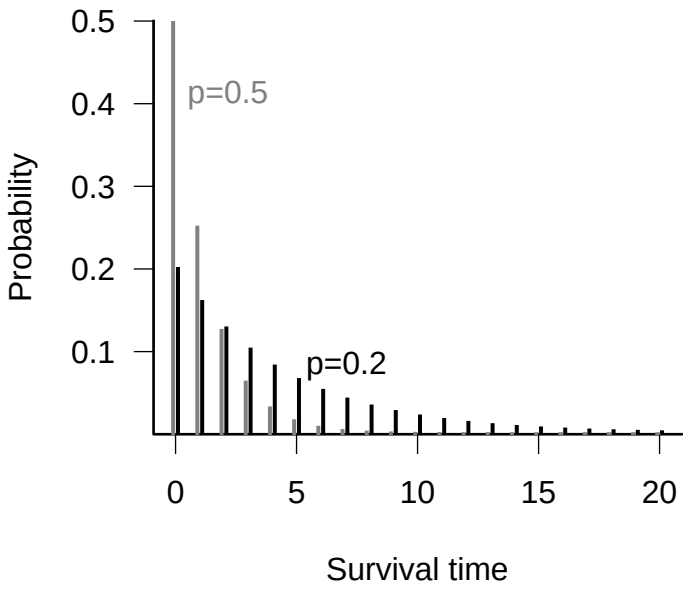

Figure 4.9 Geometric distribution.

4.5.1.5 Beta-binomial

Just as one can compound the Poisson distribution with a Gamma to allow for heterogeneity in rates, producing a negative binomial, one can compound the binomial distribution with a Beta distribution to allow for heterogeneity in per-trial probability, producing a *Beta-binomial* distribution (Crowder, 1978; Reeve and Murdoch, 1985; Hatfield et al., 1996). The most common parameterization of the beta-binomial distribution uses the binomial parameter N (trials per sample), plus two additional parameters a and b that describe the beta distribution of the per-trial probability. When $a = b = 1$ the per-trial probability is equally likely to be any value between 0 and 1 (the mean is 0.5), and the beta-binomial gives a uniform (discrete) distribution between 0 and N. As $a + b$ increases, the variance of the underlying heterogeneity decreases and the beta-binomial converges to the binomial distribution. Morris (1997) suggests a different parameterization that uses an overdispersion parameter θ, like the k parameter of the negative binomial distribution. In this case the parameters are N, the per-trial probability p $(= a/(a + b))$, and θ $(= a + b)$. When θ is large (small overdispersion), the beta-binomial becomes binomial. When θ is near zero (large overdispersion), the beta-binomial becomes U-shaped (Figure 4.10).

Summary:

range	discrete, $0 \leq x \leq N$
R	`dbetabinom`, `rbetabinom` [`emdbook` package] (`pbetabinom` and `qbetabinom` are missing)
density	$\frac{\Gamma(\theta)}{\Gamma(p\theta)\Gamma((1-p)\theta)} \cdot \frac{N!}{x!(N-x)!} \cdot \frac{\Gamma(x+p\theta)\Gamma(N-x+(1-p)\theta)}{\Gamma(N+\theta)}$
parameters	p (real, positive), probability: average per-trial probability [`prob`] θ (real, positive), overdispersion parameter [`theta`] **or** a and b (shape parameters of Beta distribution for per-trial probability) [`shape1` and `shape2`] $a = \theta p$, $b = \theta(1 - p)$
mean	Np
variance	$Np(1-p)\left(1 + \frac{N-1}{\theta+1}\right)$
CV	$\sqrt{\frac{(1-p)}{Np}\left(1 + \frac{N-1}{\theta+1}\right)}$

Examples: as for the binomial.

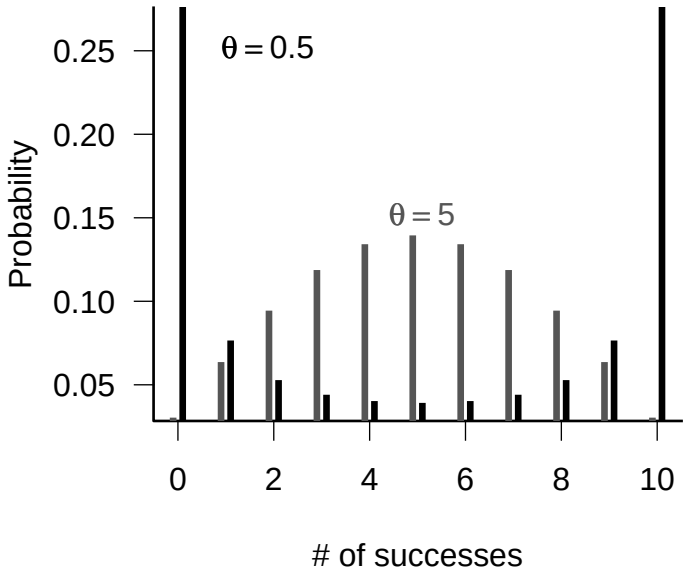

Figure 4.10 Beta-binomial distribution. Number of trials (N) equals 10, average per-trial probability (p) equals 0.5 for all distributions.

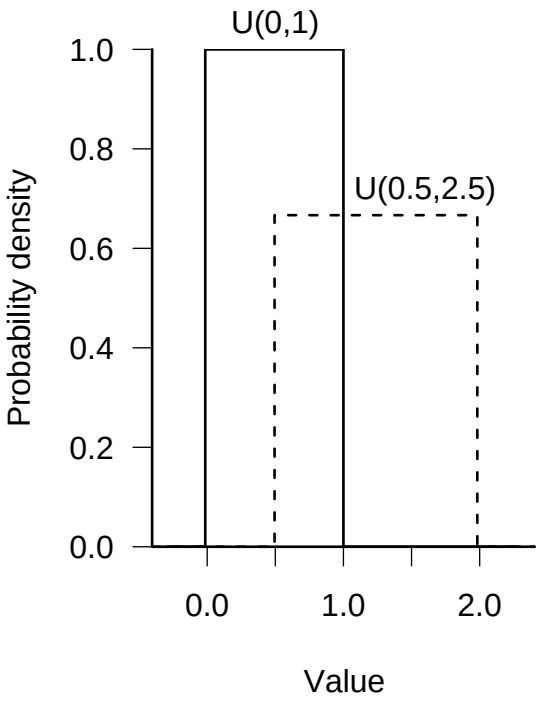

Figure 4.11 Uniform distribution.

4.5.2 Continuous distributions

4.5.2.1 Uniform distribution

The uniform distribution with limits a and b, denoted $U(a, b)$, has a constant probability density of $1/(b-a)$ for $a \leq x \leq b$ and zero probability elsewhere. The standard uniform, $U(0, 1)$, is very commonly used as a building block for other distributions, but is surprisingly rarely used in ecology otherwise.

Summary:

range	$a \leq x \leq b$
distribution	$1/(b-a)$
R	`dunif`, `punif`, `qunif`, `runif`
parameters	minimum (a) and maximum (b) limits (real) [`min`, `max`]
mean	$(a+b)/2$
variance	$(b-a)^2/12$
CV	$(b-a)/((a+b)\sqrt{3})$

4.5.2.2 Normal distribution

Normally distributed variables are everywhere, and most classical statistical methods use this distribution. The explanation for the normal distribution's ubiquity is the *Central Limit Theorem*, which says that if you add a large number of independent samples from the same distribution the distribution of the sum will be approximately normal. "Large", for practical purposes, can mean as few as 5. The central limit theorem does *not* mean that "all samples with large numbers are normal". One obvious counterexample is two different populations with different means that are lumped together, leading to a distribution with two peaks (p. 183). Also, adding isn't the only way to combine samples: if you multiply independent samples from the same distribution, you get a log-normal distribution instead of a normal distribution (p. 178).

Many distributions (binomial, Poisson, negative binomial, gamma) become approximately normal in some limit (Figure 4.17). You can usually think about this as some form of "adding lots of things together".

The normal distribution specifies the mean and variance separately, with two parameters, which means that one often assumes constant variance (as the mean changes), in contrast to the Poisson and binomial distribution where the variance is a fixed function of the mean.

Examples: practically everything.

Summary:

range	all real values
distribution	$\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$
R	`dnorm`, `pnorm`, `qnorm`, `rnorm`
parameters	μ (real), mean [`mean`]
	σ (real, positive), standard deviation [`sd`]
mean	μ
variance	σ^2
CV	σ/μ
Conjugate prior	Normal (μ); Gamma ($1/\sigma^2$)

4.5.2.3 Gamma

The *Gamma* distribution is the distribution of *waiting times* until a certain number of events take place. For example, Gamma(shape $= 3, \text{scale} = 2$) is the distribution of the length of time (in days) you'd expect to have to

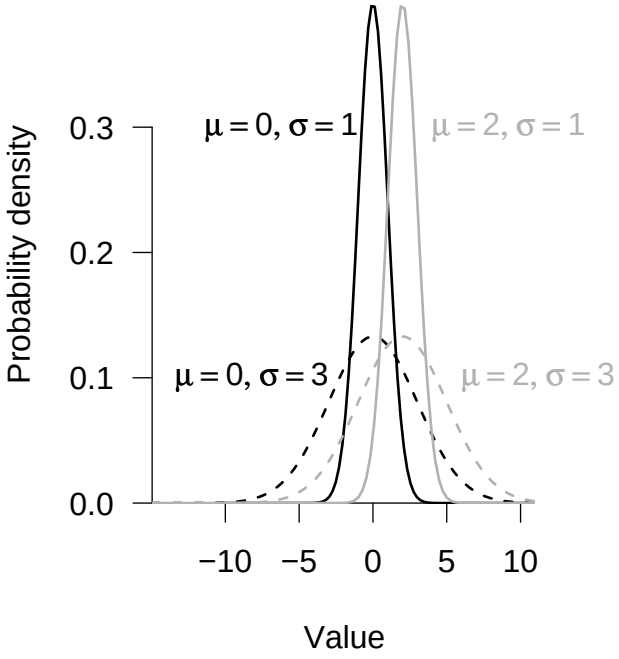

Figure 4.12 Normal distribution

wait for 3 deaths in a population, given that the average survival time is 2 days (mortality rate is 1/2 per day). The mean waiting time is 6 days=(3 deaths/(1/2 death per day)). (While the gamma *function* (`gamma` in R: see Appendix) is usually written with a capital Greek gamma, Γ, the Gamma distribution (`dgamma` in R) is written out as Gamma.) Gamma distributions with integer shape parameters are also called *Erlang* distributions. The Gamma distribution is still defined for non-integer (positive) shape parameters, but the simple description given above breaks down: how can you define the waiting time until 3.2 events take place?

For shape parameters ≤ 1, the Gamma has its mode at zero; for shape parameter $= 1$, the Gamma is equivalent to the exponential (see below). For shape parameter greater than 1, the Gamma has a peak (mode) at a value greater than zero; as the shape parameter increases, the Gamma distribution becomes more symmetrical and approaches the normal distribution. This behavior makes sense if you think of the Gamma as the distribution of the sum of independent, identically distributed waiting times, in which case it is governed by the Central Limit Theorem.

The scale parameter (sometimes defined in terms of a rate parameter instead, 1/scale) just adjusts the mean of the Gamma by adjusting the waiting time per event; however, multiplying the waiting time by a constant to adjust its mean also changes the variance, so both the variance and the mean depend on the scale parameter.

The Gamma distribution is less familiar than the normal, and new users of the Gamma often find it annoying that in the standard parameterization you can't adjust the mean independently of the variance. You could define a new set of parameters m (mean) and v (variance), with scale $= v/m$ and shape $= m^2/v$ — but then you would find (unlike the normal distribution) the shape changing as you changed the variance. Nevertheless, the Gamma is extremely useful; it solves the problem that many researchers face when they have a continuous variable with "too much variance", whose coefficient of variation is greater than about 0.5. Modeling such data with a normal distribution leads to unrealistic negative values, which then have to be dealt with in some *ad hoc* way like truncating them or otherwise trying to ignore them. The Gamma is often a more realistic alternative.

The Gamma is the continuous counterpart of the negative binomial, which is the discrete distribution of a number of trials (rather than length of time) until a certain number of events occur. Both the negative binomial and Gamma distributions are often generalized, however, in ways that don't necessarily make sense according to their simple mechanistic descriptions (e.g. a Gamma distribution with a shape parameter of 2.3 corresponds to

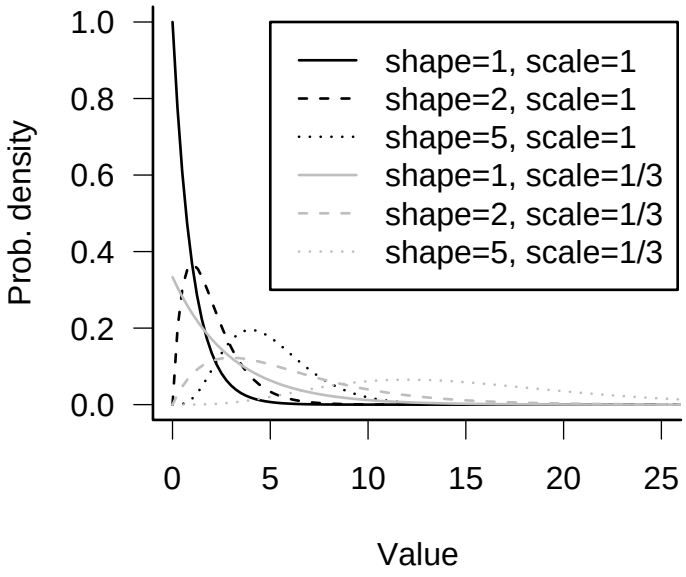

Figure 4.13 Gamma distribution

the distribution of waiting times until 2.3 events occur ...).

The Gamma and negative binomial are both commonly used phenomenologically, as skewed or overdispersed versions of the Poisson or normal distributions, rather than for their mechanistic descriptions. The Gamma is less widely used than the negative binomial because the negative binomial replaces the Poisson, which is restricted to a particular variance, while the Gamma replaces the normal, which can have any variance. Thus you might use the negative binomial for any discrete distribution with variance > mean, while you wouldn't need a Gamma distribution unless the distribution you were trying to match was skewed to the right.

Summary:

range	positive real values
R	`dgamma`, `pgamma`, `qgamma`, `rgamma`
distribution	$\frac{1}{s^a \Gamma(a)} x^{a-1} e^{-x/s}$
parameters	s (real, positive), scale: length per event [`scale`]
	or r (real, positive), rate $= 1/s$; rate at which events occur [`rate`]
	a (real, positive), shape: number of events [`shape`]
mean	as or a/r
variance	as^2 or a/r^2
CV	$1/\sqrt{a}$

Examples: almost any environmental variable with a large variance where negative values don't make sense: nitrogen concentrations, light intensity, etc..

4.5.2.4 Exponential

The exponential distribution (Figure 4.14) describes the distribution of waiting times for a single event to happen, given that there is a constant probability per unit time that it will happen. It is the continuous counterpart of the geometric distribution and a special case (for shape parameter=1) of the Gamma distribution. It can be useful both mechanistically, as a distribution of inter-event times or lifetimes, or phenomenologically, for any continuous distribution that has highest probability for zero or small values.

Examples: times between events (bird sightings, rainfall, etc.); lifespans/survival times; random samples of anything that decreases exponentially (e.g. light levels in a forest canopy).

Summary:

range	positive real values
R	`dexp`, `pexp`, `qexp`, `rexp`
density	$\lambda e^{-\lambda x}$
parameters	λ (real, positive), rate: death/disappearance rate [`rate`]
mean	$1/\lambda$
variance	$1/\lambda^2$
CV	1

4.5.2.5 Beta

The beta distribution, a continuous distribution closely related to the binomial distribution, completes our basic family of continuous distributions

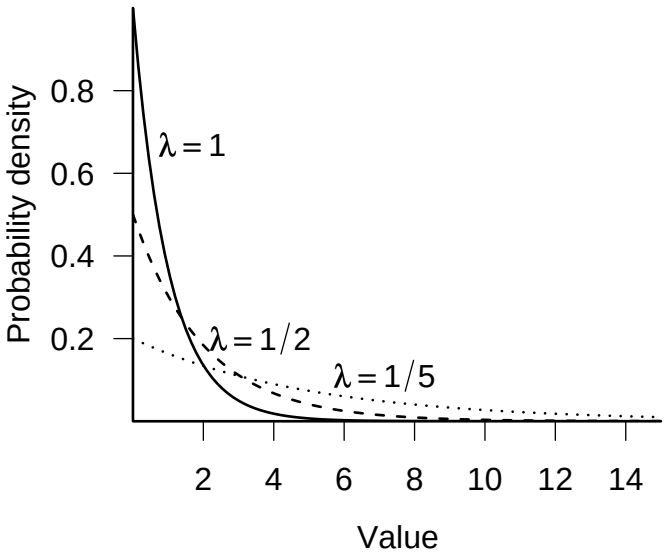

Figure 4.14 Exponential distribution.

(Figure 4.17). The beta distribution is the only standard continuous distribution (besides the uniform distribution) with a finite range, from 0 to 1. The beta distribution is the inferred distribution of the *probability* of success in a binomial trial with $a-1$ observed successes and $b-1$ observed failures. When $a = b$ the distribution is symmetric around $x = 0.5$, when $a < b$ the peak shifts toward zero, and when $a > b$ it shifts toward 1. With $a = b = 1$, the distribution is $U(0, 1)$. As $a + b$ (equivalent to the total number of trials+2) gets larger, the distribution becomes more peaked. For a or b less than 1, the mechanistic description stops making sense (how can you have fewer than zero trials?), but the distribution is still well-defined, and when a and b are both between 0 and 1 it becomes U-shaped — it has peaks at $p = 0$ and $p = 1$.

The beta distribution is obviously good for modeling probabilities or proportions. It can also be useful for modeling continuous distributions with peaks at both ends, although in some cases a finite mixture model (p. 183) may be more appropriate. The beta distribution is also useful whenever you have to define a continuous distribution on a finite range, as it is the only such standard continuous distribution. It's easy to rescale the distribution so that it applies over some other finite range instead of from 0 to 1: for example, Tiwari et al. (2005) used the beta distribution to describe the distribution of turtles on a beach, so the range would extend from 0 to the length of the beach.

Summary:

range	real, 0 to 1
R	`dbeta`, `pbeta`, `qbeta`, `rbeta`
density	$\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}$
parameters	a (real, positive), shape 1: number of successes +1 [`shape1`]
	b (real, positive), shape 2: number of failures +1 [`shape2`]
mean	$a/(a+b)$
mode	$(a-1)/(a+b-2)$
variance	$ab/((a+b)^2)(a+b+1)$
CV	$\sqrt{(b/a)/(a+b+1)}$

4.5.2.6 Lognormal

The lognormal falls outside the neat classification scheme we've been building so far; it is not the continuous analogue or limit of some discrete sampling distribution (Figure 4.17)*. Its mechanistic justification is like the normal

*The lognormal extends our table in another direction — exponential transformation of a known distribution. Other distributions have this property, most notably the *extreme value distri-*

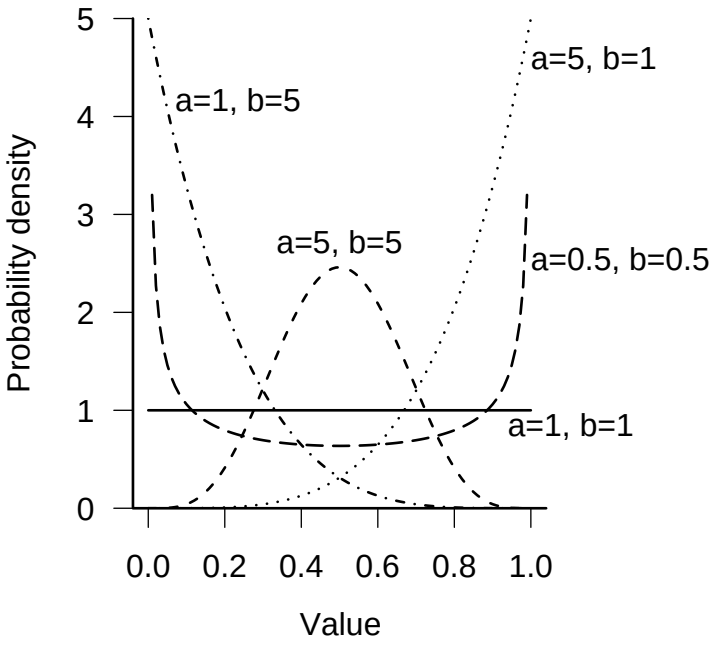

Figure 4.15 Beta distribution

distribution (the Central Limit Theorem), but for the *product* of many independent, identical variates rather than their sum. Just as taking logarithms converts products into sums, taking the logarithm of a lognormally distributed variable—which might result from the product of independent variables—converts it it into a normally distributed variable resulting from the sum of the logarithms of those independent variables. The best example of this mechanism is the distribution of the sizes of individuals or populations that grow exponentially, with a per capita growth rate that varies randomly over time. At each time step (daily, yearly, etc.), the current size is *multiplied* by the randomly chosen growth increment, so the final size (when measured) is the product of the initial size and all of the random growth increments.

One potentially puzzling aspect of the lognormal distribution is that its mean is not what you might naively expect if you exponentiate a normal distribution with mean μ (i.e. e^{μ}). Because of Jensen's inequality, and because the exponential function is an accelerating function, the mean of the lognormal, $e^{\mu+\sigma^2/2}$, is greater than e^{μ} by an amount that depends on the variance of the original normal distribution. When the variance is small relative to the mean, the mean is approximately equal to e^{μ}, and the lognormal itself looks approximately normal (e.g. solid lines in Figure 4.16, with $\sigma(\log) = 0.2$). As with the Gamma distribution, the distribution also changes shape as the variance increases, becoming more skewed.

The log-normal is also used phenomenologically in some of the same situations where a Gamma distribution also fits: continuous, positive distributions with long tails or variance much greater than the mean (McGill et al., 2006). Like the distinction between a Michaelis-Menten and a saturating exponential, you may not be able to tell the difference between a lognormal and a Gamma without large amounts of data. Use the one that is more convenient, or that corresponds to a more plausible mechanism for your data.

Examples: sizes or masses of individuals, especially rapidly growing individuals; abundance vs. frequency curves for plant communities.

Summary:

bution, which is the log-exponential: if Y is exponentially distributed, then $\log Y$ is extreme-value distributed. As its name suggests, the extreme value distribution occurs mechanistically as the distribution of extreme values (e.g. maxima) of samples of other distributions (Katz et al., 2005).

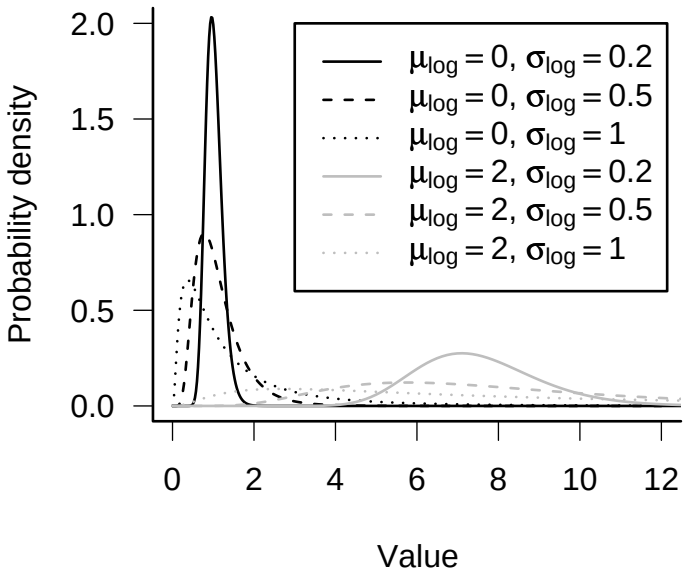

Figure 4.16 Lognormal distribution

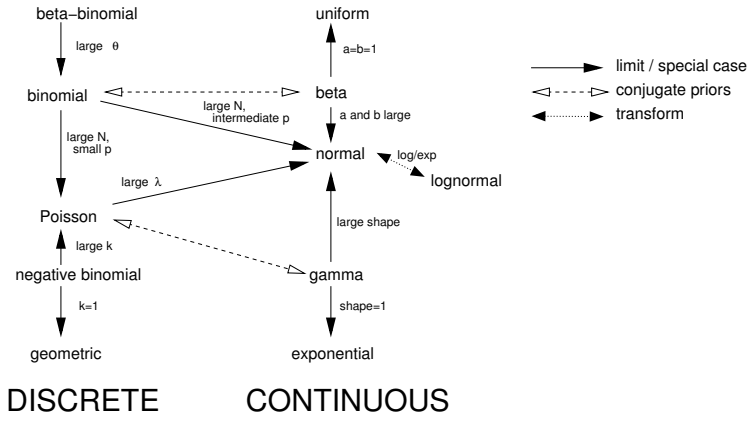

Figure 4.17 Relationships among probability distributions.

range	positive real values
R	`dlnorm`, `plnorm`, `qlnorm`, `rlnorm`
density	$\frac{1}{\sqrt{2\pi}\sigma x} e^{-(\log x - \mu)^2/(2\sigma^2)}$
parameters	μ (real): mean of the logarithm [`meanlog`]
	σ (real): standard deviation of the logarithm [`sdlog`]
mean	$\exp(\mu + \sigma^2/2)$
variance	$\exp(2\mu + \sigma^2)(\exp(\sigma^2) - 1)$
CV	$\sqrt{\exp(\sigma^2) - 1}$ ($\approx \sigma$ when $\sigma < 1/2$)

4.6 EXTENDING SIMPLE DISTRIBUTIONS; COMPOUNDING AND GENERALIZING

What do you do when none of these simple distributions fits your data? You could always explore other distributions. For example, the Weibull distribution (similar to the Gamma distribution in shape: `?dweibull` in R) generalizes the exponential to allow for survival probabilities that increase or decrease with age (p. 331). The Cauchy distribution (`?dcauchy` in R), described as *fat-tailed* because the probability of extreme events (in the tails of the distribution) is very large — larger than for the exponential or normal distributions — can be useful for modeling distributions with many outliers. You can often find useful distributions for your data in modeling papers from your subfield of ecology.

However, in addition to simply learning more distributions it can also useful to learn some strategies for generalizing more familiar distributions.

4.6.1 Adding covariates

One obvious strategy is to look for systematic differences within your data that explain the non-standard shape of the distribution. For example, a *bimodal* or *multimodal* distribution (one with two or more peaks, in contrast to most of the distributions discussed above that have a single peak) may make perfect sense once you realize that your data are a collection of objects from different populations with different means.For example, the sizes or masses of sexually dimorphic animals or animals from several different cryptic species would bi- or multimodal distributions, respectively. A distribution that isn't multimodal but is more fat-tailed than a normal distribution might indicate systematic variation in a continuous covariate such as nutrient availability, or maternal size, of environmental temperature, of different individuals.

4.6.2 Mixture models

But what if you can't identify systematic differences? You can still extend standard distributions by supposing that your data are really a mixture of observations from different types of individuals, but that you can't observe the (finite) types or (continuous) covariates of individuals. These distributions are called *mixture distributions* or *mixture models*. Fitting them to data can be challenging, but they are very flexible.

4.6.2.1 Finite mixtures

Finite mixture models suppose that your observations are drawn from a discrete set of unobserved categories, each of which has its own distribution: typically all categories have the same type of distribution, such as normal, but with different mean or variance parameters. Finite mixture distributions often fit multimodal data. Finite mixtures are typically parameterized by the parameters of each component of the mixture, plus a set of probabilities or percentages describing the amount of each component. For example, 30% of the organisms ($p = 0.3$) could be in group 1, normally distributed with mean 1 and standard deviation 2, while 70% ($1 - p = 0.7$) are in group 2, normally distributed with mean 5 and standard deviation 1 (Figure 4.18). If the peaks of the distributions are closer together, or their standard deviations are larger so that the distributions overlap, you'll see a broad (and perhaps lumpy) peak rather than two distinct peaks.

Zero-inflated models are a common type of finite mixture model (In-

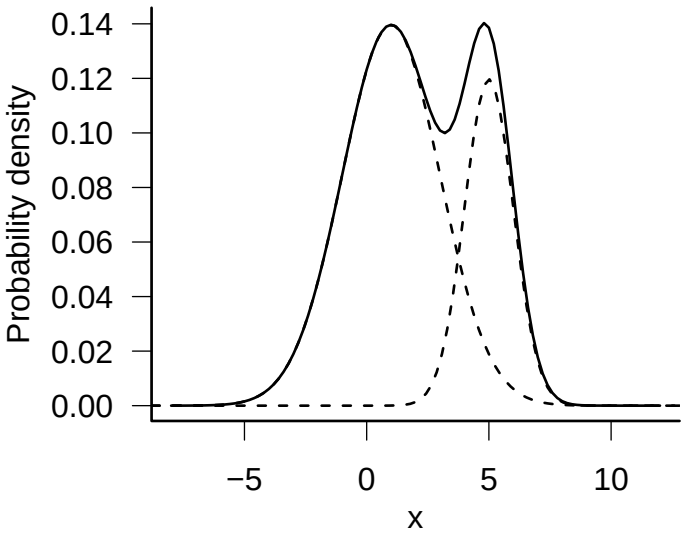

Figure 4.18 Finite mixture distribution: 70% Normal($\mu = 1, \sigma = 2$), 30% Normal($\mu = 5, \sigma = 1$).

ouye, 1999; Martin et al., 2005). Zero-inflated models (Figure 4.1). combine a standard discrete probability distribution (e.g. binomial, Poisson, or negative binomial), which typically include some probability of sampling zero counts even when some individuals are present, with some additional process that can also lead to a zero count (e.g. complete absence of the species or trap failure).

4.6.3 Continuous mixtures

Continuous mixture distributions, also known as *compounded distributions*, allow the parameters themselves to vary randomly, drawn from their own distribution. They are a sensible choice for overdispersed data, or for data where you suspect that unobserved covariates may be important. Technically, compounded distributions are the distribution of a sampling distribution $S(x, p)$ with parameter(s) p that vary according to another (typically continuous) distribution $P(p)$. The distribution of the compounded distribution C is $C(x) = \int S(x, p)P(p)dp$. For example, compounding a Poisson distribution by drawing the rate parameter λ from a Gamma distribution with shape parameter k (and scale parameter λ/k, to make the mean equal to λ) results in a negative binomial distribution (p. 165). Continuous mixture distributions are growing ever more popular in ecology as ecologists try to account for heterogeneity in their data.

The negative binomial, which could also be called the Gamma-Poisson distribution to highlight its compound origin, is the most common compounded distribution. The Beta-binomial is also fairly common: like the negative binomial, it compounds a common discrete distribution (binomial) with its conjugate prior (Beta), resulting in a mathematically simple form that allows for more variability. The *lognormal-Poisson* is very similar to the negative binomial, except that (as its name suggests) it uses the lognormal instead of the Gamma as a compounding distribution. One technical reason to use the less common lognormal-Poisson is that on the log scale the rate parameter is normally distributed, which simplifies some numerical procedures (Elston et al., 2001).

Clark et al. (1999) used the *Student t* distribution to model seed dispersal curves. Seeds often disperse fairly uniformly near parental trees but also have a high probability of long dispersal. These two characteristics are incompatible with standard seed dispersal models like the exponential and normal distributions. Clark *et al.* assumed that the seed dispersal curve represents a compounding of a normal distribution for the dispersal of any one seed with an Gamma distribution of the inverse variance of the dis-

tribution of any particular seed (i.e., $1/\sigma^2 \sim$ Gamma)*. This variation in variance accounts for the different distances that different seeds may travel as a function of factors like their size, shape, height on the tree, and the wind speed at the time they are released. Clark *et al.* used compounding to model these factors as random, unobserved covariates since they are practically impossible to measure for all the individual seeds on a tree or in a forest.

The inverse Gamma-normal model is equivalent to the Student t distribution, which you may recognize from t tests in classical statistics and which statisticians sometimes use as a phenomenological model for fat-tailed distributions. Clark *et al.* extended the usual one-dimensional t distribution (`?dt` in R) to the two-dimensional distribution of seeds around a parent and called it the $2Dt$ distribution. The $2Dt$ distribution has a scale parameter that determines the mean dispersal distance and a shape parameter p. When p is large the underlying Gamma distribution has a small coefficient of variation and the $2Dt$ distribution is close to normal; when $p = 1$ the $2Dt$ becomes a Cauchy distribution.

Generalized distributions are an alternative class of mixture distribution that arises when there is a sampling distribution $S(x)$ for the number of individuals within a cluster and another sampling distribution $C(x)$ for number of clusters in a sampling unit. For example, the distribution of number of eggs per square might be generalized from the distribution of clutches per square and of eggs per clutch. A standard example is the "Poisson-Poisson" or "Neyman Type A" distribution (Pielou, 1977), which assumes a Poisson distribution of clusters with a Poisson distribution of individuals in each.

Figuring out the probability distribution or density formulas for compounded distributions analytically is mathematically challenging (see Bailey (1964) or Pielou (1977) for the gory details), but R can easily generate random numbers from these distributions (see the R supplement for more detail).

The key is that R's functions for generating random distributions (`rpois`, `rbinom`, etc.) can take vectors for their parameters. Rather than generate (say) 20 deviates from a binomial distribution with N trials and and a fixed per-trial probability p, you can choose 20 deviates with N trials and a vector of 20 different per-trial probabilities p_1 to p_20. Furthermore, you can generate this vector of parameters from another randomizing function! For example, to generate 20 beta-binomial deviates with $N = 10$ and the per-trial probabilities drawn from a beta distribution with $a = 2$ and

*This choice of a compounding distribution, which may seem arbitrary, turns out to be mathematically convenient.

$b = 1$, you could use `rbinom(20,rbeta(20,2,1))`.

Compounding and generalizing are powerful ways to extend the range of stochastic ecological models. A good fit to a compounded distribution also suggests that environmental variation is shaping the variation in the population. But be careful: Pielou (1977) demonstrates that for Poisson distributions, every generalized distribution (corresponding to variation in the underlying density) can also be generated by a compound distribution (corresponding to individuals occurring in clusters), and concludes that (p. 123) "the fitting of theoretical frequency distributions to observational data can never by itself suffice to 'explain' the pattern of a natural population".

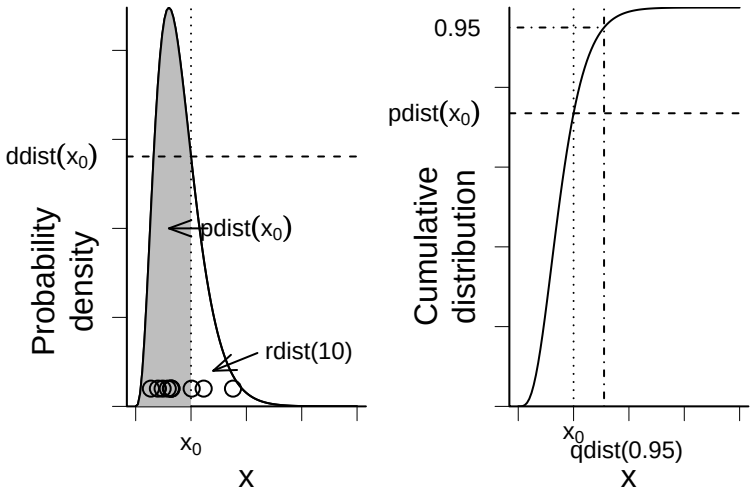

Figure 4.19 R functions for an arbitrary distribution `dist`, showing density function (`ddist`), cumulative distribution function (`pdist`), quantile function (`qdist`), and random-deviate function (`rdist`).

4.7 R SUPPLEMENT

For all of the probability distributions discussed in this chapter (and many more: try `help.search("distribution")`), R can generate random numbers drawn from the distribution ("deviates"); compute the cumulative distribution function and the probability distribution function; and compute the quantile function, which gives the x value such that $\int_0^x P(x)\,dx$ (area under the curve from 0 to x) is a specified value. For example, you can obtain the critical values of the standard normal distribution, ± 1.96, with `qnorm(0.025)` and `qnorm(0.975)` (Figure 4.19).

4.7.1 Discrete distribution

For example, let's explore the (discrete) negative binomial distribution.

First set the random-number seed for consistency:

```
> set.seed(1001)
```

Arbitrarily choose parameters $\mu = 10$ and $k = 0.9$ — since $k < 1$, this represents a strongly overdispersed population. Remember that R uses

`size` to denote k, because k is mathematically equivalent to the number of failures in the failure-process parameterization.

```
> z <- rnbinom(1000, mu = 10, size = 0.9)
```

Check the first few values:

```
> head(z)
```

```
[1] 41  3  3  0 11 14
```

Since the negative binomial has no set upper limit, we will just plot the results up to the maximum value sampled:

```
> maxz <- max(z)
```

The easiest way to plot the results is:

```
> f <- factor(z, levels = 0:maxz)
> plot(f)
```

using the `levels` specification to make sure that all values up to the maximum are included in the plot even when none were sampled in this particular experiment.

If we want the observed probabilities (freq/N) rather than the frequencies:

```
> obsprobs <- table(f)/1000
> plot(obsprobs)
```

Add theoretical values:

```
> tvals <- dnbinom(0:maxz, size = 0.9, mu = 10)
> points(0:maxz, tvals)
```

You could plot the deviations with `plot(0:maxz,obsprobs-tvals)`; this gives you some idea how the variability changes with the mean.

Find the probability that $x > 30$:

```
> pnbinom(30, size = 0.9, mu = 10, lower.tail = FALSE)

[1] 0.05725252
```

By default R's distribution functions will give you the *lower tail* of the distribution — the probability that x is less than or equal to some particular value. You could use `1-pnbinom(30,size=0.9,mu=10)` to get the uppper tail since $\text{Prob}(x > 30) = 1 - \text{Prob}(x \leq 30)$, but using `lower.tail=FALSE` to get the upper tail is more numerically accurate.

What is the upper 95th percentile of the distribution?

```
> qnbinom(0.95, size = 0.9, mu = 10)

[1] 32
```

To get the lower and upper 95% confidence limits, you need

```
> qnbinom(c(0.025, 0.975), size = 0.9, mu = 10)

[1]  0 40
```

You can also use the random sample `z` to check that the mean and variance, and 95th quantile of the sample, agree reasonably well with the theoretical expectations:

```
> mu <- 10
> k <- 0.9
> c(mu, mean(z))

[1] 10.000  9.654

> c(mu * (1 + mu/k), var(z))

[1] 121.1111 113.6539

> c(qnbinom(0.95, size = k, mu = mu), quantile(z, 0.95))

    95%
 32  31
```

4.7.2 Continuous distribution: lognormal

Going through the same exercise for the lognormal, a continuous distribution:

```
> z <- rlnorm(1000, meanlog = 2, sdlog = 1)
```

Plot the results:

```
> hist(z, breaks = 100, freq = FALSE)
> lines(density(z, from = 0), lwd = 2)
```

Add theoretical values:

```
> curve(dlnorm(x, meanlog = 2, sdlog = 1), add = TRUE,
+     lwd = 2, from = 0, col = "darkgray")
```

The probability of $x > 20$, 95% confidence limits:

```
> plnorm(30, meanlog = 2, sdlog = 1, lower.tail = FALSE)
```

```
[1] 0.08057753
```

```
> qlnorm(c(0.025, 0.975), meanlog = 2, sdlog = 1)
```

```
[1]  1.040848 52.455437
```

Comparing the theoretical values given on p. 182 with the observed values for this random sample:

```
> meanlog <- 2
> sdlog <- 1
> c(exp(meanlog + sdlog^2/2), mean(z))
```

```
[1] 12.18249 12.12708
```

Distribution	Type	Range	Skew	Examples
Binomial	Discrete	$0, N$	any	Number surviving, number killed
Poisson	Discrete	$0, \infty$	right $\rightarrow$ none	Seeds per quadrat, settlers (variance/mean ≈ 1)
Negative binomial	Discrete	$0, \infty$	right	Seeds per quadrat, settlers (variance/mean > 1)
Geometric	Discrete	$0, \infty$	right	Discrete lifetimes
Normal	Continuous	$-\infty, \infty$	none	Mass
Gamma	Continuous	$0, \infty$	right	Survival time, distance to nearest edge
Exponential	Continuous	$0, \infty$	right	Survival time, distance to nearest edge
Lognormal	Continuous	$0, \infty$	right	Size, mass (exponential growth)

Table 4.1 Summary of probability distributions

```
> c(exp(2 * meanlog + sdlog^2) * (exp(sdlog^2) - 1),
+     var(z))

[1] 255.0156 184.7721

> c(qlnorm(0.95, meanlog = meanlog, sdlog = sdlog),
+     quantile(z, 0.95))

             95%
38.27717 39.65172
```

There is a fairly large difference between the expected and observed variance. This is typical: variances of random samples have larger variances, or absolute differences from their theoretical expected values, than means of random samples.

Sometimes it's easier to deal with log-normal data by taking the logarithm of the data and comparing them to the normal distribution:

```
> hist(log(z), freq = FALSE, breaks = 100)
> curve(dnorm(x, mean = meanlog, sd = sdlog), add = TRUE,
+     lwd = 2)
```

4.7.3 Mixing and compounding distributions

4.7.3.1 Finite mixture distributions

The general recipe for generating samples from finite mixtures is to use a uniform distribution to sample which of the components of the mixture to sample, then use `ifelse` to pick values from one distribution or the other. To pick 1000 values from a mixture of normal distributions with the parameters shown in Figure 4.18 ($p = 0.3$, $\mu_1 = 1$, $\sigma_1 = 2$, $\mu_2 = 5$, $\sigma_2 = 1$):

```
> u1 <- runif(1000)
> z <- ifelse(u1 < 0.3, rnorm(1000, mean = 1, sd = 2),
+     rnorm(1000, mean = 5, sd = 1))
> hist(z, breaks = 100, freq = FALSE)
```

The probability density of a finite mixture composed of two distributions D_1 and D_2 in proportions p_1 and $1 - p_1$ is $p_1 D_1 + p_2 D_2$. We can superimpose the theoretical probability density for the finite mixture above on the histogram:

```
> curve(0.3 * dnorm(x, mean = 1, sd = 2) + 0.7 * dnorm(x,
+     mean = 5, sd = 1), add = TRUE, lwd = 2)
```

The general formula for the probability distribution of a zero-inflated distribution, with an underlying distribution $P(x)$ and a zero-inflation probability of p_z, is:

$$\begin{aligned} \text{Prob}(0) &= p_z + (1 - p_z)P(0) \\ \text{Prob}(x > 0) &= (1 - p_z)P(x) \end{aligned}$$

So, for example, we could define a probability distribution for a zero-inflated negative binomial as follows:

```
> dzinbinom = function(x, mu, size, zprob) {
+     ifelse(x == 0, zprob + (1 - zprob) * dnbinom(0,
+         mu = mu, size = size), (1 - zprob) * dnbinom(x,
+         mu = mu, size = size))
+ }
```

(the name, `dzinbinom`, follows the R convention for a probability distribution function: a `d` followed by the abbreviated name of the distribution, in this case `zinbinom` for "**z**ero-**i**nflated **n**egative **binom**ial").

The `ifelse` command checks every element of `x` to see whether it is zero or not and fills in the appropriate value depending on the answer.

Here's a random deviate generator:

```
> rzinbinom = function(n, mu, size, zprob) {
+     ifelse(runif(n) < zprob, 0, rnbinom(n, mu = mu,
+         size = size))
+ }
```

The command `runif(n)` picks `n` random values between 0 and 1; the `ifelse` command compares them with the value of `zprob`. If an individual value is less than `zprob` (which happens with probability `zprob`=p_z), then the corresponding random number is zero; otherwise it is a value picked out of the appropriate negative binomial distribution.

4.7.3.2 Compounded distributions

Start by confirming numerically that a negative binomial distribution is really a compounded Poisson-Gamma distribution. Pick 1000 values out of a Gamma distribution, then use those values as the λ (rate) parameters in a random draw from a Poisson distribution:

```
> k <- 3
> mu <- 10
> lambda <- rgamma(1000, shape = k, scale = mu/k)
> z <- rpois(1000, lambda)
> P1 <- table(factor(z, levels = 0:max(z)))/1000
> plot(P1)
> P2 <- dnbinom(0:max(z), mu = 10, size = 3)
> points(0:max(z), P2)
```

Establish that a Poisson-lognormal and a Poisson-Gamma (negative binomial) are not very different: pick the Poisson-lognormal with approximately the same mean and variance as the negative binomial just shown.

```
> mlog <- mean(log(lambda))
> sdlog <- sd(log(lambda))
> lambda2 <- rlnorm(1000, meanlog = mlog, sdlog = sdlog)
> z2 <- rpois(1000, lambda2)
```

```
> P3 <- table(factor(z2, levels = 0:max(z)))/1000
> matplot(0:max(z), cbind(P1, P3), pch = 1:2)
> lines(0:max(z), P2)
```