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 $\Delta x$ ($\text{Prob}(7.82 < X < 7.98)/0.16$, in this case). Dividing by $\Delta x$ allows the observed probability density to have a well-defined limit as precision increases and $\Delta x$ 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

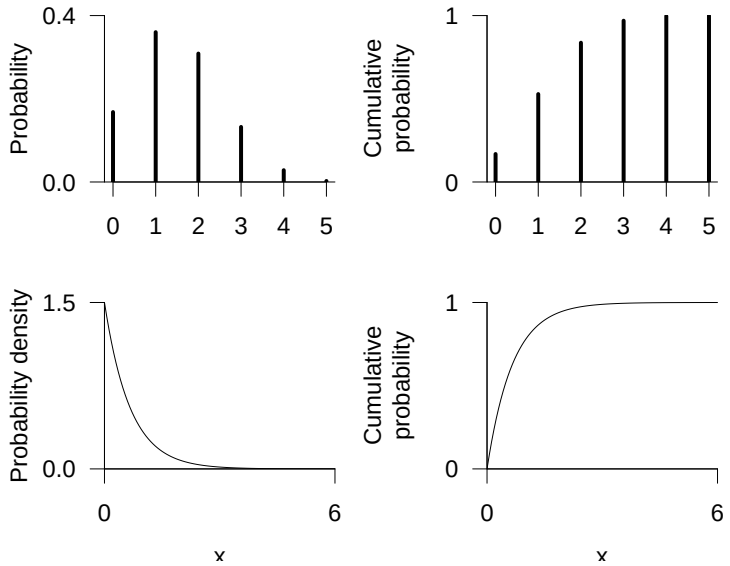


Figure 4.5 Probability, probability density, and cumulative distributions. Top: discrete (binomial: $N = 5$, $p = 0.3$) probability and cumulative probability distributions. Bottom: continuous (exponential: $\lambda = 1.5$) probability density and cumulative probability distributions.

$\{x_1, x_2, x_3, \ldots x_N\}$, 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 $\Delta x$. 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 $\Delta x$ 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 $n^{\text{th}}$ moment is $E[x^n]$, and the $n^{\text{th}}$ 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 $\mu$ (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:**

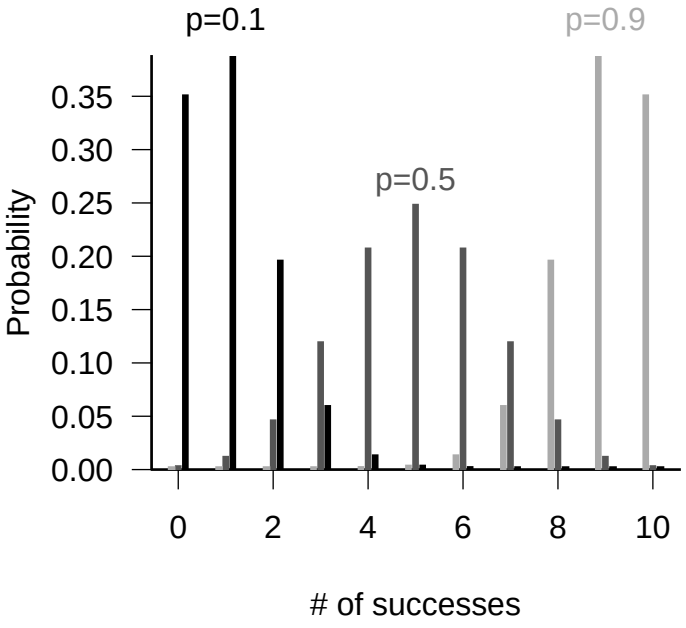


Figure 4.6 Binomial distribution. Number of trials ($N$) equals 10 for all distributions.

| | |
|---|---|
| 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, $\lambda$, 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

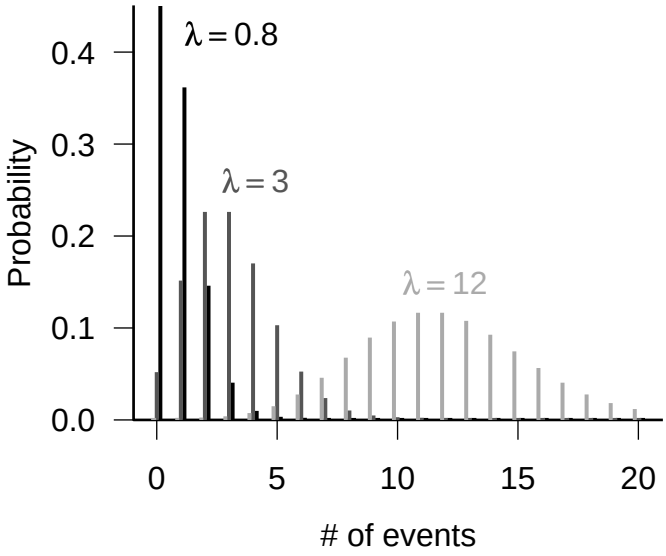


Figure 4.7 Poisson distribution.

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 | $\lambda$ (real, positive), expected number per sample [`lambda`] **or** $r$ (real, positive), expected number per unit effort, area, time, etc. (*arrival ra* |
| mean | $\lambda$ (**or** $rt$) |
| variance | $\lambda$ (**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 $\lambda$ itself varies. If the distribution of $\lambda$ is a gamma distribution (p. 172) with shape parameter $k$ and mean $\mu$, and $x$ is Poisson-distributed with mean $\lambda$, then the distribution of $x$ be a negative binomial distribution with mean $\mu$ 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) (!)

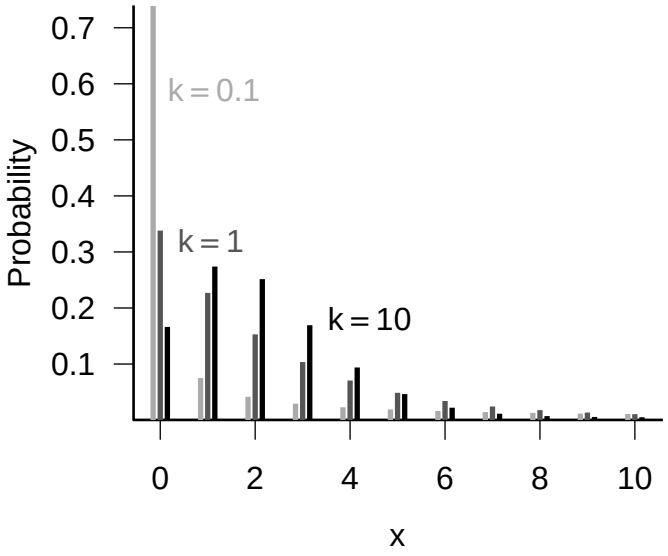


Figure 4.8 Negative binomial distribution. Mean $\mu = 2$ in all cases.

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** $\mu$ (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" ($\mu$ =`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 | $1/\sqrt{1/(1-p)}$ |

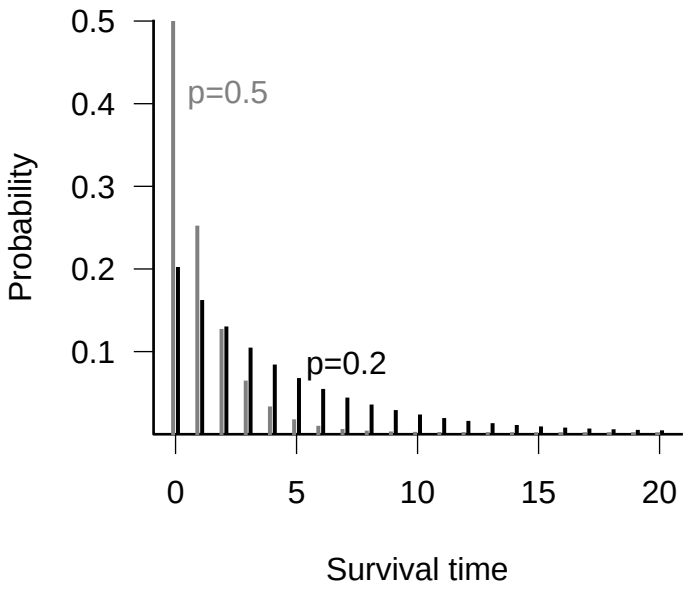


Figure 4.9 Geometric distribution.

#### 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 $\theta$, 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 $\theta$ $(= a + b)$. When $\theta$ is large (small overdispersion), the beta-binomial becomes binomial. When $\theta$ 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]<br>(`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`]<br>$\theta$ (real, positive), overdispersion parameter [`theta`]<br>**or** $a$ and $b$ (shape parameters of Beta distribution for per-trial probability) [`shape1` and `shape2`]<br>$a = \theta p$, $b = \theta(1 - p)$ |
| 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.

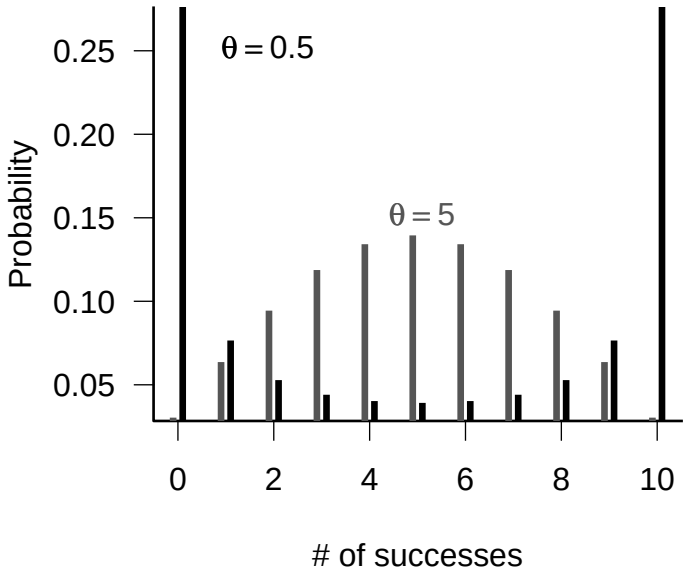


Figure 4.10 Beta-binomial distribution. Number of trials ($N$) equals 10, average per-trial probability ($p$) equals 0.5 for all distributions.

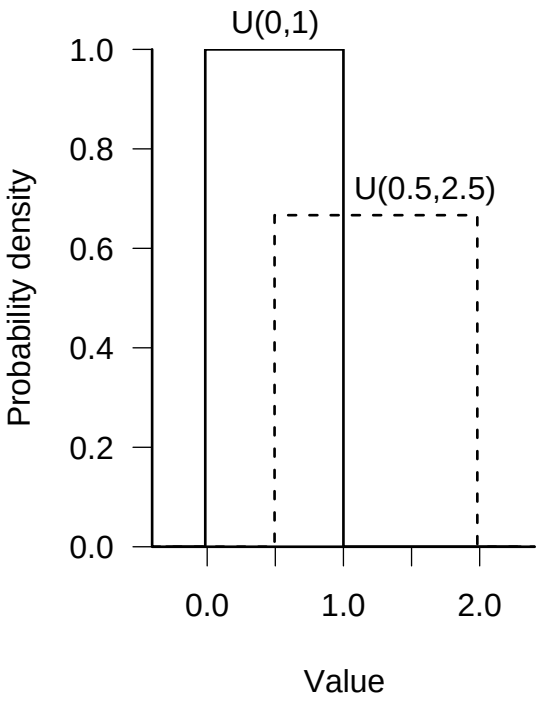


Figure 4.11 Uniform distribution.

### 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 | $\mu$ (real), mean [`mean`] |
| | $\sigma$ (real, positive), standard deviation [`sd`] |
| mean | $\mu$ |
| variance | $\sigma^2$ |
| CV | $\sigma/\mu$ |
| Conjugate prior | Normal ($\mu$); 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

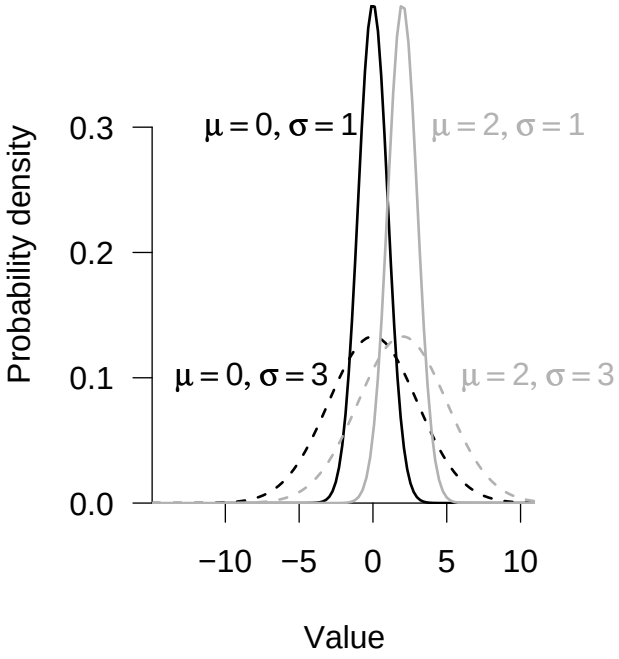


Figure 4.12 Normal distribution

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, $\Gamma$, 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 $\leq 1$, 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

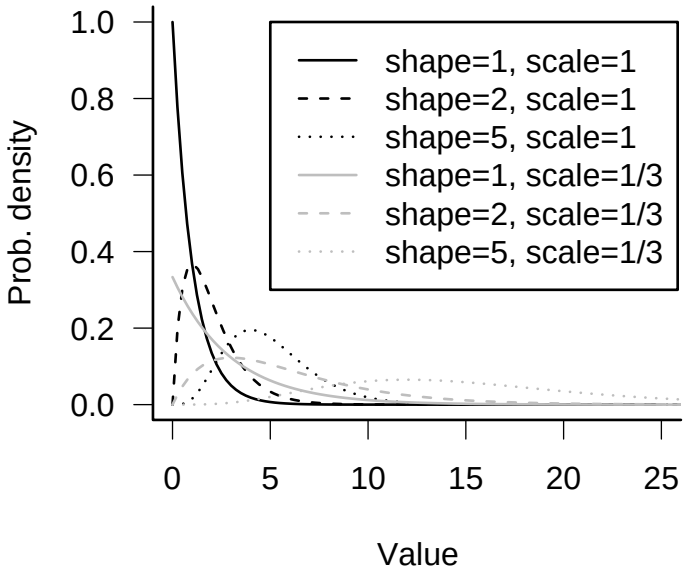


Figure 4.13 Gamma distribution

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 | $\lambda$ (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

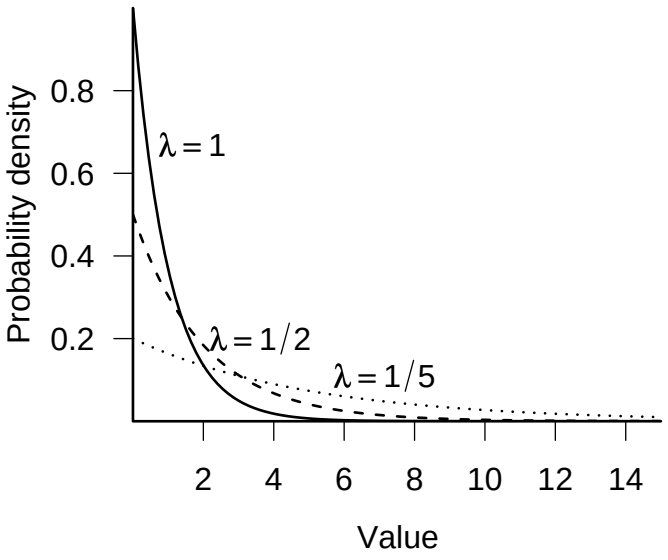


Figure 4.14 Exponential distribution.

(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-*

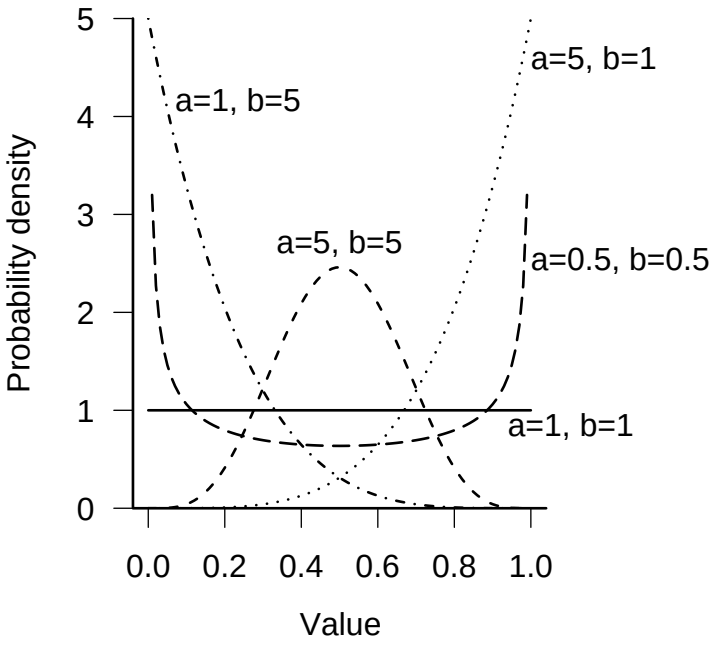


Figure 4.15 Beta distribution

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 $\mu$ (i.e. $e^{\mu}$). 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 $e^{\mu}$ 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 $e^{\mu}$, 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).

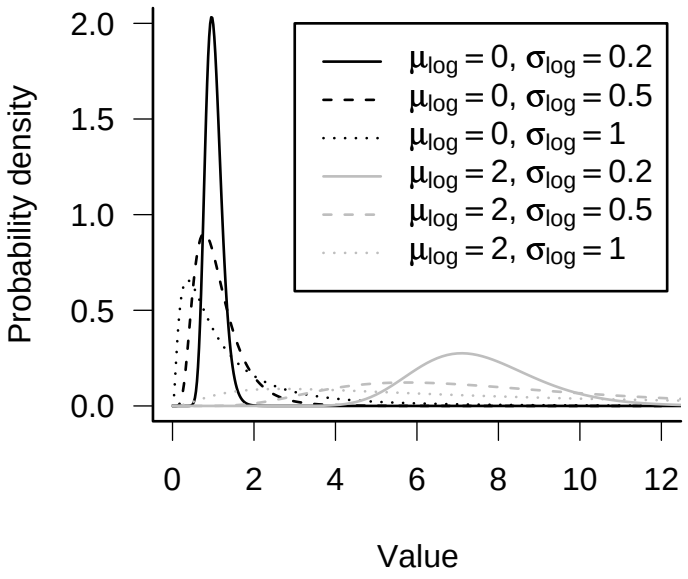


Figure 4.16 Lognormal distribution

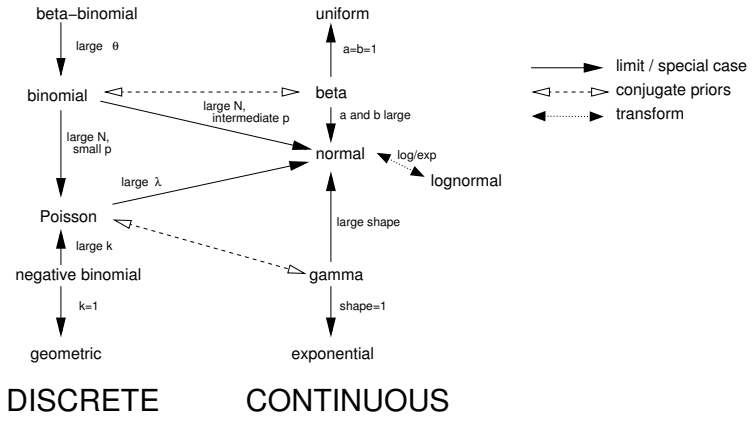


Figure 4.17 Relationships among probability distributions.

| | |
|---|---|
| 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 | $\mu$ (real): mean of the logarithm [`meanlog`] |
| | $\sigma$ (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.