Reading notes of *Ecological Models and Data in R*

Daijiang Li

April 2, 2014

Chapter 1

Introduction

1.1 Frameworks for statistical inference

1.1.1 Classical frequentist

Classical statistics, which are part of the broader frequentist paradigm, are the kind of statistics typically presented in introductory statistics classes. For a specific experimental procedure (such as drawing cards or flipping coins), you calculate the probability of a particular outcome, which is defined as *the long-run average frequency of that outcome in a sequence of repeated experiments*. Next you calculate a p-value, defined as the probability of that outcome or any more extreme outcome given a specified null hypothesis. If this so-called tail probability is small, then you reject the null hypothesis: otherwise, you fail to reject it. But you don't accept the alternative hypothesis if the tail probability is large, you just fail to reject the null hypothesis.

A frequentist would translate this biological question into statistics as "what is the probability that I would observe a result this extreme, or more extreme, given the sampling procedure?"

Working statisticians will tell you that it is better to focus on estimating the values of biologically meaningful parameters and finding their confidence limits rather than worrying too much about whether p is greater or less than 0.05 (Yoccoz, 1991; Johnson, 1999; Osenberg et al., 2002) — although Stephens et al. (2005) remind us that hypothesis testing can still be useful.

1.1.2 Likelihood

Most modern statistics uses an approach called maximum likelihood estimation, or approximations to it. For a particular statistical model, maximum likelihood finds the set of parameters (e.g. seed removal rates) that makes the observed data (e.g. the particular outcomes of predation trials) most likely to have occurred. Based on a model for both the deterministic and stochastic aspects of the data, we can compute the likelihood (the probability of the observed outcome) given a particular choice of parameters. We then find the set of parameters that makes the likelihood as large as possible, and take the resulting maximum likelihood estimates (MLEs) as our best guess at the parameters.

For mathematical convenience, we often work with the logarithm of the likelihood (the log-likelihood) instead of the likelihood; the parameters that give the maximum log-likelihood also give the maximum likelihood.

However, most modelers add a frequentist interpretation to likelihoods, using a mathematical proof that says that, across the hypothetical repeated trials of the frequentist approach, the distribution of the negative logarithm of the likelihood itself follows a χ^2 ("chi-squared") distribution.

Likelihood and classical frequentist analysis share the same philosophical underpinnings. Likelihood analysis is really a particular flavor of frequentist analysis, one that focuses on writing down a likelihood model and then testing for significant differences in the likelihood ratio rather than applying frequentist statistics directly to the observed outcomes. Classical analyses are usually easier because they are built into common statistics packages, and they may make fewer assumptions than likelihood analyses (for example, Fisher's test is exact while the LRT is only valid for large data sets), but likelihood analyses are often better matched with ecological questions.

1.1.3 Bayesian

Frequentist statistics assumes that there is a "true" state of the world (e.g. the difference between species in predation probability) which gives rise to a distribution of possible experimental outcomes. The Bayesian framework says instead that the experimental outcome — what we actually saw happen — is the truth, while the parameter values or hypotheses have probability distributions. The Bayesian framework solves many of the conceptual problems of frequentist statistics: answers depend on what we actually saw and not on a range of hypothetical outcomes, and we can legitimately make statements about the probability of different hypotheses or parameter values.

Frequentists, who believe that the true value is a fixed number and uncertainty lies in what you observe [or might have observed], and Bayesians, who believe that observations are fixed numbers and the true values are uncertain.

Chapter 2

Explorayory data analysis and graphics

Chapter 3

Deterministic functions for ecological modeling

3.1 Introduction

What functions could fit this pattern? What do their parameters mean in terms of the shapes of the curves? In terms of ecology? How do we "eyeball" the data to obtain approximate parameter values, which we will need as a starting point for more precise estimation and as a check on our results?

The Ricker function, $y = axe^{-bx}$, is a standard choice for hump-shaped ecological patterns that are skewed to the right.

3.2 Finding out about functions numerically

Calculating and plotting curves in R.

3.3 Finding out about functions analytically

3.3.1 Taking limits: what happens at either end?

Terms with larger powers of x will dwarf smaller powers, and exponentials will dwarf any power. Exponentials are stronger than powers: $x^{-n}e^x$ eventually gets big and $x^n e^{-x}$ eventually gets small as x increases, no matter how big n is (exponentials always win). For more difficult functions that contain a fraction whose numerator and denominator both approach zero or infinity in some limit (and thus make it hard to find the limiting value), you can try L'Hopital's Rule: $\lim \frac{a(x)}{b(x)} = \lim \frac{a'(x)}{b'(x)}$.

3.3.2 Taylor series approximation

The Taylor series or Taylor approximation is the single most useful, and used, application of calculus for an ecologist. Two particularly useful applications of Taylor approxima-

Table 3.1: Useful non-linear functions

Name		Equation
Asymptotic functions		
	Michaelis-Menten	$y = \frac{ax}{1+bx}$
	2-parameter asymptotic exponential	$y = a(1 - e^{-bx})$
	3-parameter asymptotic exponential	$y = a - be^{-cx}$
S-shaped functions		
	logistic 2-parameter	$y = \frac{e^{a+bx}}{1+e^{a+bx}}$
	logistic 3-parameter	$y = \frac{a}{1+be^{-cx}}$
	logistic 4-parameter	$y = a + \frac{b-a}{1+e^{(c-x)/d}}$
	Weibull	$y = a - be^{-(cx^d)}$
	Gompertz	$y = ae^{-be^{-cx}}$
Humped curves		
	Ricker curve	$y = axe^{-bx}$
	First-order compartment	$y = ke^{(-e_a x - e^{-e_b x})}$
	Bell-shaped	$y = ae^{-\lvert bx\rvert^2}$
	Biexponential	$y = ae^{bx} - ce^{-dx}$
Other functions		
	Negative exponential	$y = ae^{-bx}$
	Polynomial functions	$y = \sum_{i=0}^{n} a_i x^i$
Rational functions:polynomials in fractions		$y = (\sum a_i x^i) / (\sum b_j x^j)$
	Hyperbolic	$y = a/x$
	Michaelis-Menten (Holling type II)	$y = \frac{ax}{b+x}$
	Holling type III	$y = \frac{ax^2}{b^2+x^2}$
	Holling type IV	$y = \frac{ax^2}{b+cx+x^2}$

tion are understanding the shapes of goodnessof-fit surfaces (Chapter 6) and the delta method for estimating errors in estimation (Chapter 7). The Taylor series allows us to approximate a complicated function near a point we care about, using a simple function — a polynomial with a few terms, say a line or a quadratic curve. In practice ecologists never go beyond a quadratic expansion.

$$f(x) \approx f(x_0) + f'(x_0) \cdot (x - x_0) + f''(x_0) \cdot \frac{(x - x_0)^2}{2} + \cdots + f^n(x_0) \cdot \frac{(x - x_0)^n}{n!}$$

The Taylor expansion of the exponential, e^{rx} , around $x = 0$ is $1 + rx + (rx)^2/2 + (rx)^3/(2 \cdot 3)$.... Remembering this fact rather than working it out every time may save you time in the long run — for example, to understand how the Ricker function works for small x we can substitute $(1 - bx)$ for e^{-bx} (dropping all but the first two terms!) to get $y \approx ax - abx^2$: this tells us immediately that the function starts out linear, but starts to curve downward right away.

Calculating Taylor approximations is often tedious (all those derivatives), but we usually try to do it at some special point where a lot of the complexity goes away (such as $x = 0$ for a logistic curve).

Chapter 4

Probability and Stochastic Distributions for Ecological Modeling

Probability density function: 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 ($F(x) = Prob(X \leq x)$ 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 ± 1%. Thus continuous probability distributions are expressed as probability densities rather than probabilities — the probability that random variable X is between x and x + Δx, divided by Δx (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.

In probability theory, a ***probability density function (pdf)***, or density of a *continuous random variable*, is a function that describes the *relative likelihood* for this random variable to take on a given value. The probability of the random variable falling within a particular range of values is given by the integral of this variable's density over that range—that is, it is given by the area under the density function but above the horizontal axis and between the lowest and greatest values of the range. The probability density function is nonnegative everywhere, and its integral over the entire space is equal to one.

A ***probability density function*** is most commonly associated with absolutely continuous univariate distributions. A random variable X has ***density*** f_X, where f_X is a non-

negative *Lebesgue-integrable* function, if:

$$Pr[a \leq X \leq b] = \int_a^b f_X(x)dx$$

Hence, if F_X is the ***cumulative distribution function*** of X, then:

$$F_X(x) = \int_{-\infty}^{x} f_X(\mu)d\mu$$

and (if f_X is continuous at x)

$$f_X(x) = \frac{d}{dx}F_X(x)$$

Intuitively, one can think of $f_X(x)dx$ as being the probability of X falling within the infinitesimal interval $[x,\ x + dx]$.

Definition. Let X be a discrete random variable. Then for $x \in \mathbb{R}$, the function $p_X(x) = P\{X = x\}$ is called the *probability mass function* of X . By the axioms of probability, a probability mass function p_X satises $P\{X \in A\} = \sum_{x \in A} p_X(x)$.

Definition. Let X be a continuous random variable with distribution function $F(t) = P\{X \leq t\}$. Suppose that there exists a nonnegative, integrable function $f : \mathbb{R} \to [0, \infty)$, or sometimes f_X , such that

$$F(x) = \int_{-\infty}^{x} f(y)dy$$

Then the function f is called the *probability density function* of X . We now have that for any $A \subset \mathbb{R}$ (or, more precisely, for any $A \in \digamma$),

$$P\{X \in A\} = \int_A f_X(x)dx$$

4.1 Bestiary of Distributions

4.1.1 Discrete models

4.1.1.1 Binomial

$f(x) = \binom{N}{x} p^x(1-p)^{N-x}$, mean= Np, variance = $Np(1-p)$, depends on the number of samples per trail N. N increases, variance increases but the coefficient of variation (cv) $\sqrt{Np(1-p)}/(Np) = \sqrt{(1-p)/(Np)}$ decreases.

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 isn't too close to 0 or 1 (i.e. when Np is large, typically $Np \geq 10$, $Nq \geq 10$), then the binomial distribution is approximately **normal**. When N is large and p is small, so that the probability of getting N successes is small, the binomial approaches the **Poisson** distribution.

Table 4.1: Summary of probability distributions I.

Name	distribution	parameters	mean	variance	CV	conjugate prior
Binomial	$\binom{N}{x} p^x (1-p)^{N-x}$	p [real, 0-1], prob of success, N [positive integer], number of trials	Np	$Np(1-p)$	$\sqrt{(1-p)/(Np)}$	Beta
Poisson	$\frac{e^{-\lambda}\lambda^n}{n!}$ or $\frac{e^{-rt}(rt)^n}{n!}$	λ (real, positive) expected number per sample, or r, expected # per unit effort, area, time, etc.	λ	λ	$1/\sqrt{\lambda}$	Gamma
Negative binomial	$\frac{(n+x-1)!}{(n-1)!x!} p^n (1-p)^x$ or $\frac{\Gamma(k+x)}{\Gamma(k)x!}(\frac{k}{k+\mu})^k(\frac{\mu}{k+\mu})^x$	p probability per trial and n # of successes awaited; OR μ expected # of counts and k overdispersion parameter (shape parameter of underlying heterogeneity)	μ; nq/p	$\mu + \frac{\mu^2}{k}$;nq/p^2	$\sqrt{(1+\mu/k)/\mu}$; $1/\sqrt{nq}$	No simple
Geometric	$p(1-p)^x$	p probability of success	$\frac{1-p}{p}$	$\frac{1-p}{p^2}$	$1/\sqrt{1/(1-p)}$	
Beta Binomial	$\frac{\Gamma(\theta)}{\Gamma(p\theta)\Gamma(q\theta)} \frac{N!}{x!(N-x)!} \frac{\Gamma(x+p\theta)\Gamma(N-x+q\theta)}{\Gamma(N+\theta)}$	p, θ overdispersion parameter. Or a=θp and b=θq shape parameters of Beta distribution.	Np	$Npq(1+\frac{N-1}{\theta+1})$	$\sqrt{\frac{q}{Np}(1+\frac{N-1}{\theta+1})}$	
Uniform	$\frac{1}{b-a}$	a, b min and max	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	$\frac{b-a}{(a+b)\sqrt{3}}$	
Normal	$\frac{1}{\sqrt{2\pi\sigma}} \exp(-\frac{(x-\mu)^2}{2\sigma^2})$	μ, σ	μ	σ^2	$\frac{\sigma}{\mu}$	Normal(μ); Gamma($\frac{1}{\sigma^2}$)
Gamma	$\frac{1}{s^a\Gamma(a)} x^{a-1} e^{-x/s}$	$a > 0$ shape: # of events; $s = \frac{mean}{a} > 0$ scale: length per event or r rate$= 1/s$ rate at which events occur.	as	as^2	$\frac{1}{\sqrt{a}}$	
Exponential	$\lambda e^{-\lambda x}$	$\lambda > 0$ rate: death or disappearance rate	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	1	
Beta	$\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}$	$a > 0$ shape1: # of successes+1; $b > 0$ shape2: # of failures +1.	$\frac{a}{a+b}$	$\frac{ab}{(a+b)^2(a+b+1)}$	$\sqrt{\frac{b/a}{a+b+1}}$	
Lognormal	$\frac{1}{x\sigma\sqrt{2\pi}} \exp(-\frac{(log(x)-\mu)^2}{2\sigma^2})$	μ, σ	$e^{(\mu+\frac{\sigma^2}{2})}$	$e^{(2\mu+\sigma^2)}(e^{\sigma^2}-1)$	$\sqrt{e^{\sigma^2}-1}$	

Table 4.2: Summary of probability distributions II.

Name	Type	range	Skew	Examples
Binomial	Discrete	$0 \leq x \leq N$	any	# of surviving individuals out of an initial sample; # of infested/infected animals, fruits, etc. in a sample; # of a particular class (haplotype, subspecies, etc.) in a larger population.
Poisson	Discrete	$x \geq 0$	right$\rightarrow$none	# of seeds/seedlings falling in a gap; # of offsprings produced in a season; # of prey caught per unit time.
Negative binomial	Discrete	$x \geq 0$	right	essentially the same as the Poisson distribution, but allowing for heterogeneity. *Numerically it is really a compunded Poisson-Gamma distribution*. # of individuals per patch; # of seedlings in a gap or per unit area.
Geometric (memoryless)	Discrete	$x \geq 0$	right	# of successful/survived breeding seasons for a seasonally reproducing organism. Lifespans measured in discrete units.
Beta Binomial	Discrete	$0 \leq x \leq N$	any	as for the Binomial
Uniform	Continuous	$a \leq x \leq b$	none	frequently used as a building block for other distributions.
Normal	Continuous	$-\infty, \infty$	none	many continuous symmetrically distributed measurements – temperature, pH, nitrogen concentration.
Gamma	Continuous	$x > 0$	right	almost any environmental variable with a large variance where negative values don't make sense: nitrogen concentrations, light intensity, growth rates etc..
Exponential (memoryless)	Continuous	$x > 0$	right	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).
Beta	Continuous	$0 \leq x \leq 1$	any	good for modeling probabilities or proportions; modeling continuous distributions with peaks at both ends; define a continuous distribution on a finite range.
Lognormal	Continuous	$x > 0$	right	sizes or masses of individuals, especially rapidly growing individuals; abundance vs. frequency curves for plant communities.

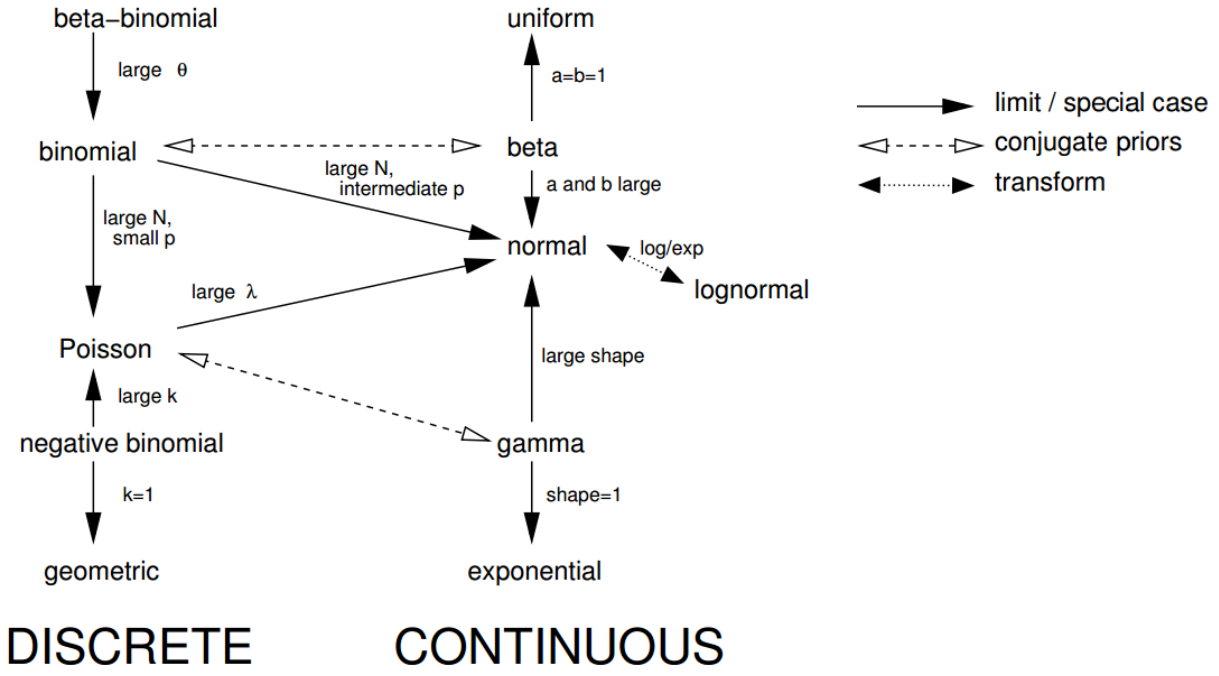


Figure 4.1: Relationships among probability distributions.

4.1.1.2 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 Poisson distribution *only makes sense for count data*. For $\lambda < 1$ the Poisson's mode is at zero. When the expected number of counts gets large (e.g. $\lambda > 10$) the Poisson becomes approximately normal.

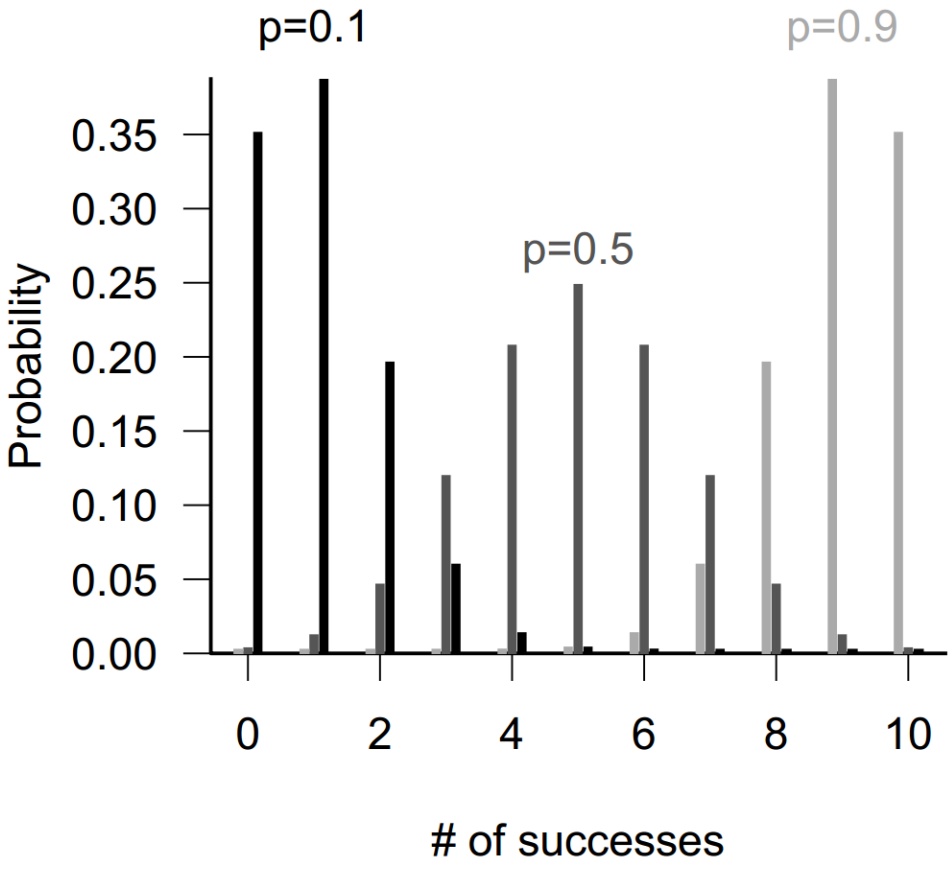


Figure 4.2: Binomial distribution. $N = 10$.

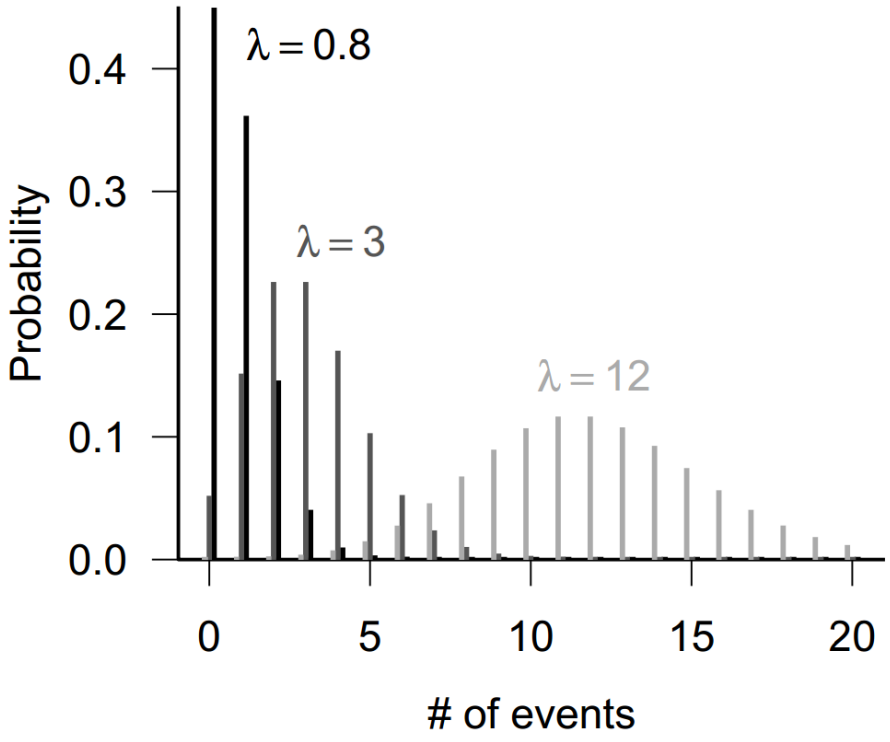


Figure 4.3: Poisson distribution.

4.1.1.3 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*. (i.e. $p \to \mu$, $n \to k$). 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 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.

```
lambda <- rgamma(1000, shape = k, scale = mu/k)  # scale=mean/shape
z <- rpois(1000, lambda)
p2 <- dnbinom(0:max(z), mu = mu, size = k)
# z = p2 !
```

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.

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)`).

4.1.1.4 Geometric

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

4.1.1.5 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

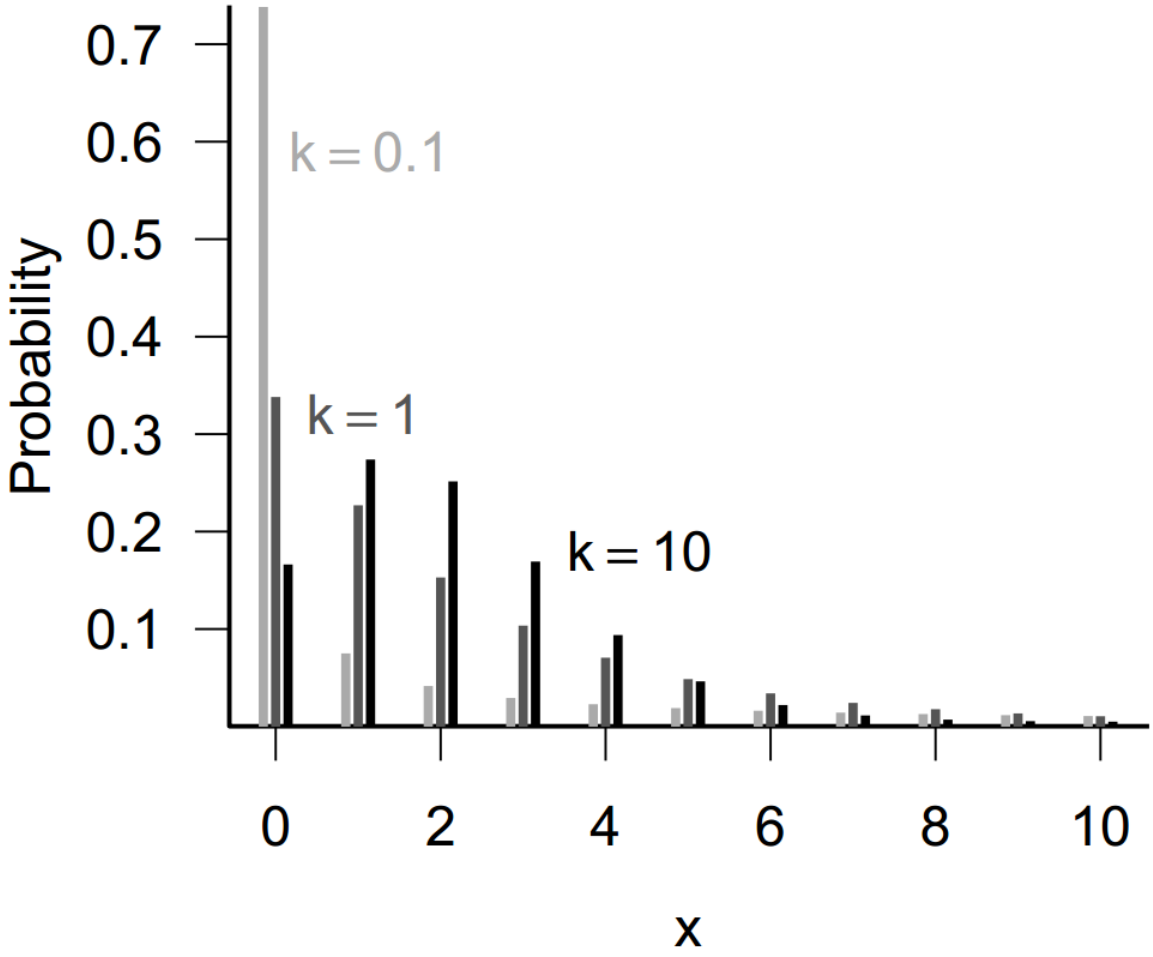


Figure 4.4: Negative binomial distribution. Mean $\mu = 2$ in all cases.

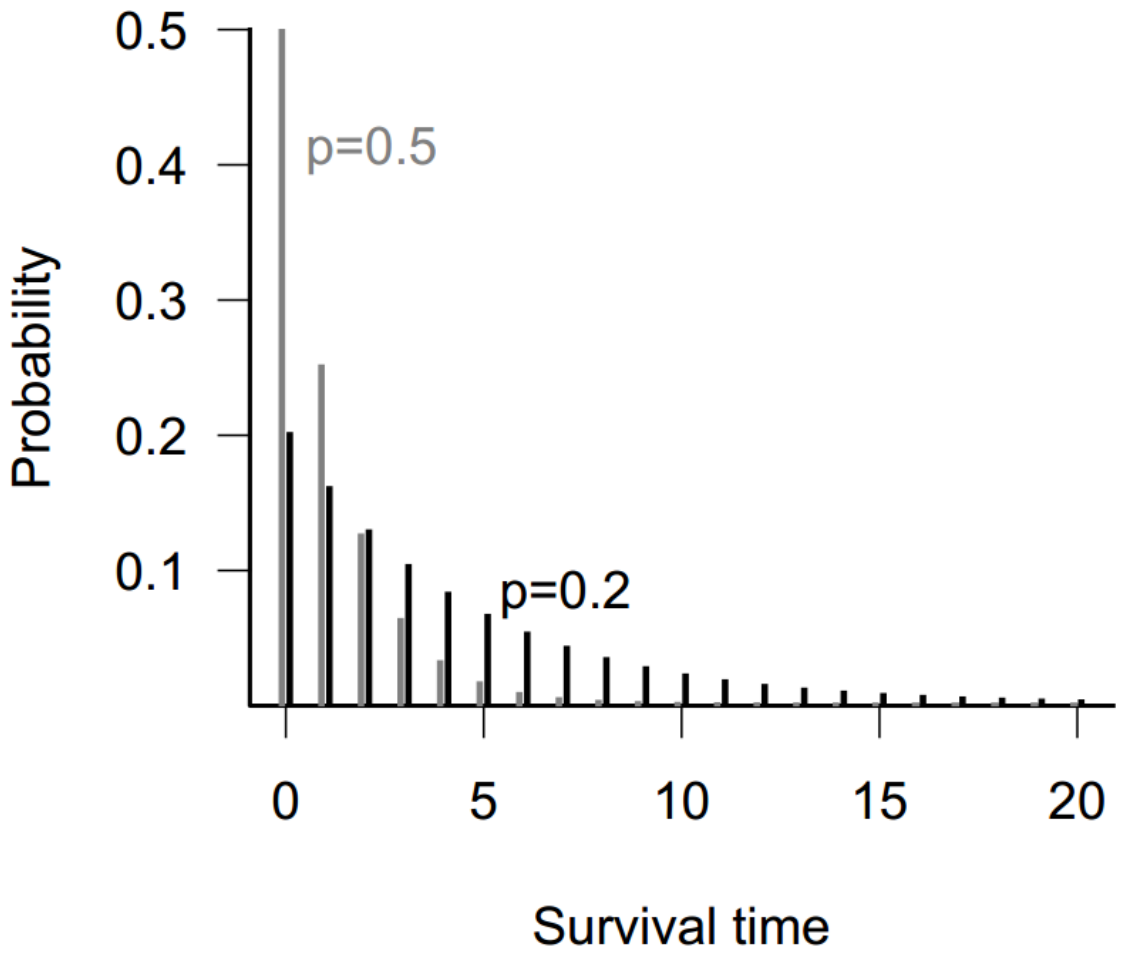


Figure 4.5: Geometric distribution.

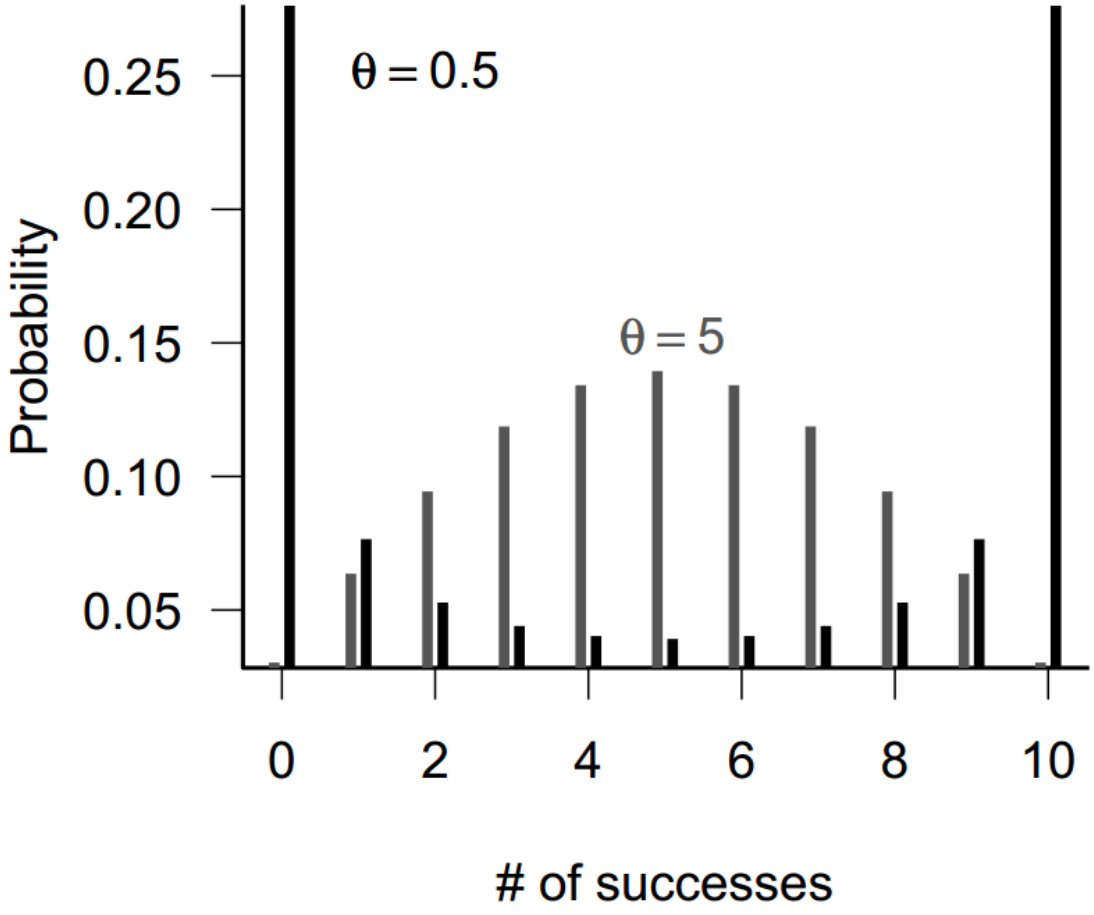


Figure 4.6: Beta-binomial distribution. Number of trials (N) equals 10, average per-trial probability (p) equals 0.5 for all distributions.

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 $\theta = a + b$. When θ is large (small overdispersion), the beta-binomial becomes binomial. When θ is near zero (large overdispersion), the beta-binomial becomes U-shaped.

4.1.2 Continuous Distributions

4.1.2.1 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.

4.1.2.2 Normal distribution

4.1.2.3 Gamma distribution

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

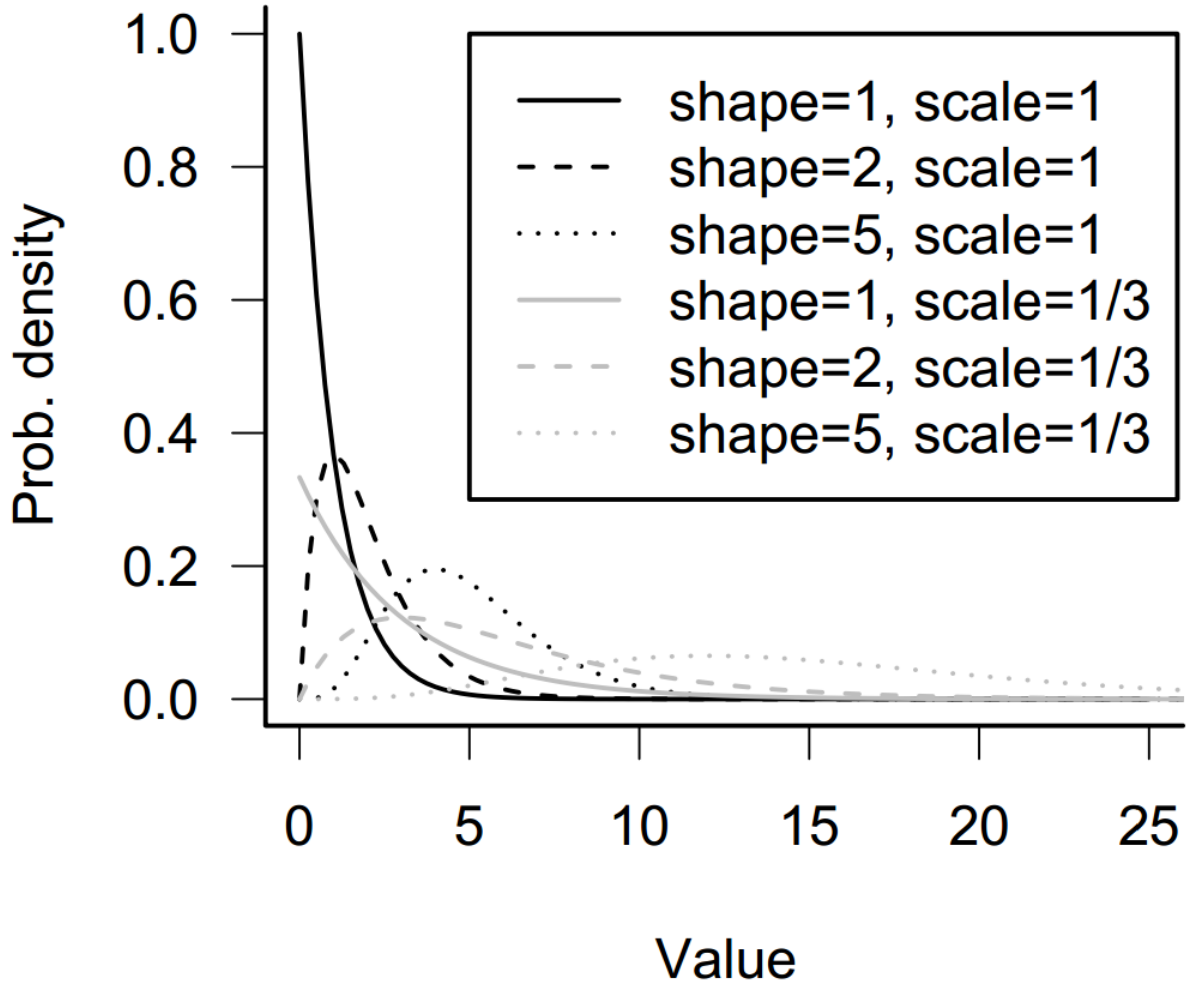


Figure 4.7: Gamma distribution

anistic descriptions (e.g. a Gamma distribution with a shape parameter of 2.3 corresponds to 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.**

4.1.2.4 Exponential

The exponential distribution 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

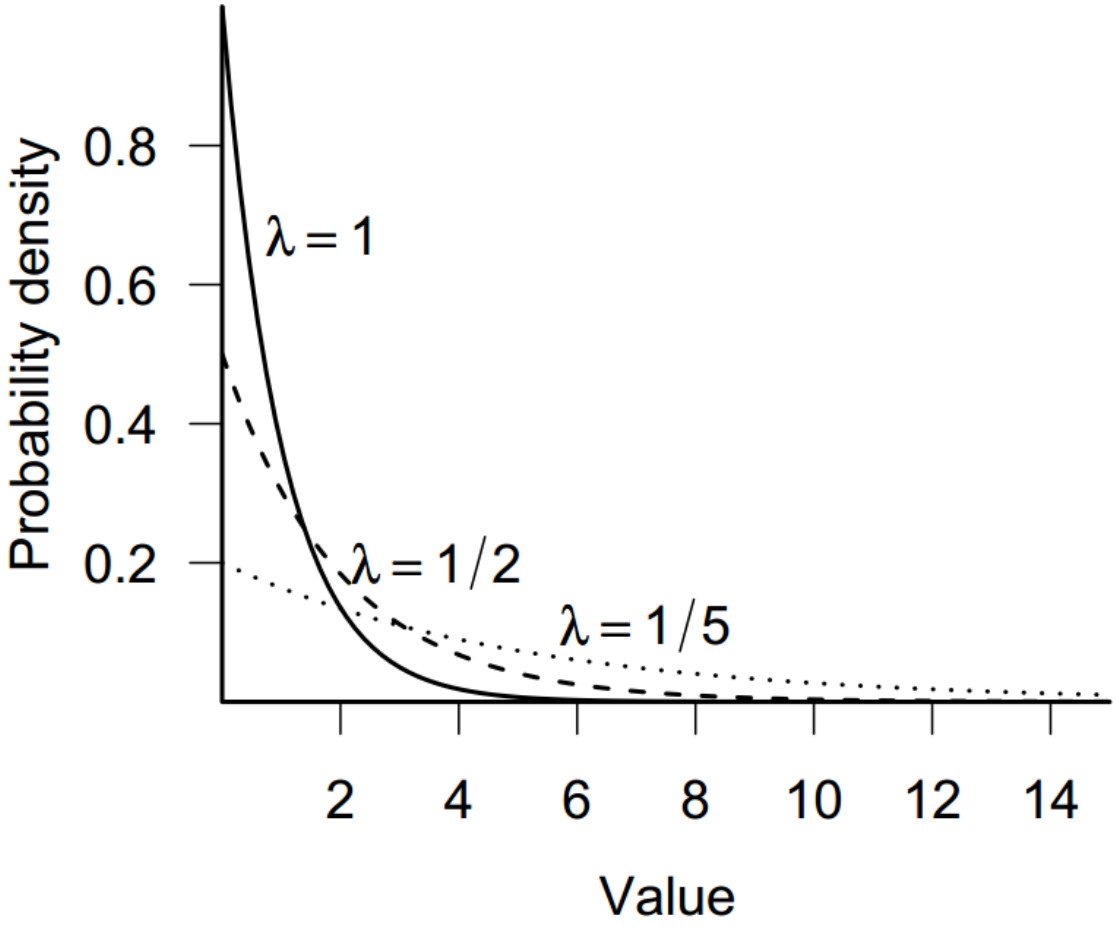


Figure 4.8: Exponential distribution.

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.

4.1.2.5 Beta

The beta distribution is a continuous distribution closely related to the binomial distribution. *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

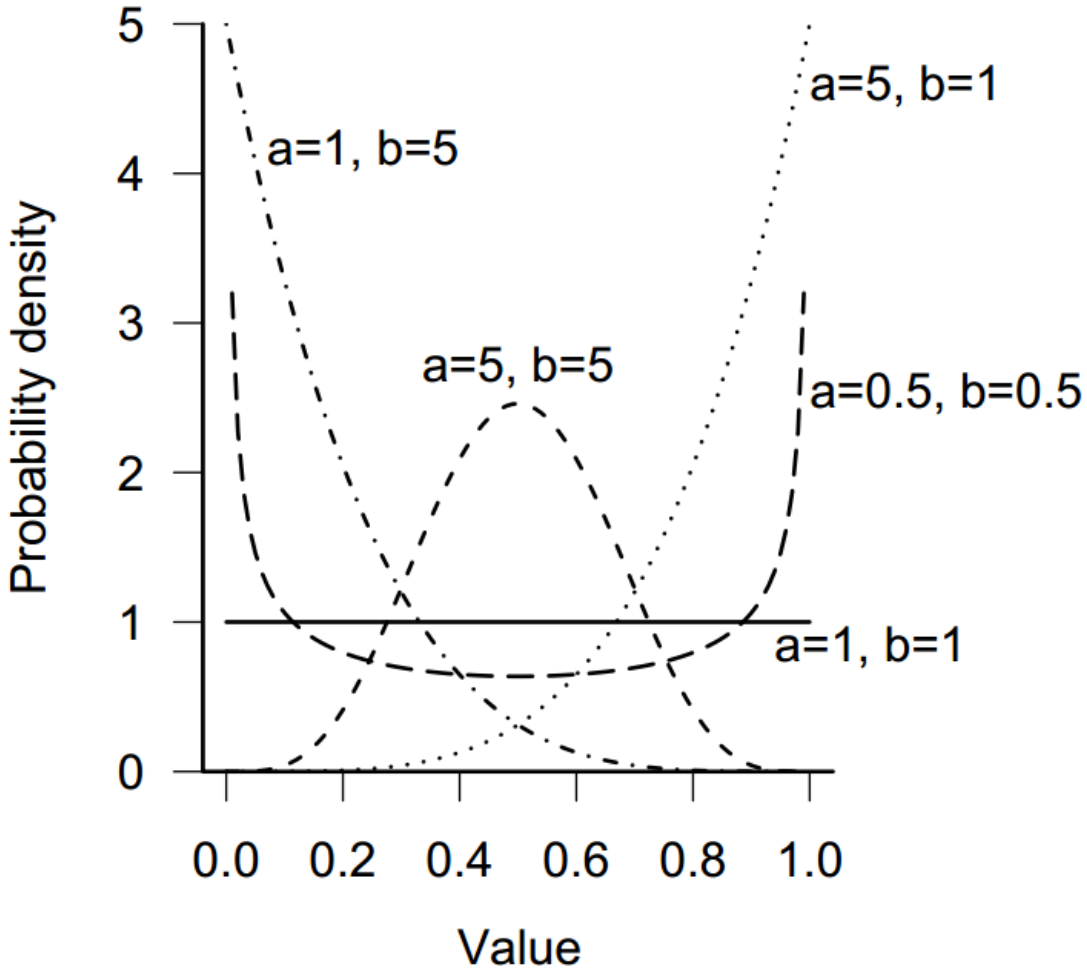


Figure 4.9: Beta distribution

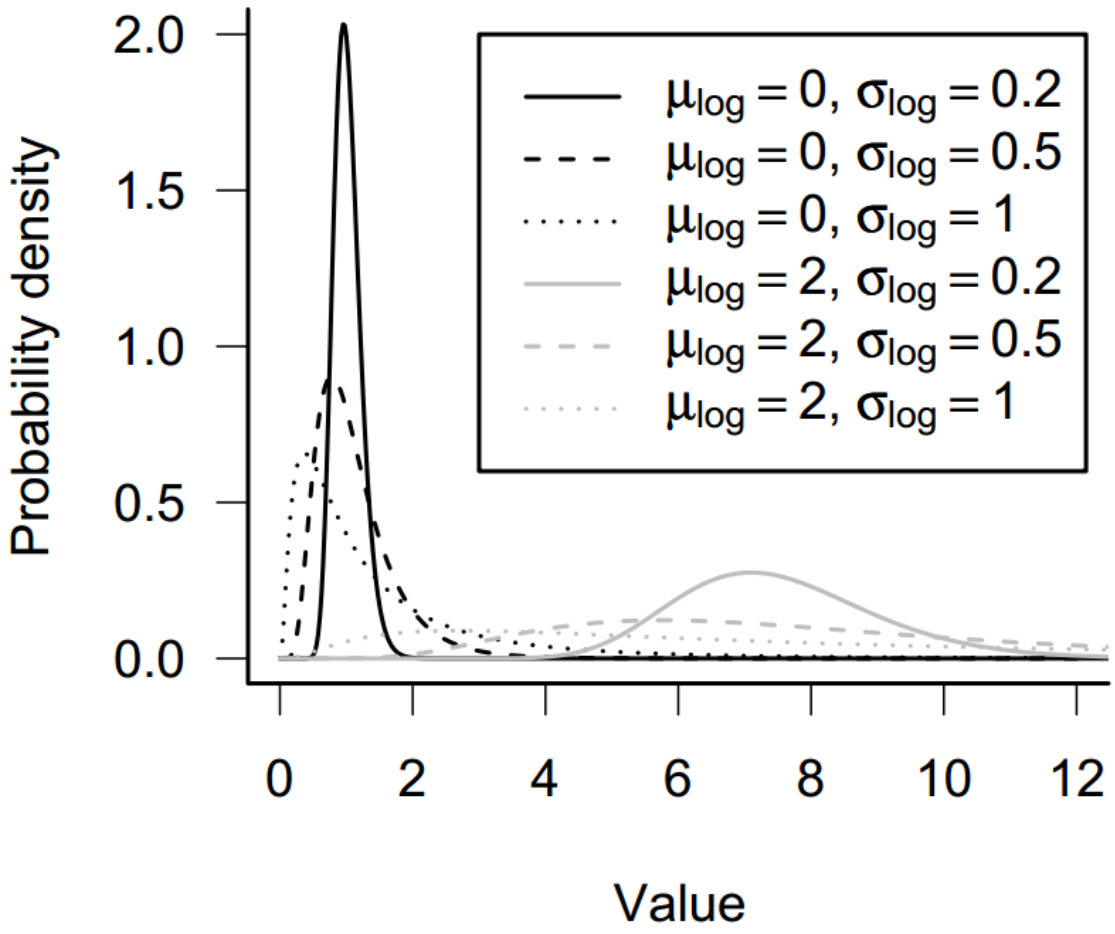
well-defined, and when a and b are both between 0 and 1 it becomes U-shaped — it has peaks at $p = 0$ and $p = 1$.

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 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.

4.1.2.6 Lognormal

Its mechanistic justification is like the normal 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.

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.



a=20;b=1;k=5

Figure 4.10: Lognormal distribution

Chapter 5

Stochastic Simulation and Power Analysis

Simulation is sometimes called *forward* modeling, to emphasize that you pick a model and parameters and work forward to predict patterns in the data. Parameter estimation, or *inverse* modeling (the main focus of this book), starts from the data and works backward to choose a model and estimate parameters.

5.1 Stochastic Simulation

5.1.1 Simple examples

$Y \sim Normal(a + bx, \sigma^2)$, or can also be written as $y_i = a + bx_i + \varepsilon_i$, $\varepsilon_i \sim N(0, \sigma^2)$, specifying that the i th value of Y , y_i , is equal to $a + bx_i$ plus a normally distributed error term with mean zero. But the first form is more general. Normally distributed error is one of the few kinds that can simply be added onto the deterministic model in the second way.

```
x <- 1:20
a <- 2
b <- 1
y_det <- a + b * x
y <- rnorm(20, mean = y_det, sd = 2)
```

A simple simulation uses hyperbolic functions ($y = ab/(b + x)$) with negative binomial error: in symbols, $Y \sim NegBin(\mu = ab/(b + x), k)$. The function is parameterized so that a is the intercept term (when $x = 0, y = ab/b = a$). *This simulation might represent the decreasing fecundity of two different species with increasing population density; the hyperbolic function is a natural expression of the decreasing quantity of a limiting resource per individual.* In this case, we cannot express the model as the deterministic function "plus error". Instead, we have to incorporate the deterministic model as a control on one of the parameters of the error distribution—in this case, the mean μ. (Although the negative binomial is a discrete distribution, its parameters μ and k are continuous.)

```r
a <- 20
b <- 1
k <- 5
x <- runif(50, min = 0, max = 5)
y_det <- a * b/(b + x)
y <- rnbinom(50, mu = y_det, size = k)
```

Chapter 6

Likelihood and all that

Previous chapters have introduced all the ingredients you need to define a model — mathematical functions to describe the deterministic patterns and probability distributions to describe the stochastic patterns — and shown how to use these ingredients to simulate simple ecological systems. However, you need to learn not only how to construct models but also how to estimate parameters from data, and how to test models against each other.

In general, to estimate the parameters of a model we have to find the parameters that make that model fit the data best. To compare among models we have to figure out which one fits the data best, and decide if one or more models fit sufficiently much better than the rest that we can declare them the winners. Our goodness-of-fit metrics will be based on the ***likelihood, the probability of seeing the data we actually collected given a particular model*** — which in this case will mean both the general form of the model and the specific parameter values.

6.1 Parameter estimation: Single distributions

6.1.1 Maximum likelihood

We want the *maximum likelihood estimates* of the parameters — those parameter values that make the observed data most likely to have happened. *i.i.d* data, so the joint likelohood of the wholde data set is the product of the likelihood of each individual observation. For mathmatical convience, we take the logarithm of the likelihood. Since the logarithm is a monotonically increasing function, the maximum log-likelihood estimate is the same as the maximum likelihood estimate. Actually, it is conventional to ***minimize*** the *negative log-likelihood* rather than maximizing the log-likelihood. For continuous probability distributions, we compute the probability density of observing the data rather than the probability itself. Since we are interested in relative (log)likelihoods, not the absolute probability of observing the data, we can ignore the distinction between the density ($P(x)$) and the probability (which includes a term for the measurement precision: $P(x)dx$).

6.1.1.1 Binomial maximum likelihood

Analytical approach:

$$l = \prod_{i=1}^{n} \binom{N}{k_i} p^{k_i} (1-p)^{N-k_i}$$

$$L = log(l) = \sum_{i=1}^{n} \left(log \binom{N}{k_i} + k_i log(p) + (N-k_i) log(1-p) \right)$$

then solve $\frac{dL}{dp} = 0$, get $\hat{p} = \frac{\sum_{i=1}^{n} k_i}{nN}$.

Numerics:

```
library(emdbook)
data(ReedfrogPred)
binomNLL1 <- function(p, k, N) {
    -sum(dbinom(k, prob = p, size = N, log = T))
}
x <- subset(ReedfrogPred, pred == "pred" & density == 10 & size == "small")
k <- x$surv
opt1 <- optim(fn = binomNLL1, par = c(p = 0.5), N = 10, k = k, method = "BFGS")
# use method = 'BFGS' for a single-parameter fit.
opt1$par  # the estimated value of the probability
exp(-opt1$value)  # the estimated value of max. likelihood.
# But usually we only care about the relative value.
library(bbmle)
m1 <- mle2(minuslogl = binomNLL1, start = list(p = 0.5), data = list(N = 10,
    k = k))
m1
mle2(k ~ dbinom(prob = p, size = 10), start = list(p = 0.5))
```

6.1.1.2 Gamma likelihhod

The likelihood equation for Gamma-distributed data is hard to maximize analytically, so we'll go straight to a numerical solution.

```
data(MyxoTiter_sum)
myxdat <- subset(MyxoTiter_sum, grade == 1)
gammaNLL1 <- function(shape, scale) {
    -sum(dgamma(myxdat$titer, shape = shape, scale = scale, log = T))
}
# starting paramters for gamma distribution are hard to find.  we can use
# those from the mothod of moments
gm <- mean(myxdat$titer)
gvm <- var(myxdat$titer)/mean(myxdat$titer)
```

```
m3 <- mle2(gammaNLL1, start = list(shape = gm/gvm, scale = gvm))
m3
mle2(myxdat$titer ~ dgamma(shape, scale = scale), start = list(shape = gm/gvm,
    scale = gvm), data = myxdat)
fitdistr(myxdat$titer, "gamma")  # report the rate = 1/scale
# fitdistr from MASS package, good for a single paramter fit.
```

6.1.2 Bayesian analysis

Bayesian estimation also uses the likelihood, but it differs in two ways from maximum likelihood analysis. **First**, we combine the likelihood with *a prior probability distribution* in order to determine a posterior probability distribution. **Second**, we often report the *mean* of the posterior distribution rather than its mode (which would equal the MLE if we were using a completely uninformative or "flat" prior). Unlike the mode, which reflects only local information about the peak of the distribution, the mean incorporates the entire pattern of the distribution, so it can be harder to compute.

6.1.2.1 Binomial distribution conjugate priors: Beta distribution

Conjugate priors also allow us to interpret the strength of the prior in simple ways. For example, the conjugate prior of the binomial likelihood that we used for the tadpole predation data is the Beta distribution. If we pick a Beta prior with shape parameters a and b, and if our data include a total of $\sum k$ "successes" (predation events) and $nN - \sum k$ "failures" (surviving tadpoles) out of a total of nN "trials" (exposed tadpoles), the posterior distribution is a Beta distribution with shape parameters $a + \sum k$, that is $(a + successes)$ and $b + (nN - \sum k)$, that is $b + failures$. If we interpret $a - 1$ as the total number of previously observed successes and $b - 1$ as the number of previously observed failures, then the new distribution just combines the total number of successes and failures in the complete (prior plus current) data set. When $a = b = 1$, the Beta distribution is flat, corresponding to no prior information ($a - 1 = b - 1 = 0$). When $a = b \to 0$, we get a very peculiar prior distribution with infinite spikes at 0 and 1. As a and b increase, the prior distribution gains more information and becomes peaked. We can also see that, as far as a Bayesian is concerned, it doesn't matter how we divide our experiments up. Many small experiments, aggregated with successive uses of Bayes' Rule, give the same information as one big experiment (provided of course that there is no variation in pertrial probability among sets of observations, which we have assumed in our statistical model for both the likelihood and the Bayesian analysis). To sum up:

1. The data: $X \sim Bin(n, p)$, $f(x|p) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$
2. The prior: $p \sim Beta(a, b)$, $\pi(p) = \frac{p^{a-1}(1-p)^{b-1}}{Beta(a,b)}$, $E(p) = \frac{a}{a+b}$, $Var(p) = \frac{ab}{(a+b)^2(a+b+1)}$. If both a, b are integers, then $Beta(a, b) = \frac{(a-1)!(b-1)!}{(a+b-1)!}$.
3. The posterior: $p|x \sim Beta(a + x, b + n - x)$

6.1.2.2 Poisson distribution conjugate priors: Gamma distribution

If our data were Poisson, we could use a conjugate prior Gamma distribution with shape α and scale s and interpret the parameters as α =total counts in previous observations and $1/s$ =number of previous observations. Then if we observed C counts in our data, the posterior would be a Gamma distribution with $\alpha' = \alpha + C$, $1/s' = 1/s + 1$. Sum up:

1. The data: $X \sim Poisson(\lambda)$, $f(x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$
2. The prior: $\lambda \sim Gamma(shape = \alpha, rate = \frac{1}{s})$. Some define the Gamma distribution in terms of $scale = s$. $\pi(\lambda) = \frac{(1/s)^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda}$. $\Gamma(\alpha)$ is the Gamma function. $E(\lambda) = \alpha s$, $Var(\lambda) = \alpha s^2$. If α is a positive integer, then $\Gamma(\alpha) = (\alpha - 1)!$.
3. The posterior: $\lambda|x_1, \cdots, x_n \sim Gamma(\alpha + \sum x_i, \frac{1}{s} + n)$. In the above example, $x_1 = C, n = 1$.

6.1.2.3 Normal distribution conjugate priors

- Normal with known variance, unkown mean
  1. The data: $X \sim N(\mu, \sigma^2 = 1)$, we know σ^2, say $\sigma^2 = 1$.
  2. The prior: $\mu \sim N(\theta = 3, \tau^2 = 4)$.
  3. The posterior: $\mu|x_1, \cdots, x_n = N\left(\frac{\frac{n\overline{X}}{\sigma^2} + \frac{\theta}{\tau^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}, \frac{1}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}\right)$
- Normal with known mean, unkown variance
  1. The data: $X \sim N(\mu = 1, \sigma^2)$.
  2. The prior of precision: $\varnothing = \frac{1}{\sigma^2} = Gamma(shape = \alpha, rate = \beta)$
  3. The posterior: $\varnothing|x_1, \cdots, x_n = Gamma(shape = \alpha + \frac{n}{2}, rate = \beta + \sum(x_i - \mu)^2/2$
- Normal with unkown mean, unkown variance
  1. The data: $X \sim N(\mu, \sigma^2)$.
  2. The prior: $\mu|\sigma^2 \sim N(\mu_0, \sigma^2/\lambda)$. Note that this prior for μ depends on the unknown σ^2 . σ^2~ Inverse χ^2 with df v and scale σ_0^2 .
  3. $\mu|x_1, \cdots, x_n \sim t$ with df $v + n$. The "center" of the t distribution is $\frac{\lambda}{\lambda+n}\mu_0 + \frac{n}{\lambda+n}\bar{x}$. The "scale" of the t posterior is a weighted average of σ_0^2, the sample variance s^2, and $(\bar{x} - \mu_0)^2$.

6.1.2.4 Gamma distribution: nonconjugate prior.

Unfortunately simple conjugate priors aren't always available, and we often have to resort to numerical integration to evaluate Bayes' Rule. Just plotting the numerator of Bayes' Rule, $(prior(p) \times L(p))$, is easy: for anything else, we need to integrate (or use summation to approximate an integral).

Bayesians often use the Gamma as a prior distribution for parameters that must be positive. Using a small shape parameter gives the distribution a large variance (corresponding to little prior information) and means that the distribution will be peaked at small values but is likely to be flat over the range of interest. Finally, the scale is usually set large enough to make the mean of the parameter (= shape · scale) reasonable.

For a Gamma distribution with shape= a and scale= s. Example: Prior(a)∼Gamma($shape = 0.01, scale = 100$); Prior(s)=Gamma($shape = 0.1, scale = 10$); Prior(a,s)=Prior(a)×Prior(s).

$$Posterior(a,s) = \frac{Prior(a,s) \times \mathcal{L}(a,s)}{\iint Prior(a,s) \times \mathcal{L}(a,s) da\, ds}$$

$$\bar{a} = \iint Prior(a,s) \times \mathcal{L}(a,s)\, a\, da\, ds; \bar{s} = \iint Prior(a,s) \times \mathcal{L}(a,s)\, s\, da\, ds$$

```
prior.as <- function(a, s) {
    dgammaa(a, shape = 0.01, scale = 100) * dgamma(s, shape = 0.1, scale = 10)
}
unscaled.posterior <- function(a, s) {
    prior.as(a, s) * exp(-gammaNLL1(shape = a, scale = s))
}
```

and use integrate (for 1-dimensional integrals) or adapt (in the adapt package; for multi-dimensional integrals) to do the integration. More crudely, we can approximate the integral by a sum, calculating values of the integrand for discrete values. However, integrating probabilities is tricky for two reasons. (1) Prior probabilities and likelihoods are often tiny for some parameter values, leading to roundoff error; tricks like calculating log-probabilities for the prior and likelihood, adding, and then exponentiating can help. (2) You must pick the number and range of points at which to evaluate the integral carefully. Too coarse a grid leads to approximation error, which may be severe if the function has sharp peaks. Too small a range, or the wrong range, can miss important parts of the surface. In practice, brute-force numerical integration is no longer feasible with models with more than about two parameters. The only practical alternatives are Markov chain Monte Carlo approaches.

6.2 Estimation for more complex functions

6.2.1 Maximum likelihood

```
mle2(titer ~ dgamma(shape, scale = a * day * exp(-b * day)/shape), start = list(a = 1,
    b = 0.2, shape = 50), data = myxdat, method = "Nelder-Mead")

## Error:  could not find function "mle2"
```

6.2.2 Bayesian analysis

MCMC. BUGS's modeling language is similar but not identical to R. For example, BUGS requires you to use <- instead of = for assignments. BUGS uses shape and rate parameters to define the Gamma distribution rather than shape and scale parameters: differences in parameterization are some of the most important differences between the BUGS and R languages. tilde (~) means "is distributed as".

6.3 Likelihood surfaces, profiles, and confidence intervals

6.3.1 Frequentist analysis: likelihood curves and profiles

The most basic tool for understanding how likelihood depends on one or more parameters is the *likelihood curve* or *likelihood surface*, which is just the likelihood plotted as a function of parameter values. By convention, we plot the negative log-likelihood rather than log-likelihood, so the best estimate is a minimum rather than a maximum. On a negative log-likelihood curve or surface, higher points represent worse fits.

If we want to deal with models with more than two parameters, or if we want to analyze a single parameter at a time, we have to find a way to isolate the effects of one or more parameters while still accounting for the rest. A simple, but usually **wrong**, way of doing this is to calculate a likelihood *slice*, fixing the values of all but one parameter (usually at their maximum likelihood estimates) and then calculating the likelihood for a range of values of the focal parameter.

Instead, we calculate likelihood *profiles*, which represent "ridgelines" in parameter space showing the minimum negative log-likelihoods for particular values of a single parameter. To calculate a likelihood profile for a focal parameter, we have to set the focal parameter in turn to a range of values, and for each value optimize the likelihood with respect to all of the other parameters.

Slices are always steeper than profiles, because they don't allow the other parameters to adjust to changes in the focal parameter.

6.3.1.1 The likelihood ratio test

Take some likelihood function $L(p_1, p_2, ..., p_n)$, and find the overall best (maximum likelihood) value, $L_{abs} = L(\hat{p}_1, \hat{p}_2, ..., \hat{p}_n)$ ("abs" stands for "absolute"). Now fix some of the parameters (say $p_1 ... p_r$) to specific values ($p_1^*, ... p_r^*$), and maximize with respect to the remaining parameters to get $L_{restr} = L(p_1^*, ... p_r^*, \hat{p}_{r+1}, ..., \hat{p}_n)$ ("restr" stands for "restricted", sometimes also called a *reduced or nested* model). The Likelihood Ratio Test says that the

distribution of twice the negative log of the likelihood ratio, $-2log(L_{restr}/L_{abs})$, called the *deviance*, is approximately χ^2 ("chi-squared") distribution with r degrees of freedom.

$$2\left(-\log \mathcal{L}_{restr} - (-\log \mathcal{L}_{abs})\right) \sim \chi^2_r$$

The definition of the LRT echoes the definition of the likelihood profile, where we fix one parameter and maximize the likelihood/minimize the negative log-likelihood with respect to all the other parameters: $r = 1$ in the definition above. Thus, for univariate confidence limits we cut off the likelihood profile at (min. neg. log. likelihood + $\chi^2_1(1-\alpha)/2$), where α is our chosen confidence level (0.95, 0.99, etc.). (The cutoff is a one-tailed test, since we are looking only at differences in likelihood that are larger than expected under the null hypothesis.)

`R` can compute profiles and profile confidence limits automatically. Given an `mle2` fit `m`, `profile(m)` will compute a likelihood profile and `confint(m)` will compute profile confidence limits. `plot(profile(m2))` will plot the profile, square-root transformed so that a quadratic profile will appear V-shaped (or linear if you specify `absval=FALSE`). This transformation makes it easier to see whether the profile is quadratic, since it's easier to see whether a line is straight than it is to see whether it's quadratic.

The LRT is only correct asymptotically, for large data sets. For small data sets it is an approximation, although one that people use very freely. The other limitation of the LRT that frequently arises, although it is often ignored, is that it only works when the best estimate of the parameter is not on the edge of its allowable range (Pinheiro and Bates, 2000). For example, if you are fitting an exponential model $y = \exp(rx)$ that must be decreasing, so that $r \leq 0$, and your best estimate of r is equal to 0, then the LRT estimate for the upper bound of the confidence limit is not technically correct.

6.3.2 Bayesian approach: posterior distribution and marginal distributions

Instead of drawing likelihood curves, Bayesians draw the posterior distribution (proportional to prior × L). Instead of calculating confidence limits using the (frequentist) LRT, they define the *credible interval*, which is the region in the center of the distribution containing 95% (or some other standard proportion) of the probability of the distribution, *bounded by values on either side that have the **same** probability* (or probability density). Technically, the credible interval is the interval $[x_1, x_2]$ such that $P(x_1) = P(x_2)$ and $C(x_2) - C(x_1) = 1 - \alpha$, where P is the probability density and C is the cumulative density. The credible interval is slightly different from the frequentist confidence interval, which is defined as $[x_1, x_2]$ such that $C(x_1) = \alpha/2$ and $C(x_2) = 1 - \alpha/2$.

The *marginal probability density* is the Bayesian analogue of the likelihood profile. Where frequentists use likelihood profiles to make inferences about a single parameter while taking the effects of the other parameters into account, Bayesians use the marginal posterior probability density, the overall probability for a particular value of a focal parameter integrated over all the other parameters.

6.4 Confidence intervals for comples models: quadratic approximation

6.5 Comparing models

Parsimony (sometimes called "Occam's razor") is a general argument for choosing simpler models even though we know the world is complex. All other things being equal, we should prefer a simpler model to a more complex one — especially when the data don't tell a clear story. Model selection approaches typically go beyond parsimony to say that a more complex model must be not just better than, but a specified amount better than, a simpler model. If the more complex model doesn't exceed a threshold of improvement in fit (we will see below exactly where this threshold comes from), we typically reject it in favor of the simpler model.

Model complexity also affects our predictive ability. Walters and Ludwig (1981) simulated fish population dynamics using a complex agestructured model and showed that in many cases, when data were realistically sparse and noisy, they could best predict future (simulated) dynamics using a simpler non-age-structured model. In other words, even though they knew for sure that juveniles and adults had different mortality rates (because they simulated the data from a model with mortality differences), a model that ignored this distinction gave more accurate predictions. This apparent paradox is an example of the ***bias-variance tradeoff*** introduced in Chapter 5. As we add more parameters to a model, we necessarily get an increasingly precise fit to the particular data we have observed (the bias decreases), but our accuracy for predicting future observations decreases as well (the variance increases). Data contain a fixed amount of information; as we estimate more and more parameters we spread the data thinner and thinner. Eventually the gain in precision from having more details in the model is outweighed by the loss in accuracy from estimating the effect of each of those details more poorly. In Ludwig and Walters's case, spreading the data out across age classes meant there was not enough data to estimate each age class's dynamics accurately.

6.5.1 Likelihood ratio test: nested models

Comparisons among different groups can also be framed as a comparison of nested models. If the more complex model has the mean of group 1 equal to $a1$ and the mean of group 2 equal to $a2$, then the nested model (both groups equivalent) applies when $a1 = a2$. It is also common to parameterize this model as $a2 = a1 + \delta 12$, where $\delta 12 = a2 - a1$, so that the simpler model applies when $\delta 12 = 0$. This parameterization works better for model comparisons since testing the hypothesis that the more complex model is better becomes a test of the value of one parameter ($\delta 12 = 0$?) rather than a test of the relationship between two parameters ($a1 = a2$?).

The Likelihood Ratio Test can compare any two nested models, testing whether the nesting parameters of the more complex model differ significantly from their null values. Put another way, the LRT tests whether the extra goodness of fit to the data is worth the added complexity of the additional parameters. To use the LRT to compare models, com-

pare the difference in deviances (the more complex model should always have a smaller deviance — if not, check for problems with the optimization) to the critical value of the χ^2 distribution, with degrees of freedom equal to the additional number of parameters in the more complex model. If the difference in deviances is greater than $\chi^2_{n2-n1}(1-\alpha)$, then the more complex model is significantly better at the $p = \alpha$ level. If not, then the additional complexity is not justified.

Choosing among statistical distributions can often be reduced to comparing among nested models. The most common use of the LRT in this context is to see whether we need to use an *overdispersed distribution such as the negative binomial or beta-binomial* instead of their *lower-variance counterparts (Poisson or binomial)*. **The Poisson distribution is nested in the negative binomial distribution when $k \to \infty$.** If we fit a model with a and b varying but using a Poisson distribution instead of a negative binomial, we can then use the LRT to see if adding the overdispersion parameter is justified: `anova(poisfit.ab, nbfit.ab)`.

6.5.2 Information criteria

One way to avoid having to make pairwise model comparisons is to select models based on *information criteria*, which compare all candidate models at once and do not require nested alternatives. These relatively recent alternatives to likelihood ratio tests are based on the expected distance (quantified in a way that comes from information theory) between a particular model and the "true" model (Burnham and Anderson, 1998, 2002). In practice, all information-theoretic methods reduce to the finding the model that minimizes some criterion that is the sum of a term based on the likelihood (usually twice the negative log-likelihood) and a *penalty term* which is different for different information criteria.

For all information criteria, small values represent better overall fits.

The *Akaike Information Criterion*, or AIC, is the most widespread information criterion, and is defined as $AIC = -2L + 2k$, where L is the log-likelihood and k is the number of parameters in the model. For small sample sizes (n) — such as when n/k < 40 (Burnham and Anderson, 2004, p. 66)) — you should use a finite-size correction and apply the $AIC_c = AIC + \frac{2k(k+1)}{n-k-1}$

The second most common information criterion, the *Schwarz criterion* or *Bayesian information criterion* (SC/BIC) $*$, uses a penalty term of $(logn)k$. When n is greater than $e^2 \approx 9$ observations (so that log n > 2), the BIC is more conservative than the AIC, insisting on a greater improvement in fit before it will accept a more complex model.

Information criteria do not allow frequentist significance tests based on the estimated probability of getting more extreme results in repeated experiments (some statisticians would say this is an advantage). **With ICs, you cannot say that there is a statistically significant difference between models**; a model with a lower IC is better, but there is no p-value associated with how much better it is. Instead, there are ***commonly used rules of thumb***: **models with ICs less than 2 apart (ΔIC < 2) are more or less equivalent; those with ICs 4-7 apart are clearly distinguishable; and models with ICs more than 10 apart are definitely different**. Richards (2005) concurs with these recommendations,

but cautions that simply dropping models with ΔAIC > 2 (as some ecologists do) will probably discard useful models.

One big advantage of IC-based approaches is that they do not require nested models. You can compare all models to each other, rather than stepping through a sometimes confusing sequence of pairwise tests. In ICbased approaches, you simply compute the likelihood and IC for all of the candidate models and rank them in order of increasing IC. The model with the lowest IC is the best fit to the data; those models with ICs within 10 units of the minimum IC are worth considering. As with the LRT, the absolute size of the ICs is unimportant — only the differences in ICs matter.

6.5.3 Bayesian analyses

Bayesians are on the whole less interested in formal methods of model selection. Dropping a parameter from a model is often equivalent to testing a null hypothesis that the parameter is exactly zero, and Bayesians consider such point null hypotheses silly. They would describe a parameter's distribution as being concentrated near zero rather than saying its value is exactly zero.

The marginal likelihood of a model is the probability of observing the data (likelihood), averaged over the prior distribution of the parameters:

$$\hat{L} = \int L(x) \cdot Prior(x) dx$$

where x represents a parameter or set of parameters (if a set, then the integral would be a multiple integral). The marginal likelihood (the average probability of observing a particular data set exactly) is often very small, and we are really interested in the relative probability of different models. If we have two models with marginal likelihoods $\hat{}L1$ and $\hat{}L2$, the *Bayes factor* is the ratio of the marginal likelihoods, $B12 = \hat{}L1/\hat{}L2$, or the odds in favor of model 1.

A more recent criterion, conveniently built into WinBUGS, is the DIC or *deviance information criterion*, which was designed particularly for models containing random effects where even specifying the number of parameters is confusing.

6.5.4 Model weighting and averaging

Bayesians themselves would say that you should not simply select one model. Taking the best model and ignoring the rest is equivalent to assigning a probability of 1.0 to the best and 0.0 to the rest. Model averaging methods take the average of the predictions of different models, weighted by the probability of the models or by some other index.