CHAPTER 12

Mixture Models for Overdispersed Data

Jonathan R. Rhodes, The University of Queensland, School of Geography, Planning and Environmental Management and ARC Centre of Excellence for Environmental Decisions, Brisbane, QLD 4072, Australia. Email: j.rhodes@uq.edu.au.

12.1 Introduction

Much of ecological statistics, including most of the methods in this book, rely on parametric statistics. Parametric statistics make specific assumptions about the nature of the probability distributions that our data arise from, in contrast to non-parametric statistics that make far fewer such assumptions. These assumptions can be an advantage because it allows us to make clearly defined and transparent assumptions about the processes generating our data (Royle and Dorazio 2008). This, in turn, allows us to test explicit hypotheses about the ecological processes that led to observed data. However, a disadvantage is that ecological data can fail to meet the assumptions of the standard probability distributions (e.g., normal, Poisson and binomial distributions) used in parametric statistics. A particularly common problem in this context is a phenomenon known as overdispersion that arises when data are more variable than can be accommodated by the parametric distribution being used to describe it (McCullagh and Nelder 1989). This chapter is about how to deal with overdispersion when using parametric statistics for ecological inference. In particular, I show how a class of models known as mixture models (Mengersen et al. 2011) can be used to help ensure that our statistical tests are valid when overdispersion is present and to better understand the drivers of overdispersion for improved ecological inference.

To illustrate the idea of overdispersion, imagine you go out to a number of randomly selected sites and count the number of individuals of a species at each site. What would these data look like if the distribution of the species was highly spatially aggregated, occurring at high densities at a few locations where habitat is suitable, but being absent from other areas where habitat is unsuitable? If this were the case, we would expect the data to consist predominantly of high values in sites where habitat is suitable, but zero values elsewhere. That is, we would tend to observe data at the two extremes of the distribution, with values in between being much less common. A consequence of the data lying at the two extremes is that its variance will be higher than the theoretical variance for the standard parametric distribution used to model count data, which is the Poisson distribution. This happens because the Poisson distribution assumes that the data lie predominantly around the center of the distribution, rather than the extremes. In this example, it is the spatial aggregation process that leads to overdispersion, but overdispersion in ecological data can be caused by a range of ecological, observation and modeling processes (Haining et al. 2009, Linden and Mantyniemi 2011).

A major issue with overdispersion is that it generates bias in statistical tests. Overdispersion means that the true variances of the data are larger than the theoretical variances assumed by parametric distributions. This leads to incorrect model likelihoods and, because variances are underestimated, we will tend to incorrectly reject the null hypothesis (i.e., make Type I errors) more often than we should (also see Chapters 2 and 3). Consequently, we need tools that allow us to account for overdispersion in our statistical models so that biases in our statistical tests are reduced or eliminated. Anderson et al. (1994) illustrate one way to do this using quasi-likelihood methods (see Chapter 6) to adjust Akaike's Information Criteria (AIC) values to correct for overdispersion in capture-recapture data. This approach tells us little about the nature of the overdispersion itself, it just accounts for it, but the nature of overdispersion can also provide important information about underlying ecological processes and/or observation processes (Martin et al. 2005). For example, Rhodes et al. (2008b) model the effects of marine pollution on fecundity in the copepod *Tigriopus japonicus* and use a mixture model that explicitly models the processes driving an excess of zeroes (and therefore overdispersion) in their data. By being explicit about the processes driving overdispersion they were able to make inferences about the effect of pollution on two different processes: the number of individuals that entirely fail to breed (causing an excess of zeroes), and the number of young per successful breeder. In this case, overdispersion is not just a nuisance that we want to control for, but reflects a key ecological process of interest. Consequently, we often want methods to explicitly model the overdispersion process when these processes themselves are of direct interest.

Mixture models are a particular class of statistical model that allow us to control both for overdispersion in our statistical tests and to model explicitly the processes that drive overdispersion, resulting in improved ecological inference. These models allow for greater variability than standard distributions by allowing the parameters (and sometimes the structure) of standard statistical models to vary randomly, rather than being fixed. To illustrate the idea of a mixture model, let us go back to our hypothetical example of the spatially aggregated count data that we looked at above. The standard way to model this type of data would be using a Poisson distribution that has a single parameter, λ, representing the mean. However, since our count data have a high frequency of zeroes and a high frequency of high values, the true variance of the data will be greater than the theoretical variance of the Poisson distribution. To deal with this, we could use a mixture model that assumes that the λ parameter can vary randomly and take one of two values: either zero, or a fixed value greater than zero, with either case occurring with a given probability. This model explicitly accounts for the high frequencies of zeroes and high frequencies of high values, by allowing the mean to be either zero or fixed value greater than zero respectively. This mixture model is known as a zero-inflated Poisson distribution (Lambert 1992). Importantly, by allowing λ to vary randomly we characterize sites as suitable ($\lambda > 0$) and unsuitable ($\lambda = 0$) and therefore the process that leads to the overdispersion is explicitly characterized. A nice property of a mixture of this type, therefore, is that not only do we control for overdispersion, but we can make inferences about the processes that lead to that overdispersion, such as estimating the proportion of sites where habitat is suitable versus unsuitable.

Although I have used count data and the problem of estimating the distribution and abundance of a species to illustrate the idea of a mixture model, we can apply mixture models to other classes of problems. In fact, mixture models represent a highly flexible approach for dealing with overdispersion across a very wide range of classes of statistical problems (Mengersen et al. 2011). In ecology, mixture models have been successfully applied to deal with overdispersion and heterogeneity in a range of applications, including: modeling species' distributions and abundance (Tyre et al. 2003, Royle 2004, Wenger and Freeman 2008), survival analysis (Pledger and Schwarz 2002), population dynamics (Kendall and Wittmann 2010); disease ecology and parasitology (Calabrese et al. 2011); community ecology (Colwell et al. 2004); and dispersal ecology (Clark et al. 1999). However, there are three common problems in ecology where mixture models are particularly useful: (1) accounting for an excess of zeroes in data (arising either due to ecological or observation processes); (2) accounting for heterogeneity among sampling units (e.g., individuals or social groups); and (3) making explicit inferences about two or more ecological or observation processes that jointly give rise to overdispersed data (e.g., short- and long-distance dispersal processes that both contribute to the distribution of dispersal distances).

In this chapter I present mixture models as a powerful and flexible way to deal with overdispersion in ecological data and discuss how this approach can be used to account for overdispersion and facilitate improved inference by understanding the overdispersion process itself. Although mixture models are not new and some mixture models (e.g., the negative-binomial distribution) are commonly used in ecology, the routine consideration of mixture models as a flexible approach for modeling complex ecological data is rare. The discipline can therefore benefit greatly from a better-informed use of mixture models that will lead to improved ecological inference. Further, faster computers and new computational methods now make it possible for most ecologists to routinely fit complex statistical models to ecological data. Ecologists are therefore in a unique position to extend their toolkit to the more general use of mixture models.

The remainder of the chapter is divided into four sections. In the first section I define overdispersion and provide guidance on how it can be detected. In the second section I discuss mixture models in more detail and highlight the main types of mixture models used in ecology. I then present two empirical examples. The first example is a survival analysis problem where I use a mixture model to deal with heterogeneity among groups. In the second example I present a problem where the aim is to estimate species' abundance from count data that is zero-inflated. In this example, I show how mixture models can be used to model both ecological and observational sources of the zero inflation. I end with a discussion of the benefits and challenges of using mixture models for ecological inference, especially in comparison to alternative approaches, and highlight the key things to consider when using mixture models.

12.2 Overdispersion

12.2.1 What is overdispersion and what causes it?

Data are defined as overdispersed if the variance of the data is greater than the theoretical variance of the probability distribution being used to describe the data generation process (Hinde and Demetrio 1998). In other words, overdispersion is always relative to a specified probability distribution. Overdispersion is often most apparent in count and presence/absence data because the variance of a standard Poisson or binomial distribution is a function of the mean, rather than estimated independently from the data. In the case of the Poisson distribution, the variance is equal to the mean (i.e., $\sigma^2 = \lambda$) and, in the case of the binomial distribution, the variance is equal to the number of trials multiplied by the success probability multiplied by the failure probability (i.e., $\sigma^2 = np[1-p]$). If the data fail to conform to these characteristics of the variance, then the true variance of the data can be higher than the theoretical variance and therefore overdispersed. Such overdispersion in ecological data can arise from ecological processes, observation processes and/or misspecification of the mean (Table 12.1, Haining et al. 2009, Linden and Mantyniemi 2011). It is these sources of overdispersion that we will look at next.

Two important ways in which ecological mechanisms can lead to overdispersion include: (1) causing spatial/temporal clustering or aggregation (see also Chapter 10) and/or (2) introducing heterogeneity among sampling units. As I highlighted in section 12.1, spatial clustering tends to generate data with too many high counts (from locations where the species is present) and/or too many zero counts (from locations where the species is absent), resulting in a variance that is greater than the theoretical variance of the Poisson distribution. Cunningham and Lindenmayer (2005), for example, show that, in the Central Highlands of Victoria, Australia, counts of the threatened Leadbeater's posssum (*Gymnobelideus leadbeateri*) are overdispersed primarily due to an excess of zero values. This is driven by the species' distribution being highly spatially clustered in only a few areas of the landscape where its habitat occurs. In a similar way, temporal clustering can also lead to overdispersion. For example, disturbance events that impact on ecosystems such as cyclones (hurricanes or typhoons for those not fortunate enough to live in Australia) can be highly temporally clustered and therefore overdispersed (Mumby et al. 2011). Heterogeneity among sampling units (e.g., genetic or phenotypic variation among individuals) can also result in an excess of high and/or low values and therefore overdispersion relative to the Poisson or binomial distributions. A typical example of this is where breeding success varies among individuals, leading to highly variable reproductive output and overdispersion in data on reproductive output (Quintero et al. 2007, Kendall and Wittmann 2010).

Observation processes commonly result in data inaccuracies that can also lead to overdispersion by increasing variability in the data. For example, presence / absence data collected where there is imperfect detection (which is almost always the case) can lead to an excess of zeroes and overdispersion relative to the binomial distribution (Tyre et al. 2003). However, although zero-inflation caused by detection error can appear similar to zero inflation caused by ecological processes, the inferences we make from the data are normally quite different. This is because, in the presence of detection error, we are usually interested in making ecological inferences after stripping out the process (detection error) causing overdispersion. On the other hand, when ecological processes are the cause of zero-inflation, we are commonly interested making inferences about overdispersion as a component of the ecological processes of interest.

The final way overdispersion can arise is when the mean is misspecified. Most statistical models specify the mean of the appropriate distribution as a function of covariates (e.g., in linear regression, the mean of the normal distribution is specified as a function of covariates). In the case of the Poisson and binomial distributions, if the function that links the mean to the covariates is misspecified in a way that results in the mean being underestimated (e.g., due to important covariates, or non-linear terms, being missed), then the variance will also be underestimated. This leads to overdispersion because the estimated variance is lower than the true variance of the data. Missing covariates are likely to be common in ecology, particularly in applications such as modelling the distribution of species where the factors driving distributions are often not well understood, or even when they are understood, often cannot be directly measured (Barry and Elith 2006).

12.2.2 Detecting overdispersion

Prior to and during the development of statistical models for data that may be overdispersed, it is important to be able to identify whether the data are in fact overdispersed or not. There are three primary ways in which we can detect overdispersion: (1) inspect histograms of the raw data; (2) inspect quantile-quantile plots of the residuals of the model; and/or (3) conduct formal hypothesis tests, or model selection. Often, simply inspecting a histogram of the raw data and comparing this against expected frequencies based on the relevant standard distribution can reveal important information on whether data are overdispersed or not. Table 12.1 illustrates what histograms look like relative to the Poisson distribution for different causes of overdispersion in count data, but we can construct similar plots for any distribution. However, overdispersion can sometimes be accounted for by the relationship with a covariate included in the model, rendering residuals that are not overdispersed. Therefore, a preferred and more sophisticated, approach is to inspect a quantile-quantile plot of the residuals; is a standard method for visually comparing the distribution of data versus the expected distribution. Quantile-quantile plots show the actual ordered residuals from the model against the expected ordered residuals of the model; a plot lying close to the 1:1 line represents good agreement between the distribution of the data and the expected distribution. Quantile-quantile plots of overdispersed data will tend to lie below the 1:1 line at the lower end of the distribution and/or lie above the 1:1 line at the higher end of the distribution. This reflects the tendency for overdispersed data to contain more extreme values than expected, but the exact pattern will depend upon the nature of the overdispersion present. Table 12.1 illustrates what quantile-quantile plots look like relative to the Poisson distribution for different causes of overdispersion. Landwehr et al. (1984) develop a useful simulation approach for constructing quantile-quantile plots for logistic regression, but the approach is flexible enough to be applied to any model. The inspection of histograms of the raw data and quantile-quantile plots of the residuals represent qualitative approaches for detecting overdispersion. One advantage of this approach is that it allows a visual representation of the distribution of the data relative to the expected distribution that can help in pinpointing how overdispersion arises in the data. However, it is also possible to take a more formal approach and explicitly test for overdispersion by conducting hypothesis (score) tests or to use multi-model selection methods (see also Chapter 3; Dean 1992, Richards 2008). I will expand on and illustrate these approaches in the empirical examples later in the chapter.

12.3 Mixture Models

In the previous section I discussed the nature of overdispersion and described how to identify whether your data are overdispersed or not. Now we are going to turn our attention to mixture models as a way of dealing with overdispersion in our statistical models. In this section I define what a mixture model is, identify some typical mixture models used in ecology and briefly mention the different ways in which mixture models can be fit to data. A complete technical treatment of mixture models is not possible within a single book chapter, but McLachlan and Peel (2000), Johnson et al. (2005) and Mengersen et al. (2011) provide more comprehensive and technical treatments of mixture models.

12.3.1 What is a mixture model?

To illustrate the idea of a mixture model, let us go back to our hypothetical example of the surveys of the highly spatially aggregated species that I described in the Introduction. There I discussed the idea that we could formulate a mixture model in a way that allowed the mean to vary randomly between having a value of zero and having a value greater than zero. That is to say, some data points would come from a distribution with a mean of zero and some will come from a distribution with a mean greater than zero. But let us now look at this more formally. First assume that, regardless of whether the mean is zero or greater than zero, the data are Poisson distributed (see Appendix LIKELIHOODS for the definition of the probability density function for the Poisson distribution). Then, when the mean is greater than zero, the probability density function is

$$f(y \mid \lambda > 0) = \frac{e^{-\lambda}\lambda^{y}}{y!}, \qquad (12.1)$$

and, by substituting zero for the mean λ when the mean equals zero, the probability density function is

$$\begin{aligned} f(y \mid \lambda = 0) &= \frac{e^{-0}0^{y}}{y!} \\ &= \begin{cases} 1 \text{ when } y = 0 \\ 0 \text{ when } y > 0 \end{cases}. \end{aligned} \qquad (12.2)$$

If we let p be the probability that the mean is zero, so the probability that the mean is greater than zero is $1 - p$, then the probability density function is

$$\begin{aligned} g(y \mid \lambda) &= pf(y \mid \lambda = 0) + (1-p)f(y \mid \lambda > 0) \\ &= p\frac{e^{-0}0^y}{y!} + (1-p)\frac{e^{-\lambda}\lambda^y}{y!} \\ &= \begin{cases} p + (1-p)e^{-\lambda} \text{ when } y = 0 \\ (1-p)\dfrac{e^{-\lambda}\lambda^y}{y!} \text{ when } y > 0 \end{cases} \end{aligned} \quad . \qquad (12.3)$$

This is the probability density function for the zero-inflated Poisson distribution (Lambert 1992); a mixture model that has a mean equal to $(1-p)\lambda$ and variance equal to $(1-p)(\lambda + p\lambda^2)$. The variance of the zero-inflated Poisson distribution is always greater than the mean (in contrast to the Poisson), since $(1-p)(\lambda + p\lambda^2) > (1-p)\lambda$ when $p > 0$. Therefore, if we were to use a Poisson distribution to model these data we would underestimate the variance. Using instead a zero-inflated Poisson distribution corrects this problem and allows for overdispersion to be accommodated in our model (in this case, in the form of zero-inflation).

Equation 12.3 shows that the mixture model is essentially a weighted sum of two probability density functions (the mixture components), with weights p and $1-p$ (the mixture weights). This is what is known as a **finite mixture distribution** because it is a finite sum of distributions. We can generalize this idea to a K-component finite mixture model, $g(y \mid \boldsymbol{\Theta})$, which is any convex combination of K probability density functions such that

$$g(y \mid \boldsymbol{\Theta}) = \sum_{i=1}^{K} \omega_i f_i(y \mid \boldsymbol{\theta}_i) \quad \text{subject to} \quad \sum_{i=1}^{K} \omega_i = 1, \qquad (12.4)$$

where $f_i(y \mid \boldsymbol{\theta}_i)$ is a probability density function, with parameters $\boldsymbol{\theta}_i$, representing mixture component i, ω_i is the mixture weight for component i, and $\boldsymbol{\Theta} = (\omega_1, \ldots, \omega_k, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k)$. In ecology, mixtures of more than two distributions may be appropriate when we want to capture more than two processes generating the data. For example, Kendall and Wittmann (2010) use a finite mixture model with more than two components to model multiple process that drive reproductive output in birds, mammals and reptiles. Many of the processes they consider (that include, nest building success, number of eggs laid or births, chance of nest destruction, and offspring survival) can result in overdispersion in reproduction data, but they explicitly account for them using a finite mixture model with more than two components. In general, finite mixture models provide a highly flexible approach for modeling non-standard distributions and are well suited to accounting for many kinds of overdispersion where model parameters can take a finite number of discrete values (Mengersen et al. 2011).

Another kind of mixture model arises when one or more of the parameters of a probability distribution varies randomly and can take an infinite number of values. In this case, the number of mixture components (the number of different possible probability distributions) is not discrete anymore and becomes infinite, reflecting the infinite number of possible values for the parameter(s). In the example above we assumed that the mean, λ, could take one of two discrete values: (1) a value of zero; or (2) a fixed value greater than zero. But, what if the mean can actually take any random value between zero and infinity? In this case, rather than just being able to take two discrete values, λ could be an infinite number of values, with the appropriate model changing from a finite to an **infinite mixture**. This could arise, for example, if mean abundance or density varies randomly across a landscape due to some ecological processes, such as variation in habitat quality. This would likely result in a pattern different from zero-inflation, but still cause overdispersion in the data.

If the parameter with random variation (such as the mean) varies according to a discrete distribution (e.g., Poisson), the resulting mixture is known as a **countable mixture**, but if it varies randomly according to a continuous distribution (e.g., normal or gamma), the resulting mixture is known as a **continuous mixture**. An example of a continuous mixture is when modeling the spatial distribution of a species' abundance, but where there is continuous random variation in the mean abundance of the species across a landscape. Here, random variation in mean abundance would best be described by a continuous distribution, so a continuous mixture would be used. An example of a countable mixture is when spatially modeling the proportion of individuals of a species calling across a landscape, but where the number of individuals present at sites varies randomly across the landscape. Here, variation in the number of individuals must be described by a discrete distribution (since the number of individuals must be an integer) and so a countable mixture would be used.

To more formally define these types of mixture models, consider a distribution that has only one parameter, θ, and that this parameter varies randomly according to some probability density function, $h(\theta \mid \boldsymbol{\psi})$, where $\boldsymbol{\psi}$ is a vector of the parameters of $h(\theta \mid \boldsymbol{\psi})$. If the distribution of θ is discrete, then a countable mixture results and the probability density function is

$$g(y \mid \boldsymbol{\psi}) = \sum_i f(y \mid \theta_i) h(\theta_i \mid \boldsymbol{\psi}), \qquad (12.5)$$

where θ_i are the discrete values of θ and the sum is over all possible value of θ_i. This can be thought of as analogous to a finite mixture model, except the mixture weights, ω_i, are replaced by $h(\theta_i \mid \boldsymbol{\psi})$ and there are an infinite number of possible mixture components, $f(y \mid \theta_i)$. If, on the other hand, the distribution of θ is continuous, then a continuous mixture results and the probability density function is

$$g(y \mid \boldsymbol{\psi}) = \int f(y \mid \theta) h(\theta \mid \boldsymbol{\psi}) d\theta, \qquad (12.6)$$

where the integration is over all possible values of θ. This is similar to a countable mixture, but we integrate over continuous values of θ rather than summing over discrete values of θ. Countable and continuous mixture models are important because they provide an explicit and flexible framework for modeling heterogeneity across sampling units (Johnson et al. 2005). In the empirical examples later in the chapter I will illustrate the use of finite and infinite (countable and continuous) mixture models in an ecological context.

12.3.2 Mixture models used in ecology

Most standard distributions used in ecology have corresponding overdispersed versions based on mixture distributions (Table 12.2). The Poisson distribution, which is typically used to model count data, has an overdispersed version known as the negative-binomial. This distribution is a Poisson-gamma continuous mixture that explicitly models heterogeneity in the Poisson rate parameter through a gamma distribution. The zero-inflated Poisson distribution that I discussed above is another overdispersed version of the Poisson distribution that is increasingly being used to model excess zeroes and observation error in ecological data (Lambert 1992, Martin et al. 2005). The binomial distribution, which is typically used to model presence / absence data, has an overdispersed version known as the beta-binomial that accounts for heterogeneity in the binomial probability among sampling units. This distribution is a beta-binomial continuous mixture that models heterogeneity in the binomial probability through a beta distribution. Similarly to the Poisson distribution, the binomial distribution also has a zero-inflated version; the zero-inflated binomial distribution (Hall 2000, Martin et al. 2005). The beta-binomial distribution can also be generalized to an overdispersed multinomial distribution; the Dirichlet-multinomial distribution.

One of the most commonly used of these mixture models in ecology is the negative-binomial distribution, which tends to be the default strategy for dealing with overdispersion in count data (Linden and Mantyniemi 2011). However, the negative-binomial distribution does not deal with overdispersion that arises due to zero-inflation that is common in ecological data (Martin et al. 2005). This is because, although it models variation in the mean of the Poisson distribution, it does not model any process that specifically leads to zero values. Zero-inflated models are more appropriate in this case, but the fact that zero-inflation can arise through either ecological and/or observation processes complicates the choice of mixture model and inference. Recognition of the particular problem of observation error has led to a rapidly growing area of statistical ecology that uses mixture models to explicitly account for observation error in both count and presence/absence data (MacKenzie et al. 2006, Royle and Dorazio 2008).

12.4 Empirical Examples

In this section I present two empirical examples to illustrate the use of mixture models to account for overdispersion that arises from both ecological processes and observation processes. Throughout the examples I illustrate the process of detecting overdispersion and choosing appropriate models. I also emphasize how mixture models allow us to reduce bias in our statistical tests, but also importantly that they allow us to make inferences that would otherwise not be possible. In the first example, I apply mixture models to data from koala dung decay trials to illustrate how these models can be used to disentangle the role of two distinct decay processes that lead to overdispersion in the data. In the second example, I use data on lemur counts from eastern Madagascar for two species, to illustrate how mixture models can be used to distinguish between, and control for, overdispersion that arises from ecological and observation processes in modeling abundance.

There are a number of packages available in R for fitting mixture models, including: `Flexmix`; `mixtools`; and `mclust` (mixtures of normal distributions only). However, in the empirical examples I present here I use R to construct likelihoods for the appropriate mixture models and fit these models to data by maximizing the likelihood using the package `bbmle`. One advantage of writing your own likelihood functions is that it allows for greater flexibility, but here it also illustrates how we can fit mixture models by maximum likelihood in a way that expands on the concepts developed in Chapter 3. However, for many applications, existing mixture model packages available in R may be perfectly sufficient.

12.3.1 Using binomial mixtures to model dung decay

For many species that are highly cryptic, the only realistic way we can determine their presence/absence is indirectly through signs that they leave. However, some signs, such as dung and snow tracks, decay and disappear over time and the rates at which they decay can vary substantially both spatially and temporally. This causes problems when using these types of indirect signs to estimate presence or absence, for two main reasons. First, if signs decay very slowly we may detect old signs in locations where the species is no longer present and we will make false-positive errors. On the other hand, if signs decay very quickly, we may fail to detect signs in areas where the species is still present, and we will make false-negative errors. If there are high levels of spatial and temporal variation in decay rates, this will likely introduce bias into our estimates of presence/absence, and thus will bias our estimates of how presence/absence depends on habitat variables. One way to try to deal with this is to model decay rates of the signs and then adjust for those decay rates in our estimates of presence/absence.

Koala dung is a primary means by which koala distributions are estimated since our ability to detect koalas using direct observations is poor (McAlpine et al. 2008, Rhodes et al. 2008a). This is because the species is highly cryptic and occurs at low densities over much of its range. Nonetheless, dung decay rates can vary both spatially and temporally, thus introducing biases into estimates of presence and absence, so there is a need to understand how spatially and temporally variable dung decay rates are. Here, I illustrate how we can use overdispersed binomial mixture models to analyses data from koala dung decay trials that were designed to understand and estimate dung decay rates under a range of habitat and climatic conditions (Rhodes et al. 2011).

The decay trial data were collected in Coffs Harbour, New South Wales, Australia between April 1996 and March 1997 (Appendix A). The trials were conducted at five sites, with three plots nested within each site, and at each site, plots were located in different topographic positions: one on a ridge; one on a mid-slope; and one in a gully. Each month, a group of 10 fresh koala dung pellets were laid out in each plot and the number of pellets that had disappeared were counted at approximately fortnightly intervals. Data were also available on daily rainfall, daily average humidity, and daily average temperatures for the study area. The data therefore consisted of the number of pellets that had disappeared (decayed) in each recording interval and a series of possible covariates for predicting the probability of dung decay, namely: site, topography, pellet age, rainfall, humidity, and temperature.

The raw decay data are overdispersed, with many more low values and more high values than would be expected under the standard binomial distribution (Fig. 12.1). This is confirmed by looking at the relative variances of the raw data versus the expected values based on the binomial distribution (0.064 for the raw data versus 0.044 for the binomial distribution). In particular, there are many more zeroes in the data (data points where no pellets disappeared in a time interval), than we would expect under the binomial distribution. There is also a slightly greater frequency of high values (where all pellets disappeared in a time interval), than would be expected under the binomial distribution. However, this difference is small relative to the zero-inflation.

One way to model the overdispersion would be to use the overdispersed version of the binomial distribution, the beta-binomial, which is a continuous mixture model that assumes that the binomial probability varies randomly according to the beta distribution (sensu Rhodes et al. 2011). This would account for extra-binomial variation that may have arisen due to inherent variation in decay rates among groups, possibly due to random variation in environmental conditions or pellet susceptibility to decay. However, this approach may fail to adequately account for the high levels of zero-inflation, which could have arisen via an entirely different process. It is possible, based on our understanding of the decay process, that there is another mechanism operating, related to whether the agents that cause pellet decay (e.g., insect, animal and/or bacterial activity) are present or not. If agents are not present, then no decay will occur, resulting in zero values in the data. On the other hand, if agents are present, decay will occur, but the rate of decay may then still depend on environmental variation or variation in susceptibility to decay. One way to proceed to deal with this is to use a zero-inflated binomial or zero-inflated beta-binomial mixture model, rather than the standard binomial or beta-binomial models to account for the extra process of agents being present or absent.

One thing to note at this point is that here we are paying careful attention to the possible processes that may drive patterns in the data, and this is critical in informing the choice of mixture model and its interpretation. I will return to this issue in the Discussion, but I mention it here to stress that this is an important habit to get into, to ensure that your choices of models are ecologically sensible.

The first thing that we will do now is explore the relative support from the data for four possible decay models: (1) binomial; (2) beta-binomial; (3) zero-inflated binomial; and (4) zero-inflated beta-binomial. The first model is the standard binomial model, with the other three models being mixture models representing different mechanisms through which overdispersion may arise, as discussed above. The binomial model has one parameter, *s*, representing the expected daily pellet survival probability (with 1 – *s* being the decay rate). The beta-binomial model has an additional parameter, γ, that controls the level of overdispersion, with low values of γ representing high levels of overdispersion and high values of γ representing low levels of overdispersion. The two zero-inflated models then have a further parameter, *q*, representing the level of zero-inflation, with low values of *q* representing high levels of zero-inflation and high values of *q* representing low levels of zero-inflation. The formulations of the likelihoods of these models are described in Appendix B. In the context of this empirical example, we interpret the parameter *q* as the probability that agents causing decay (e.g., insect, animal and/or bacterial activity) are present.

In addition to estimating the support for each of these models, we are also interested in whether pellet survival rates, *s*, and the probability that decay agents are present, *q*, vary with environmental factors or remain roughly constant. To incorporate covariates we model *s* and *q* as functions of covariates, rather than assuming they are constant. Specifically, we model them using the standard logit link function, such that the daily survival probability of a pellet in group *i* on day *t* of interval *j* is

$$s_{ijt} = \frac{\exp\left(\alpha + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_{ijt}\right)}{1 + \exp\left(\alpha + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_{ijt}\right)}, \tag{12.7}$$

where α is an intercept, $\boldsymbol{\beta}$ is a vector of regression coefficients, and $\mathbf{X}_{ijt}$ is a vector of covariates for pellet group *i* in interval *j* on day *t*. We model the probability that decay agents are present in pellet group *i* in interval *j* is

$$q_{ij} = \frac{\exp\left(\eta + \boldsymbol{\upsilon}^{\mathrm{T}}\mathbf{Y}_{ij}\right)}{1 + \exp\left(\eta + \boldsymbol{\upsilon}^{\mathrm{T}}\mathbf{Y}_{ij}\right)}, \tag{12.8}$$

where η is an intercept, $\boldsymbol{\upsilon}$ is a vector of regression coefficients, and $\mathbf{Y}_{ij}$ is a vector of covariates for pellet group *i* in interval *j*.

We can construct expressions for the likelihoods for each of these models (see Appendix B for details) and then use the function `mle2` from the package `bbmle` in R to find the maximum likelihood parameter estimates. The `mle2` function accepts, as one of its arguments, a function for the negative log-likelihood and then finds the parameter values that minimize this function using numerical optimization (note that minimizing the negative log-likelihood is identical to maximizing the log-likelihood, so this finds the maximum likelihood estimates of the model parameters). Although this can also be achieved by using the `optim` function (see Chapter 3), the `mle2` function provides additional functionality, such as the generation of standard errors for the parameter estimates that is very useful. In Appendix C, I provide R code for the likelihood-functions and example code for fitting the models using the `mle2` function.

If we fit the models described above with all covariates (site, topography, pellet age, rainfall, humidity and temperature) as predictors of s_{ijt}, but with *q* constant initially (i.e., assuming that environmental variables determine decay rates, but not the presence or absence of decay agents), the best-supported model (based on Akaike's Information Criteria (AIC) – see Chapter 3) is the beta-binomial model, but the zero-inflated beta-binomial model also has considerable support (having an AIC only 1.2 units larger than the beta-binomial model) (Table 12.3). On the other hand, the binomial and zero-inflated binomial models have almost no support from the data, with AIC values much greater than either of the two best models. This provides strong evidence that random variation in pellet survival is a key process driving overdispersion (i.e., both of the top two models contain the beta-binomial mixture representing random variation in pellet survival). However, there is some evidence that the presence or absence of decay agents may operate together with random variation in survival rates (i.e., the zero-inflated beta-binomial model also has support relative to the beta-binomial model). If we plot the quantile-quantile plots of residuals for the binomial and beta-binomial models (Figure 12.2), we can see that the beta-binomial model adequately accounts for overdispersion in the data, but the binomial model does not (i.e., Figure 12.2A shows the characteristic quantile-quantile plot shape for overdispersed data). Moreover, the standard errors of the coefficient estimates for the beta-binomial models are larger than for the binomial models, which is expected because it is accounting for overdispersion, and standard errors are no longer underestimated (Table 12.3).

So far we have been able to account for overdispersion and say something about the relative support for each hypothesized mechanism driving overdispersion. But now let us look at predictors of pellet survival rates and the presence of decay agents. We will do this for the two best supported models: the beta-binomial and the zero-inflated beta-binomial models. The variables hypothesized to be potentially important drivers of pellet decay include spatial variables (site and topography) and temporal variables (pellet age, rainfall, humidity, and temperature). A sensible question therefore might be to ask is, "What do the data tell us about the importance of spatial versus temporal variables?" For the beta-binomial model we can ask this by constructing models (through Equation 12.7) that contain: none of the variables, either the spatial or temporal variables, or both and comparing the four resulting models using AIC. For the zero-inflated beta-binomial model, however, there is the possibility the variables may determine the expected survival rate, s, and/or the probability that decay agents are present, q. In this case, there are sixteen possible combinations of models representing the different ways in which the spatial and temporal variables could influence s and q (through Equations 12.7 and 12.8) and the support for each of these models can also be explored using AIC. Note that, in constructing these models I always include pellet age as a covariate for s, but never include it as a covariate for q (since there is no reason to expect pellet age to determine whether decay agents are present or not), and the temporal covariates for q are quantified based on their mean values within each time interval.

So what does this tell us? For the beta-binomial model, there is very strong indication that both temporal and spatial variables drive pellet decay (Table 12.4). For the zero-inflated beta-binomial model spatial and temporal variables driving pellet decay (variables for *s*) is still strongly supported, but there is also strong evidence that the temporal climatic variables are important determinants of whether decay agents are present for not (variables for *q*; Table 12.4). Interestingly, in this case, the best zero-inflated beta-binomial model has a considerably lower AIC than the best beta-binomial model (2,345.64 versus 2,360.60). Hence, once we include covariates for both *s* and *q*, there is compelling evidence for two different processes operating to drive pellet decay; one that determines whether decay agents are present and one that determines the decay rate if decay agents are present.

This example provides an illustration of the power of mixture models to account for overdispersion and to allow inferences about the processes that drive that overdispersion. By grounding our model construction explicitly in terms of hypotheses about the ecological mechanisms that drive overdispersion, we are able to say something useful about the support for each of those mechanisms. This may be particularly important here since it appears that the drivers of whether decay agents (e.g., insect, animal and/or bacterial activity) are present may be different from the drivers of decay rates if decay agents are present. The development of approaches for using these types of models to reliably calibrate surveys of indirect signs will depend on being able to correctly identify and quantify the processes that drive the decay process. Mixture models are an important tool for helping us to do this.

12.3.2 Using Poisson mixtures to model lemur abundance

Count data are one of the most commonly collected types of data for estimating species' distributions and abundance. However, as I have already pointed out, these data commonly exhibit overdispersion that precludes analysis based on standard distributions. Although both ecological processes and observation errors can lead to overdispersion in these types of data (Table 12.1), correct inference relies on distinguishing between these two sources of overdispersion. In the context, Royle (2004) demonstrates how to account for detection errors in count data by using so called *N*-mixture models. These are countable mixture models that are explicit about zero-inflation arising from detection errors and the distribution of the true underlying abundances (which can also be represented by a mixture model if necessary). For example, Royle (2004) adopt the Poisson distribution to describe the true underlying abundances, but also illustrate how the negative-binomial may be used instead so that the model accounts for both zero-inflation that arises due to detection errors and overdispersion in the underlying abundances. Wenger and Freeman (2008) extend the approach to allow the true underlying abundances to be described by zero-inflated models. This allows for the possibility of simultaneously representing zero-inflation that arises from observation errors and zero-inflation that arises from ecological processes in the same model.

In this second empirical example, I illustrate the use of *N*-mixture models to make inferences about the abundance of two lemur species (the common brown lemur *Eulemur fulvus fulvus* and the black and white ruffed lemur *Varecia variegata variegata*) at two sites in the Zahamena Reserve in eastern Madagascar. I use this example to illustrate how we can construct mixture models to account for and make inferences about overdispersion that arises from both observation and ecological sources. You will see as we go through the example that once again, thinking carefully about the sources of overdispersion is central to successful model construction. I will once again use mixture models to try to distinguish between two sources of overdispersion; one that relates to overdispersion arising from observation error and one that relates to overdispersion arising from an ecological process. I will show that source of overdispersion has profound implications for ecological inference. This is because, in the case of the observation error process, we actually want to 'strip out' the effect of that process so as to reduce bias in ecological inference, while in the case of the ecological process, we are interested in the process itself and it is therefore retained as a component of ecological inference.

The data I use were collected in 1999 and 2000 at two sites in the Zahamena reserve in eastern Madagascar; one in mid-altitude rainforest (Antenina; elevation 900 m) and one in lowland rainforest (Namarafana; elevation 450 m). The data consists of direct group counts of lemurs along

300 m or 400 m long transect sections at each site. Although data were collected on all lemur species at the sites, I will only focus here on counts of groups of *Eulemur f. fulvus* and *Varecia v. variegata* (Appendix D) and we will aim to quantify differences in abundance for these species between the two sites. Once again, let us start by looking at histograms of the data (Figure 12.3). Histograms of the raw data reveal substantial zero-inflation, but there is also some suggestion of an excess of high values too. Overdispersion is also indicated by the variance of the data relative to the variance of the expected values (0.056 versus 0.016 for *Eulemur f. fulvus* and 0.091 versus 0.018 for *Varecia v. variegata*). However, it is unclear on inspection of the histograms whether the overdispersion occurs primarily due to observation error, or as a result of an ecological process, such as a highly clumped spatial distribution. Understanding this is critical because it will likely make a major difference to our interpretation. We will now begin to explore these issues starting with a simple model and then adding complexity.

The simplest way to model these data is to ignore any observation error and overdispersion and use a standard Poisson distribution with one parameter, λ, representing the mean. However, it would make sense to try to account for the zero-inflation in some way. A straightforward way to accommodate the zero-inflation is to use a zero-inflated Poisson model with two parameters, λ and *q*. If we assume that there is no observation error then we can interpret *q* as the probability that habitat is suitable and then λ is the mean abundance, given that habitat is suitable (remember we discussed this idea earlier in the chapter). However, since we also seem to have an excess of high values in the data, as well as zero-inflation, it could also make sense to extend this to the zero-inflated negative-binomial model which has one further parameter, κ, that represents the level of overdispersion in the negative-binomial component of the mixture (high values of κ imply high levels of overdispersion and low values κ imply low levels of overdispersion).

The likelihoods for the Poisson, zero-inflated Poisson, and zero-inflated negative-binomial models are described in Appendix E and can, once again, be fitted to the data using the function `mle2`. As in the first empirical example we can make the model parameters functions of covariates.

We are interested in the difference in abundance between sites so it would make sense to introduce a covariate for site. However, because total survey effort varies between transect sections, we need to control for this by incorporating survey effort as a covariate too. Survey effort can be controlled for in a simple way: let λ (mean abundance, given suitable habitat in the case of the zero-inflated models) depend on survey effort, with $\lambda = \gamma S$, where $S > 0$ is the survey effort and $\gamma > 0$ is the expected count per unit of survey effort (given suitable habitat in the case of the zero-inflated models). Then, to introduce the site covariate, we model γ and q (the probability that habitat is suitable) as functions of the site using the standard log and logit link functions respectively, such that

$$\gamma_i = \exp(\alpha + \beta X_i), \tag{12.9}$$

where γ_i is the expected count per unit of survey effort (our index of abundance) for transect section *i*, α is an intercept, β is a regression coefficient, and X_i is a categorical covariate representing the site within which transect section *i* is located. Finally, let

$$q_i = \frac{\exp(\eta + \upsilon X_i)}{1 + \exp(\eta + \upsilon X_i)}, \tag{12.10}$$

where q_i is the probability that the habitat in transect section *i* is suitable, η is an intercept, υ is a regression coefficient, and X_i is a categorical covariate representing the site within which transect section *i* is located. In Appendix F, I provide R code for the likelihood-functions and example code for fitting the models using the `mle2` function.

Fitting the Poisson model to the data and inspecting the quantile-quantile plots of the residuals reveals high levels of overdispersion in the data for both species, with the characteristic pattern of too many low values and too many high values in the data (Figures 12.4A and 12.4C). The zero-inflated Poisson model reduces the level of overdispersion, but some points in the quantile-quantile plot still lie outside the 95% confidence intervals, suggesting some remaining overdispersion. On the other hand, the zero-inflated negative-binomial model adequately accounts for overdispersion, with the quantile-quantile plot lying close to the expected 1:1 line and within the 95% confidence intervals (Figures 12.4B and 12.4D). Therefore, a model whereby overdispersion is represented by both zero-inflation (representing whether habitat is suitable or not) and heterogeneity among transect sections (represented by the negative-binomial component of the mixture) appears to be adequate for accounting for the overdispersion.

Although we have accounted for overdispersion here through both zero-inflation and heterogeneity among sections and the model seems to fit well, we have not considered the possibility that the zero-inflation may arise due to detection error (i.e., where the probability of detecting a species, or individual, that it is present, is less than one), rather than through the processes of habitat being suitable or not. If detection errors are present, then our estimates of abundance will be biased if not accounted for, especially if detection errors vary between the two sites. In recent years there has been substantial progress made in the development of methods for dealing with detection errors in ecological data (MacKenzie et al. 2006, Royle and Dorazio 2008). In general, these methods have, at their core, a mixture model that enables the explicit representation of zero-inflation or missed counts, arising from the failure to detect individuals that are actually present. For example, MacKenzie et al. (2002) and Tyre et al. (2003) use zero-inflated binomial models to estimate occupancy while accounting for a failure to detect occupancy, thus reducing bias in occupancy estimates. However, to distinguish false-negatives (i.e., a failure to detect individuals that are actually present) from true-negatives (i.e., true absences) requires repeat surveys of sites within a short enough time period that the true occupancy or abundance state can be assumed to be unchanged. Fortunately, the lemur data consists of repeat surveys of each transect sections and therefore we can take advantage of this to explicitly account for detection error and reduce bias in abundance estimates.

Earlier I mentioned Royle (2004)'s *N*-mixture model for dealing with detection errors in count data and we are going to use this model to examine the implications of detection error on our inferences about abundance at the two sites. The details of the likelihood for an *N*-mixture model are given in Appendix E, but I will describe the model briefly here. The model is a countable mixture model based on an observation process defined by a binomial distribution that represents the probability of detecting an individual given that it is present. In the binomial distribution, the binomial probability, *p*, represents the probability of detection, while the number of trials, *N*, represents the true number of groups present at a site and this is assumed to vary randomly according to a Poisson distribution with mean λ (although other distributions, such as the negative-binomial are also possible). Covariates for *q* and λ can be included in a similar way to Equations 12.9 and 12.10. In these mixture models we are explicit about the observation process, via the binomial distribution, and explicit about the true underlying abundance, via the Poisson or negative-binomial distribution. I provide R code for the likelihood functions of the *N*-mixture models that I use here and example code for fitting these models to the lemur count data using the `mle2` function in Appendix F.

The interpretation of *N* and *p* is worth a note here before moving on. We interpret *N* for this case study as the number of groups that use a transect section, rather than the usual interpretation that would be the number of groups present in a transect section at the time of survey. Due to the mobile nature of the species, the number of groups present on a transect section may be different from day to day. An important assumption of these models is that the state of the system does not change between repeat surveys (an assumption known as the closure assumption) and this is clearly broken here possibly leading to bias (Kendall and White 2009, Rota et al. 2009). This is because detection errors can occur for two reasons that are confounded in the estimate of *p*: (1) a group may be present at the time of a survey, but not observed and (2) a group that uses the section my not be present at the time of the survey. However, if we interpret *p* as the probability that we detect a group that uses a transect section, rather than the probability that we detect a group that is present at the time of the survey, this issue is resolved. This is because the number of groups that used a transect section over the study period would have been relatively constant and so by making inferences at this level, the closure assumption holds and the confounding of sources of detection error does not matter. This also means, however, that we must interpret *N* as the number of groups using a transect section over the study period, rather than being the number of groups present in a transect section at the time of the survey.

We will now look at to what extent the use of *N*-mixture models (i.e., assuming that the zero-inflation arises due to observation error), as opposed to using the zero-inflated Poisson or zero-inflated negative-binomial models (i.e., assuming that the zero-inflation arises due to the availability of habitat) modify our conclusions about differences in abundance between the two sites. If we fit both Poisson and negative-binomial N-mixture models to the lemur data assuming that both detection errors and abundance can vary between sites (i.e., we include a site covariate on p and λ) and compare these models to the zero-inflated models we see a number of key differences (Table 12.6). The first thing to note is that, although the negative-binomial distribution has better support than the Poisson distribution for the zero-inflated models (based on AIC), this is not necessarily the case for the *N*-mixture models. The second thing to note is that, although for *Eulemer f. fulvus* the zero-inflated models suggest that abundance is lower at the lowland site that the mid-elevation site (although not significantly so, based on the standard error estimate), the *N*-mixture models suggest that abundance is greater at the lowland site than at the mid-elevation site. This is because, although sighting rates are lower at the lowland site, the *N*-mixture model estimates that probability of detection is much lower at the lowland site than at the mid-elevation site. The lower probability of detection more than compensates for the lower sighting rates at the lowland site, resulting in a higher estimate of abundance. For *Varecia v. variegata* the two types of model are in agreement, with abundance estimated to be higher in the lowland than the mid-elevation site, but the probability of detection is similarly estimated to be lower in the lowland than the mid-elevation site.

This example shows that our assumptions about sources of overdispersion can have profound implications for the inferences we make. In developing our inferences in this case, we need to make a decision about whether we believe that zero-inflation arises through observation error, or through some ecological processes related to the availability of habitat. It is unlikely in this example that observation error is zero; it is almost certainly the case that groups that are present on a transect section could have been missed and the mobile nature of the species means that a group that uses a transect section may not be present at the time of survey. There are two possible reasons for detection errors being higher in the lowland site than the mid-elevation site. The first reason is that groups present on the transect sections are not detected more often at the lowland than the mid-elevation site. The lowland site has a more dense understory and higher canopy than the mid-elevation site (J. Rhodes, personal observation), which would tend to make lemur observations more difficult, so this is consistent with the *N*-mixture models. However, this could also be driven by differences in field personnel between the two sites. The second reason, is that groups that use a transect section are more often absent from a transect section at the lowland site than the mid-elevation site. This could occur, for example, if groups tend to move more frequently at the lowland site than the mid-elevation site. Although we have no information about the relative movement frequencies at the two sites, the *N*-mixture models make sense in terms of the likely presence of detection errors and variation in forest structure and personnel between the two sites. Nonetheless, in this example, a mixture modeling approach has allowed us to be explicit about the mechanisms driving overdispersion and, importantly, to understand the implications of the assumptions we have made.

12.5 Discussion

Mixture models should become a critical component of the ecologist's statistical tool box. Ecological data arise from complex interacting processes; they rarely conform nicely to the assumptions of standard statistical distributions. When they do not, this often manifests itself as overdispersion, playing havoc with our statistical tests and inference. Fortunately, mixture models provide a flexible way to deal with overdispersion, but they are useful for much more than simply controlling for overdispersion. This is because they allow us to make inferences about the causes of overdispersion, leading to greatly improved ecological inference. In particular, it allows us to have a much more mechanistic understanding of the processes that lead to the observed data. In this chapter I have outlined what mixture models are and illustrated their use in two quite different applications. The applications demonstrate how careful consideration of the mechanisms driving overdispersion in the data and can lead to a much richer understanding of the underlying ecological and observation processes. Although the applications I have presented come from survival analysis and abundance estimation, mixture models are applicable to almost any area of ecology. As such, they are an important and widely applicable approach in ecological statistics.

In this chapter I have focused on some of the more typical and standard mixture models. However, it is possible to more generally construct complex mixtures of distributions to represent a wide range of ecological mechanisms that may be hypothesized to generate any observed data. For example, I mentioned earlier, Kendall and Whittmann (2010)'s stochastic model of breeding success that explicitly models the probability of laying eggs, nest survival, clutch size and offspring survival as mechanisms leading to the observed data on reproductive output. They apply it to 53 vertebrate species and to achieve this, they model the number of offspring as a finite mixture distribution, with mixing weights defined by the probability that eggs are laid and then model the probability of nest survival, given eggs are laid, that is itself mixture model. The model for nest survival reflects the contribution of clutch size and offspring survival to the number of offspring and is assumed to be a countable mixture with offspring survival defined by a binomial distribution and the number of trials specified by a Poisson distribution (in a similar way to an *N*-mixture model). This provides inference about these separate component processes that would otherwise not be possible without the use of a mixture model. More broadly, flexible mixture models form the basis of so-called state-space models that aim to represent ecological and observation processes in a mechanistic fashion (for nice examples see Buckland et al. 2004, Patterson et al. 2008). Specifying these models often results in complex mixtures, but because they are explicit about the ecological and observation processes that generate the observed data, they provide a powerful and flexible framework for ecological inference that is becoming increasingly popular.

I have demonstrated how we should ground our choice of mixture model in mechanistic hypotheses about the processes that may have led to the data. An alternative is to adopt a, so called, quasi-likelihood approach. Rather than characterizing the full likelihood of the data, quasi-likelihood approaches characterize a quasi-likelihood function that depends only on the mean and variance, but behaves in a similar way to the full likelihood (see Chapter 6; McCullagh and Nelder 1989, Burnham and Anderson 2002). Essentially, what this means is that the quasi-likelihood does not characterize the full distribution of the data, but simply adjusts the variance to account for overdispersion. In contrast, mixture models use information about the full distribution of the data and this is what allows us the make more mechanistic inferences that would not necessarily be possible using quasi-likelihood methods. For example, in the koala dung decay example, our mixture model uses the amount of zero-inflation in the data to distinguish between the processes driving the presence of decay agents versus processes driving decay rates where decay agents are present. This type of analysis would not be possible with a quasi-likelihood; inferences about the presence of decay agents would not be possible, although it would still control for overdispersion allowing us to perform correct statistical tests.

Although inference about mechanisms is a major strength of the mixture modeling approach it can also be problematic if we have no, or little, *a priori* information about potential causes of overdispersion. However, in cases where we are not specifically interested in the causes of overdispersion, or are unable to develop sensible mixture models, then quasi-likelihood approaches are often a suitable alternative to mixture models for dealing with overdispersion. In fact, Richards (2008) shows that a quasi-likelihood approaches to model selection produce very similar results to the negative-binomial mixture model based on Akaike's Information Criteria (AIC). None the less, we still need to think critically about our choice of model and model assumptions because this can have important bearing on inference. For example, Ver Hoef and Boveng (2007) show that quasi-

Poisson (a quasi-likelihood version of the Poisson distribution) and negative-binomial models can produce quite different parameter estimates, using an example of harbor seals in Alaska. They show that regression coefficient estimates are affected by the choice of model because they make different assumptions about how the variance of the data changes with abundance. The quasi-Poisson model assumes that the variance increases linearly with abundance, while the negative-binomial model assumes that the variance increases quadratically with abundance. In their case, they find that the quasi-Poisson is the better model for their data, and they suggest plotting abundance versus variance of data to get an idea of which model may be most appropriate. The important take-home message here is that, even if we use a quasi-likelihood approach, we should ensure that the assumed variance-mean relationship is sensible for our data.

Despite the great promise of mixture-models, they should be used carefully and with some caution. One issue is that they make strong assumptions about the distribution of the data and if these assumptions do not hold this could result in biased parameter estimates. Therefore, I recommend that careful *a priori* consideration be given to the choice of assumed mechanisms as I have done in this chapter. First, you should think carefully about what the implications of failing to meet those assumptions might be. For example, in the lemur example, inference is strongly dependent upon whether you assume that zero-inflation arises from observation or ecological processes. Second, because mixture models often contain a large number of unobserved (latent) variables, model parameters can often fail to be identifiable. Model parameters are not identifiable when the data are insufficient to distinguish between the values for two or more parameters because their values are confounded. For example, in the lemur case study I was unable to fit an *N*-mixture model based on a zero-inflated Poisson distribution because there was insufficient information in the data to separate zero-inflation that arises due to detection error from zero-inflation that arises due to the availability of suitable habitat. This issue can limit the extent to which mixture models are able to be applied to specific ecological questions.

Mixture models are closely related to the mixed-effects (or random-effects) models that are commonly used in ecology (see Chapter 13). Mixed-effects models introduce random variation in model parameters (through the specification of random-effects) that can account for additional variation in the data in a similar way to mixture models. However, whereas mixture models introduce random variation at the level of individual data points, variation in mixed-effects models is usually specified at a hierarchical level above that of the individual data points. For example, in a mixed-effects model we may have random variation among sites, but not among data points within sites as in a mixture model. For this reason, mixed-effects models are most often used to account for hierarchical structure, or dependencies, in the data rather than overdispersion. For example, Thomas et al. (2006) use individual-level random-effects to model variation in habitat selection among individuals in their analysis of caribou location data. The purpose for doing so was to account for dependencies in the data within individuals and variation among individuals, rather than dealing with and understanding overdispersion per se. Therefore, despite the close links between the two approaches, their use in ecology is quite different.

What should you report in a paper using mixture models? One of the most critical aspects is to be clear about how you constructed your mixture models, the mechanisms you hypothesize the different components of the mixture represent, and what assumption you have made. In describing your models this is critical so that the reader understands what your models represent. In this context, Kendall and Whittmann (2010) provides an excellent example where the rationale for the model is very clearly described. You should also present evidence that your models have dealt adequately with overdispersion in the data using techniques such as quantile-quantile plots as I have used in this chapter. Finally, providing inference in terms of the different components of the mixture model is important so that readers can relate this inference back to the proposed mechanisms. For example, if you use a negative binomial distribution to deal with random variation in habitat quality then, in addition to reporting the regression parameters, report and interpret the overdispersion parameter in the terms of the hypothesized source of overdispersion. This is will provide readers with a richer understanding of the ecological processes that would be possible if the overdispersion parameter was not interpreted in this way.

Acknowledgements

The writing of this chapter was funded in part by the Australian Research Council Centre of Excellence for Environmental Decisions. Thank you to Dan Lunney for providing access to the koala dung decay trial data.

Summary for online only (120-150 words)

When the variance of data has a higher variance than that of standard distributions used in statistical tests this is known as overdispersion. Overdispersion is ubiquitous in ecological data, leading to the underestimation of variances and bias in statistical tests unless the overdispersion is accounted for. Consequently, having methods for dealing with overdispersion is an essential component of the ecologist's statistical toolbox. A powerful approach for dealing with overdispersion are mixture models; it is powerful because it allows us to be explicit about the processes that drive overdispersion in the data and enabling a deeper understanding of ecological processes. In this chapter I introduce mixture models and illustrate their approach using examples from survival analysis and the analysis of population abundance. I specifically focus on demonstrating how mixture models can both account for overdispersion and allows ecological inferences that would not otherwise be possible without the use of a mixture model approach.

Keywords for online only (5-10 words)

overdispersion, mixture models, detection error, finite mixture, countable mixture, continuous mixture, mechanistic model

References

Anderson, D. R., K. P. Burnham, and G. C. White. 1994. AIC model selection in overdispersed capture-recapture data. Ecology **75**:1780-1793.

Barry, S. and J. Elith. 2006. Error and uncertainty in habitat models. Journal of Applied Ecology **43**:413-423.

Buckland, S. T., K. B. Newman, L. Thomas, and N. B. Koesters. 2004. State-space models for the dynamics of wild animal populations. Ecological Modelling **171**:157-175.

Burnham, K. P. and D. R. Anderson. 2002. Model selection and multimodel inference: a practical information-theoretic approach. Second edition. Springer-Verlag, New York, USA.

Calabrese, J. M., J. L. Brunner, and R. S. Ostfeld. 2011. Partitioning the aggregation of parasites on hosts into intrinsic and extrinsic components via an extended Poisson-gamma mixture model. Plos One **6**.

Clark, J. S., M. Silman, R. Kern, E. Macklin, and J. HilleRisLambers. 1999. Seed dispersal near and far: Patterns across temperate and tropical forests. Ecology **80**:1475-1494.

Colwell, R. K., C. X. Mao, and J. Chang. 2004. Interpolating, extrapolating, and comparing incidence-based species accumulation curves. Ecology **85**:2717-2727.

Cunningham, R. B. and D. B. Lindenmayer. 2005. Modeling count data of rare species: some statistical issues. Ecology **86**:1135-1142.

Dean, C. B. 1992. Testing for Overdispersion in Poisson and Binomial Regression Models. Journal of the American Statistical Association **87**:451-457.

Haining, R., J. Law, and D. Griffith. 2009. Modelling small area counts in the presence of overdispersion and spatial autocorrelation. Computational Statistics & Data Analysis **53**:2923-2937.

Hall, D. 2000. Zero-inflated Poisson and binomial regression with random-effects: a case study. Biometrics **56**:1030-1039.

Hinde, J. and C. G. B. Demetrio. 1998. Overdispersion: models and estimation. Computational Statistics & Data Analysis **27**:151-170.

Johnson, N. L., A. W. Kemp, and S. Kotz. 2005. Univariate Discrete Distributions. John Wiley & Sons, Hoboken, USA.

Kendall, B. E. and M. E. Wittmann. 2010. A stochastic model for annual reproductive success. American Naturalist **175**:461-468.

Kendall, W. L. and G. C. White. 2009. A cautionary note on substituting spatial subunits for repeated temporal sampling in studies of site occupancy. Journal of Applied Ecology **46**:1182-1188.

Lambert, D. 1992. Zero-inflated Poisson regression, with an application to defects in manufacturing. Technometrics **34**:1-14.

Landwehr, J. M., D. Pregibon, and A. C. Shoemaker. 1984. Graphical methods for assessing logistic regression models. Journal of the American Statistical Association **79**:61-71.

Linden, A. and S. Mantyniemi. 2011. Using the negative binomial distribution to model overdispersion in ecological count data. Ecology **92**:1414-1421.

MacKenzie, D. I., J. D. Nichols, G. B. Lachman, S. Droege, J. A. Royle, and C. A. Langtimm. 2002. Estimating site occupancy rates when detection probabilities are less than one. Ecology **83**:2248-2255.

MacKenzie, D. I., J. D. Nichols, J. A. Royle, K. H. Pollock, L. L. Bailey, and J. E. Hines. 2006. Occupancy Estimation and Modeling Elsevier, Burlington, USA.

Martin, T. G., B. A. Wintle, J. R. Rhodes, P. M. Kuhnert, S. A. Field, S. J. Low Choy, A. J. Tyre, and H. P. Possingham. 2005. Zero tolerance ecology: improving ecological inference by modelling the source of zero observations. Ecology Letters **8**:1235-1246.

McAlpine, C. A., J. R. Rhodes, M. E. Bowen, D. Lunney, J. G. Callaghan, D. L. Mitchell, and H. P. Possingham. 2008. Can multi-scale models of species' distribution be generalised from region to region? A case study of the koala. Journal of Applied Ecology **45**:558-567.

McCullagh, P. and J. Nelder. 1989. Generalized linear models. second edition. Chapman and Hall, London, UK.

McLachlan, G. and D. Peel. 2000. Finite Mixture Models. John Wiley & Sons, New York, USA.

Mengersen, K. L., C. P. Robert, and D. M. Titterington, editors. 2011. Mixtures: Estimation and Applications. John Wiley & Sons, Chichester, UK.

Mumby, P. J., R. Vitolo, and D. B. Stephenson. 2011. Temporal clustering of tropical cyclones and its ecosystem impacts. Proceedings of the National Academy of Sciences of the United States of America **108**:17626-17630.

Patterson, T. A., L. Thomas, C. Wilcox, O. Ovaskainen, and J. Matthiopoulos. 2008. State-space models of individual animal movement. Trends in Ecology & Evolution **23**:87-94.

Pledger, S. and C. J. Schwarz. 2002. Modelling heterogeneity of survival in band-recovery data using mixtures. Journal of Applied Statistics **29**:315-327.

Quintero, H. E., A. Abebe, and D. A. Davis. 2007. Zero-inflated discrete statistical models for fecundity data analysis in channel catfish, Ictalurus punctatus. Journal of the World Aquaculture Society **38**:175-187.

Rhodes, J., D. Lunney, C. Moon, A. Matthews, and C. A. McAlpine. 2011. The consequences of using indirect signs that decay to determine species' occupancy. Ecography **34**:141-150.

Rhodes, J. R., J. G. Callaghan, C. A. McAlpine, C. de Jong, M. E. Bowen, D. L. Mitchell, D. Lunney, and H. P. Possingham. 2008a. Regional variation in habitat-occupancy thresholds: a warning for conservation planning. Journal of Applied Ecology **45**:549-557.

Rhodes, J. R., E. P. M. Grist, K. W. H. Kwok, and K. M. Y. Leung. 2008b. A Bayesian mixture model for estimating intergeneration chronic toxicity. Environmental Science & Technology **42**:8108-8114.

Richards, S. A. 2008. Dealing with overdispersed count data in applied ecology. Journal of Applied Ecology **45**:218-227.

Rota, C. T., R. J. Fletcher, R. M. Dorazio, and M. G. Betts. 2009. Occupancy estimation and the closure assumption. Journal of Applied Ecology **46**:1173-1181.

Royle, J. A. 2004. *N*-mixture models for estimating population size from spatially replicated counts. Biometrics **60**:108-115.

Royle, J. A. and R. M. Dorazio. 2008. Hierarchical Modeling and Inference in Ecology. Academic Press, London, UK.

Thomas, D. L., D. Johnson, and B. Griffith. 2006. A Bayesian random effects discrete-choice model for resource selection: population-level selection inference. Journal of Wildlife Management **70**:404-412.

Tyre, A. J., B. Tenhumberg, S. A. Field, D. Niejalke, K. Parris, and H. P. Possingham. 2003. Improving precision and reducing bias in biological surveys: estimating false-negative error rates. Ecological Applications **13**:1790-1801.

Ver Hoef, J. M. and P. L. Boveng. 2007. Quasi-Poisson vs. negative binomial regression: how should we model overdispersed count data? Ecology (Washington D C) **88**:2766-2772.

Wenger, S. J. and M. C. Freeman. 2008. Estimating species occurrence, abundance, and detection probability using zero-inflated distribution. Ecology **89**:2953-2959.

Table 12.1. Main causes of overdispersion and, for each cause, an ecological example illustrated by abundance data (counts), together with typically what a histogram of the overdispersed data would look like relative to a Poisson distribution fitted to the data and what a quantile-quantile plot of the overdispersed data would look like relative to a Poisson distribution fitted to the data. See Haining et al. (2009) and Linden and Mantymiemi (2011) for useful further discussion of the causes of overdispersion.

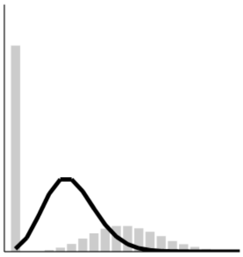
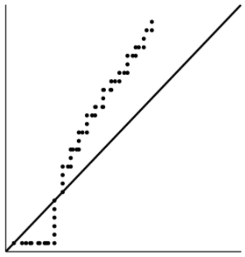
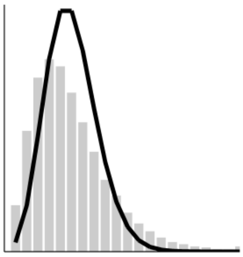
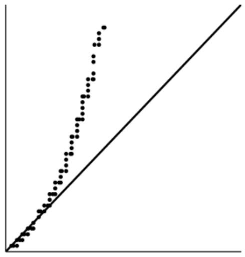
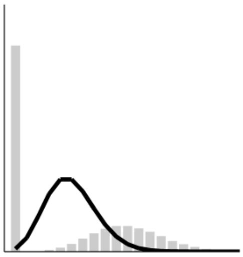
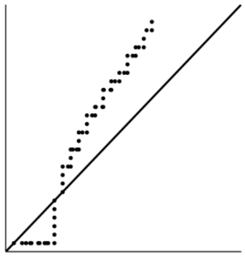
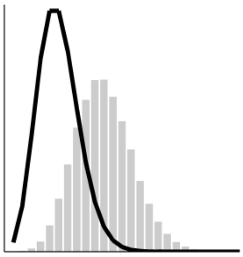
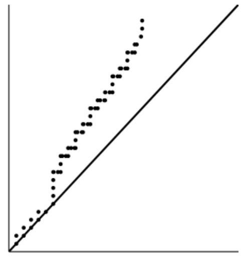
Cause of Overdispersion	Ecological Example	Histograms of Overdispersed Count Data Relative to the Poisson	Quantile-quantile Plots of Overdispersed Count Data Relative to the Poisson
Spatial/temporal clustering	A species only occurs in a small part of the landscape sampled		
Heterogeneity among sampling units	The expected abundance of a species varies randomly across a landscape based on variation in habitat quality		
Measurement error	Individuals that are truly present sometimes fail to be detected		
Misspecification of the mean	The relationship between abundance and habitat quality is actually non-linear but we model it as linear, resulting in underestimation of the mean for some habitat qualities		

Table 12.2. Commonly used probability distributions and some overdispersed mixture distribution equivalents. See Appendix LIKELIHOODS for formal definitions of the likelihood functions for these distributions.

Distribution	Equivalent Overdispersed Mixture Distributions
Normal	**Student's *t*-distribution** - *normal distribution with variance following an inverse gamma distribution (continuous mixture)*
Poisson	**Negative binomial distribution** - *Poisson distribution with the rate parameter following a gamma distribution (continuous mixture)* **Zero-inflated Poisson distribution** - *Poisson distribution with a Bernoulli distribution determining value of rate parameter of either zero or greater than zero (finite mixture)*
Binomial	**Beta-binomial distribution** - *binomial distribution with binomial probability parameter following a beta distribution (continuous mixture)* **Zero-inflated binomial distribution** - *binomial distribution with a Bernoulli distribution determining value of binomial probability parameter of either zero or greater than zero (finite mixture)*
Multinomial	**Dirichlet-multinomial distribution** - *multinomial distribution with multinomial probability parameters following a Dirichlet distribution (continuous mixture)*

Table 12.3. Akaike's Information Criteria (AIC) and coefficient estimates for the binomial (Bin), beta-binomial (BBin), zero-inflated binomial (ZIBin), and the zero-inflated beta-binomial (ZIBBin) models fitted to the koala dung decay data (standard errors shown in parentheses).

	Model			
Value	Bin	BBin	ZIBin	ZIBBin
AIC	3215.7	2360.6	2771.4	2361.8
ΔAIC	855.1	0.0	410.8	1.2
α	5.30 (0.110)	5.07 (0.174)	4.18 (0.125)	5.04 (0.181)
β_{age}	0.02 (0.001)	0.01 (0.001)	0.01 (0.001)	0.01 (0.001)
β_{site}	-1.19 (0.096) 0.14 (0.104) 0.71 (0.083)	-1.19 (0.162) 0.31 (0.168) 0.59 (0.13)	-0.89 (0.103) -0.04 (0.113) 0.58 (0.094)	-1.21 (0.164) 0.31 (0.170) 0.60 (0.136)
β_{mid}	-0.59 (0.088)	-0.53 (0.146)	-0.32 (0.102)	-0.53 (0.148)
β_{gully}	-0.52 (0.088)	-0.59 (0.144)	-0.14 (0.099)	-0.59 (0.146)
β_{rain}	-0.01 (0.001)	-0.01 (0.002)	-0.01 (0.002)	-0.01 (0.002)
β_{hum}	-0.08 (0.008)	-0.08 (0.013)	-0.06 (0.008)	-0.08 (0.0127)
β_{temp}	-0.12 (0.014)	-0.08 (0.021)	-0.13 (0.014)	-0.09 (0.022)
log(γ)		0.52 (0.098)		0.57 (0.119)
logit(q)			0.13 (0.099)	3.31 (1.316)

ΔAIC = difference between model AIC and model with the lowest AIC; α = intercept; β_{site} = coefficients for the sites; β_{mid} = coefficient for mid-slope; β_{gully} = coefficient for gully; β_{age} = coefficient for pellet age; β_{rain} = coefficient for rainfall; β_{hum} = coefficient for humidity; β_{temp} coefficient for temperature; log(γ) = logarithm of the overdispersion parameter in the beta-binomial distribution; and logit(q) = logit of the probability that decay agents are present for the zero-inflated models. Continuous covariates were centered based on: median age = 71 days, median rainfall = 0 mm, median humidity = 71.50%, and median temperature = 19.48 Celsius.

Table 12.4. Akaike's Information Criteria (AIC) values for each alternative beta-binomial model fitted to the koala dung decay data.

Model Rank	Spatial Variables	Temporal variables	AIC	ΔAIC
1	X	X	2,360.60	0.00
2	X		2,436.30	102.70
3		X	2,499.94	139.34
4			2,583.33	222.73

X = variables present in the model and ΔAIC = difference between model AIC and model with the lowest AIC. Continuous covariates were centered based on: median age = 71 days, median rainfall = 0 mm, median humidity = 71.50%, and median temperature = 19.48 Celsius.

Table 12.5. Akaike's Information Criteria (AIC) values for each alternative zero-inflated beta-binomial model fitted to the koala dung decay data.

Model Rank	Covariates for *s*		Covariates for *q*		AIC	ΔAIC
	Spatial	Temporal	Spatial	Temporal		
1	X	X		X	2,345.64	0.00
2	X	X	X	X	2,354.72	9.08
3	X	X			2,361.77	16.13
4	X	X	X		2,366.30	20.66
5	X		X	X	2,414.09	68.45
6		X	X		2,430.38	84.74
7		X	X	X	2,431.38	85.74
8	X			X	2,442.54	96.90
9	X		X		2,455.50	109.86
10	X				2,465.30	119.66
11			X	X	2,472.17	126.53
12		X		X	2,484.54	138.90
13			X		2,519.61	173.97
14				X	2,551.20	205.56
15			X		2,558.11	212.47
16					2,585.34	239.70

X = variables present in the model ("Variables for *s*" represent explanatory variables for the survival rates and "Variables for *q*" represent explanatory variables for the probability that decay agents are present) and ΔAIC = difference between model AIC and model with the lowest AIC. Continuous covariates for *p* were centered based on: median age = 71 days, median rainfall = 0 mm, median humidity = 71.5%, and median temperature = 19.48 Celsius. Continuous covariates for *q* were centered based on: median rainfall = 2.87 mm, median humidity = 71.02%, and median temperature = 19.79 Celsius.

Table 12.6. Model coefficients, Akaike's Information Criteria (AIC) values and the estimated difference in abundance between the lowland and mid-elevation site for each of the zero-inflated and N-mixture models fitted to the lemur count data (values in parentheses are standard errors).

Model	Coefficients for p or q		Coefficients for γ		$\ln(\kappa)$	AIC	$\hat{N}_{lowland} -$ $\hat{N}_{mid-elevation}$ (groups km^{-1})
	Intercept	Lowland site	Intercept	Lowland site			
Eulemer f. fulvus							
Zero-inflated Poisson	1.522 (0.493)	19.616 (2 x 10^{-10})	-0.986 (0.010)	-1.632 (0.028)		197.77	-0.233 (0.029)
Zero-inflated negative-binomial	1.770 (0.659)	11.935 (715.179)	-1.030 (0.141)	-1.575 (0.323)	1.624 (0.728)	194.86	-0.175 (0.168)
N-mixture (Poisson)	-2.335 (0.297)	-2.767 (0.741)	1.253 (0.284)	1.244 (0.682)		738.02	8.645 (2.629)
N-mixture (negative-binomial)	-2.336 (0.297)	-2.767 (0.741)	1.253 (0.284)	1.244 (0.682)	31.193 (2 x 10^{-10})	740.02	8.645 (2.629)
Varecia v. variegata							
Zero-inflated Poisson	-0.028 (0.366)	22.375 (2 x 10^{-10})	-1.044 (0.129)	0.143 (0.171)		222.72	0.233 (0.156)
Zero-inflated negative-binomial	0.157 (0.423)	15.460 (615.348)	-1.132 (0.221)	0.267 (0.295)	0.937 (0.454)	204.94	0.247 (0.374)
N-mixture (Poisson)	-1.805 (0.230)	-0.550 (0.450)	0.202 (0.263)	1.375 (0.444)		746.50	3.617 (1.172)
N-mixture (negative-binomial)	-1.892 (0.274)	-0.994 (0.514)	0.279 (0.302)	1.798 (0.523)	1.457 (0.828)	747.09	6.659 (1.741)

p = probability of suitable habitat in the zero-inflated models, q = detection probability in the N-mixture models, κ = the overdispersion parameter for the negative-binomial distribution, and $\hat{N}_{lowland} - \hat{N}_{mid-elevation}$ = the difference in estimated abundance between the lowland site and mid-elevation site in units of groups km^{-1} (standard errors for the differences were estimated using the delta method)

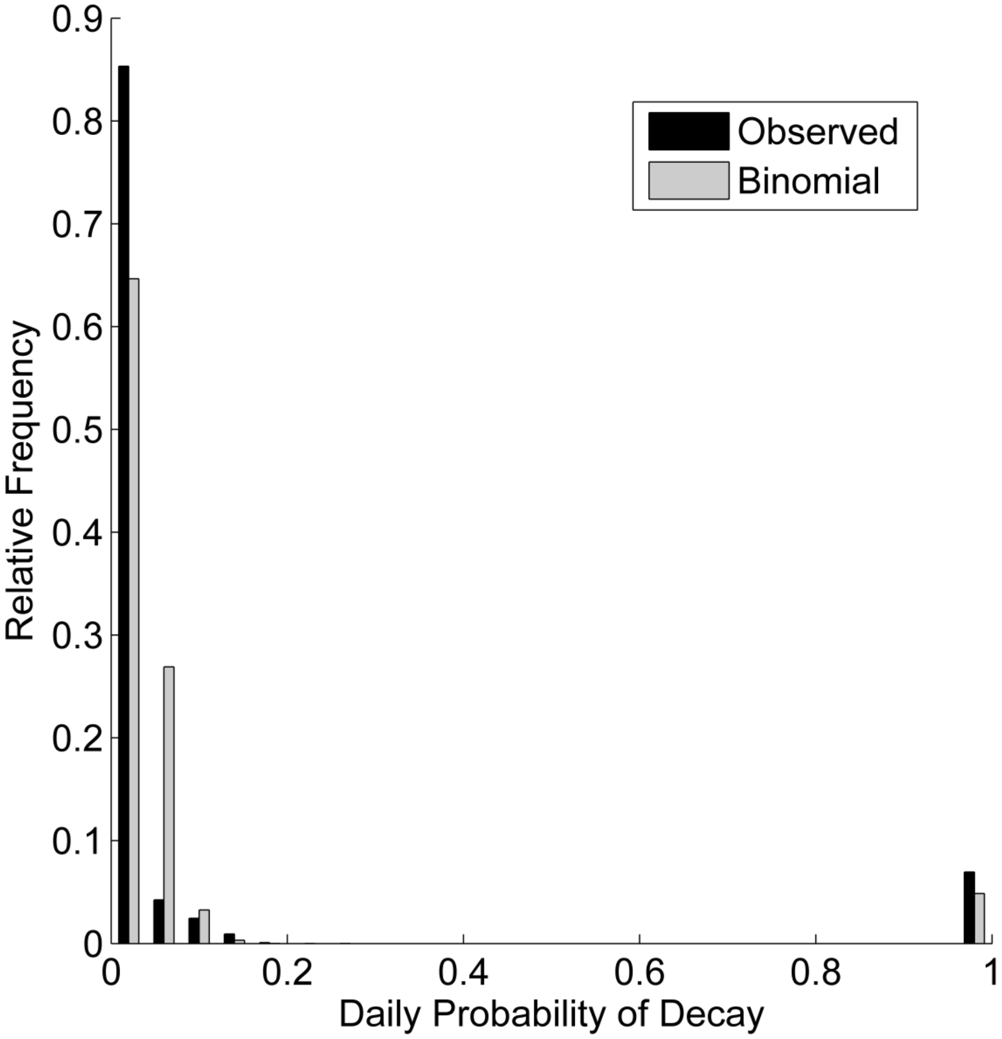


Figure 12.1. Histogram of observed daily probabilities of a koala pellet decay (black bars) versus expected values based on a binomial distribution with the same mean as the observed data (grey bars). Daily probabilities of decay relating to each recording interval were calculated as $1-(s/n)^{1/t}$, where *s* = the number of pellets that survived the interval (observed or expected), *n* = number of pellets at the start of the interval, and *t* = number of days in the interval.

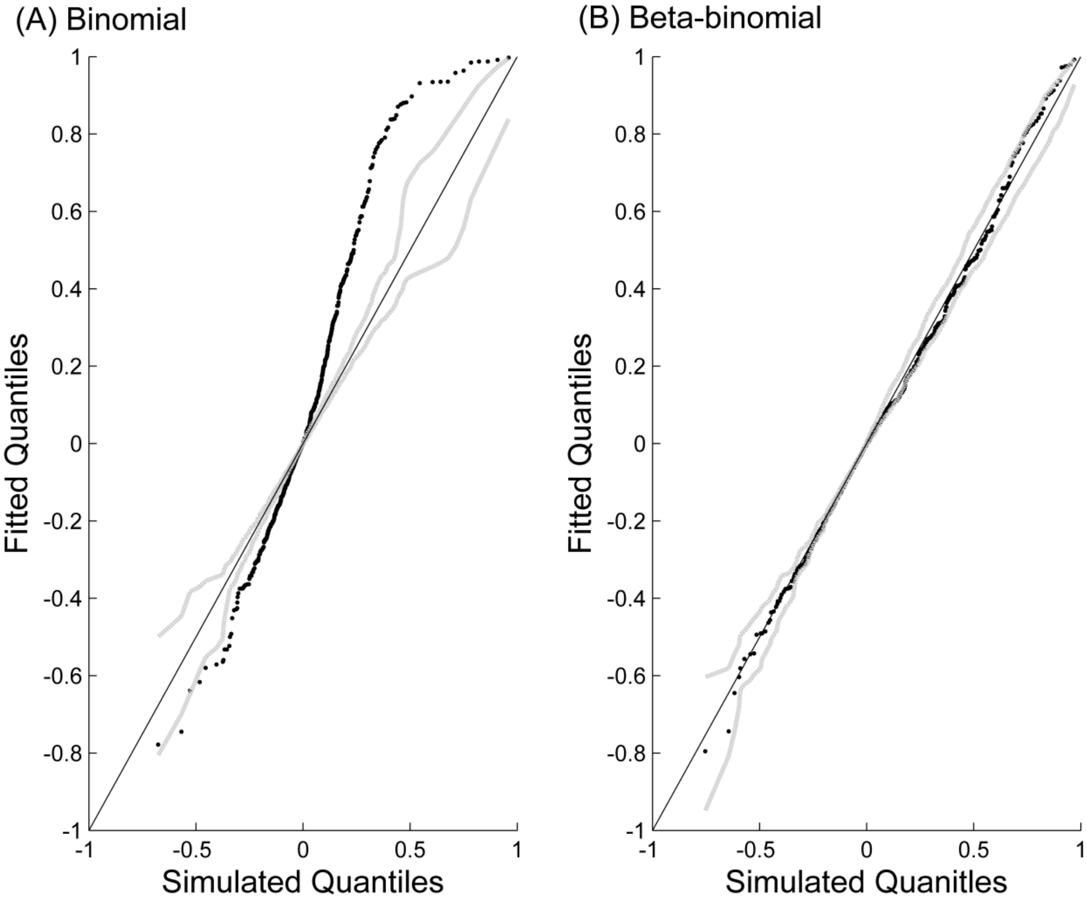


Figure 12.2. Quantile-quantile plots for: (A) binomial and (B) beta-binomial models of koala pellet decay. Dots represent the quantile-quantile plot, with the solid black line and grey lines representing the expected (1:1) relationship and 95% point-wise confidence intervals respectively. The quantile-quantile plot for the binomial model shows a pattern characteristic of overdispersion, with the lower end of the distribution lying below the confidence intervals and the upper end of the distribution lying above the confidence intervals. On the other hand, the quantile-quantile plot for the beta-binomial model lies close to the 1:1 line and within the confidence intervals, indicating little that overdispersion has been accounted for.

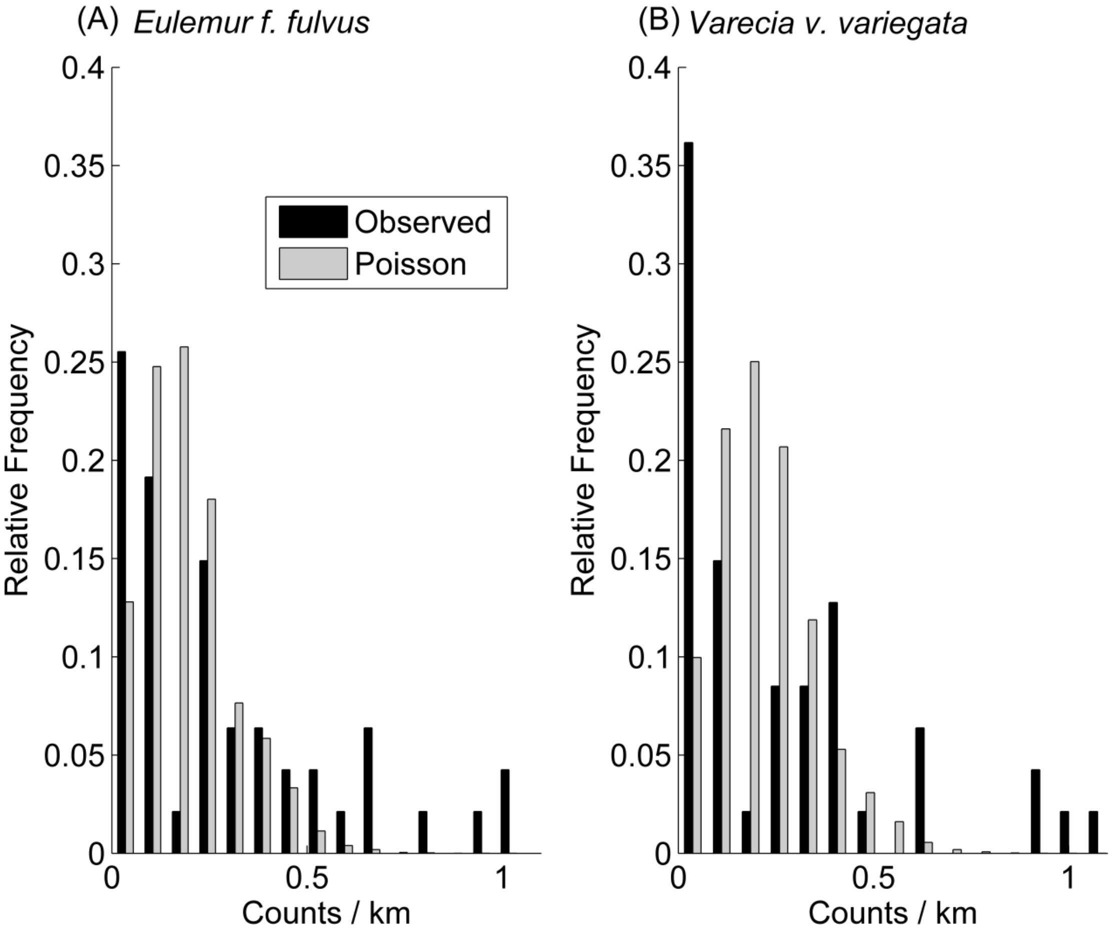


Figure 12.3. Histogram of observed counts km^{-1} of transect surveyed on each transect section (black bars) versus expected values based on a Poisson distribution with the same mean as the observed data (grey bars) for: (A) *Eulemer f. fulvus* and (B) *Varecia v. variegata*.

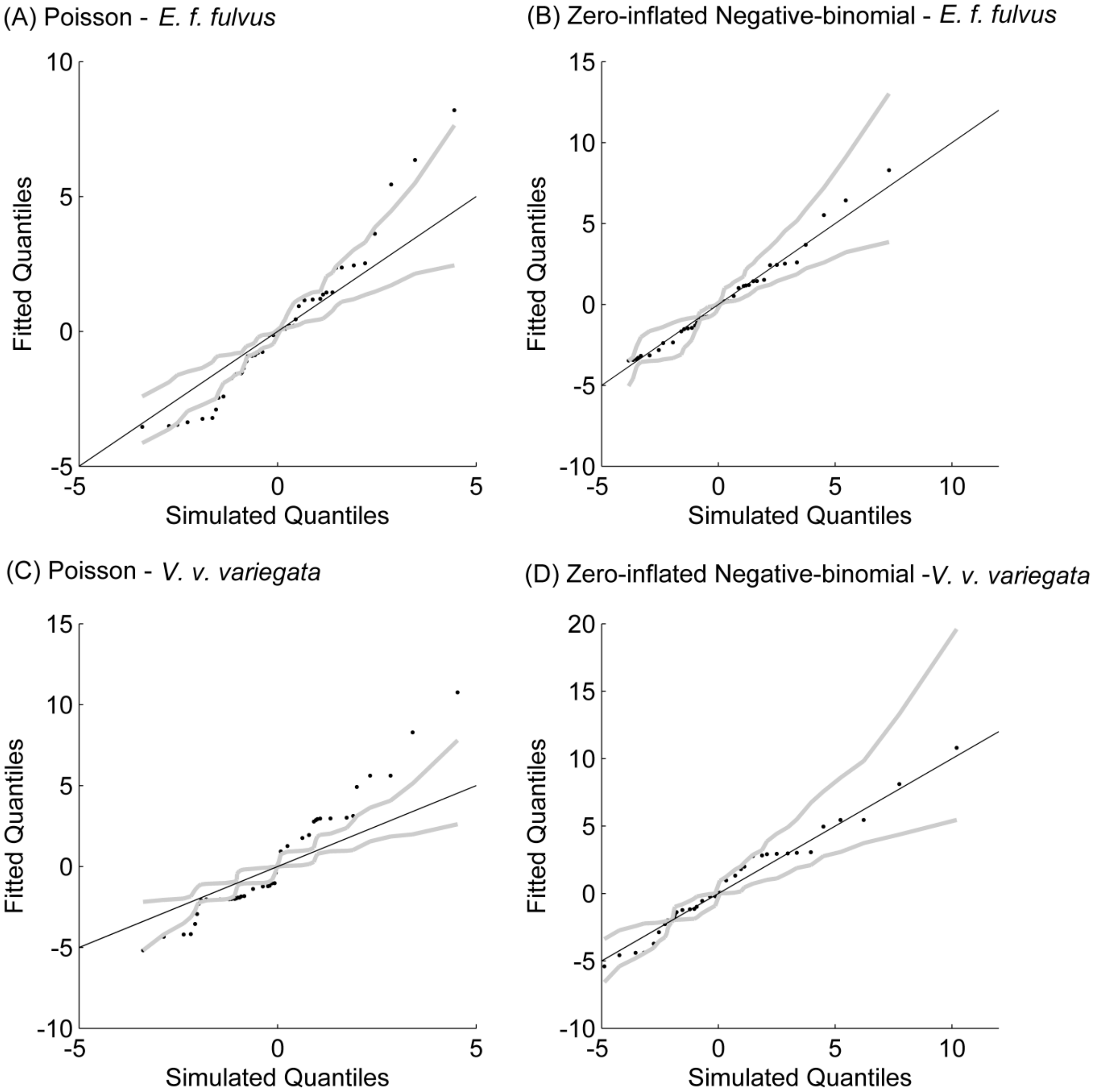


Figure 12.4. Quantile-quantile plots for the Poisson and zero-inflated negative binomial models for counts of *Eulemur f. fulvus* (A, B) for the Poisson and zero-inflated negative binomial models for counts of *Varecia v. variegata* (C, D). Dots represent the quantile-quantile plot, with solid black line and grey lines representing the expected 1:1 relationship and 95% point-wise confidence intervals respectively. The quantile-quantile plot for the Poison models show a pattern characteristic of overdispersion, with the lower end of the distribution lying below the confidence intervals and the upper end of the distribution lying above the confidence intervals. On the other hand, the quantile-quantile plots for the zero-inflated negative binomial models lie close to the 1:1 line and with points lying within the confidence intervals, indicating that overdispersion has been accounted for.