Chapter 8
Meet the Exponential Family

8.1 Introduction

In Chapters 2 and 3 and in Appendix A, linear regression and additive modelling were discussed and various extensions allowing for different variances, nested data, temporal correlation, and spatial correlation were then discussed in Chapters 4, 5, 6, and 7. In Chapters 8, 9, and 10, we discuss generalised linear modelling (GLM) and generalised additive modelling (GAM) techniques. In linear regression and additive modelling, we use the Normal (or: Gaussian) distribution. It is important to realise that this distribution applies for the response variable. GLM and GAM are extensions of linear and additive modelling in the sense that a non-Gaussian distribution for the response variable is used and the relationship (or link) between the response variable and the explanatory variables may be different. In this chapter, we focus on the first point, the distribution.

There are many reasons for using GLM and GAM instead of linear regression and additive modelling. Absence–presence data are (generally) coded as 1 and 0, proportional data are always between 0 and 100%, and count data are always non-negative. The GLM and GAM models used for $0-1$ and proportional data are typically based on the Bernoulli and binomial distributions and for count data the Poisson and negative binomial distributions are common options. For continuous data, the Gaussian distribution is the most used distribution, but you can also use the gamma distribution. So before using GLMs and GAMs, we should focus on the questions: What are these distributions, how do they look like, and when would you use them? These three questions form the basis of this chapter. We devote an entire chapter to this topic because in our experience few of our students have been familiar with Poisson, negative binomial or gamma distributions, and some level of familiarity is required before entering the world of GLMs and GAMs in the next chapter.

As we will see in the next chapter, a GLM (or GAM) consists of three steps: (i) choosing a distribution for the response variable, (ii) defining the systematic part in terms of covariates, and (iii) specifying the relationship (or: link) between the expected value of the response variable and the systematic part. This means that we have to stop for a moment and think about the nature of the response variable.

A.F. Zuur et al., *Mixed Effects Models and Extensions in Ecology with R*,
Statistics for Biology and Health, DOI 10.1007/978-0-387-87458-6_8,

In most statistics textbooks and undergraduate statistics courses, only the Normal, Poisson, and binomial distributions are discussed in any detail. However, there are various other distributions that are equally interesting for ecological data, for example, the negative binomial distribution. These are useful if the 'ordinary' GLMs do not work, and in practise, this is quite often in ecological data analysis.

Useful references for distributions within the context of GLMs are McCullagh and Nelder (1989), Hilbe (2007), and Hardin and Hilbe (2007). It should be noted that most books on GLMs discuss distributions, but these three have detailed explanations.

8.2 The Normal Distribution

We start with some revision on the Normal distribution. Figure 8.1 A shows the histogram of the weight of 1280 sparrows (unpublished data from Chris Elphick, University of Connecticut, USA). The *y*-axis in panel A shows the number per class. It is also possible to rescale the *y*-axis so that the total surface under the histogram adds up to 1 (Fig. 8.1B). The reason for doing this is to give a better representation of the density curve that we are going to use in a moment. The shape of the histogram suggests that assuming normality may be reasonable, even though the histogram is

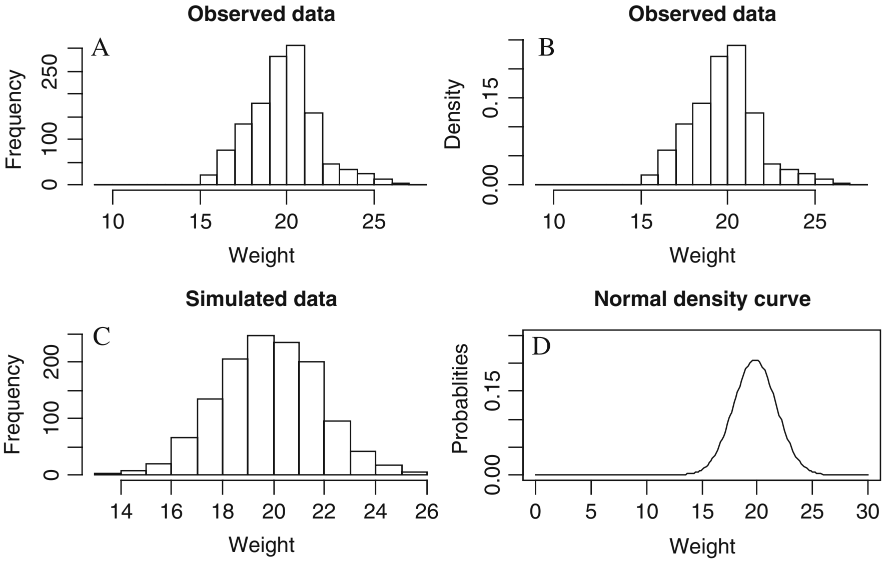

Fig. 8.1 **A**: Histogram of weight of 1281 sparrows. **B**: As panel A, but now scaled so that the total area in the histogram is equal to 1. **C**: Histogram of simulated data from a Normal distribution with mean and variance taken from the 1281 sparrows. **D**: Normal probability curve with values for the mean and the variance taken from the sample of 1281 sparrows. The surface under the Normal density curve adds up to 1

slightly skewed. Panel C shows simulated data (1280 observations) from a Normal distribution with the sample mean (18.9) and sample variance (3.7) from the 1280 sparrows. The shape of the histogram in panel C gives an impression of how a distribution looks if the data are really Normal distributed. Repeating this simulation ten times gives a good idea how much variation you can expect in the shape of the histogram.

Possible factors determining the weight of a sparrow are sex, age, time of the year, habitat, and diet, among others. But for the moment, we will not take these into account. The Normal distribution is given by the following formula:

$$f(y_i;\mu,\sigma)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(y_i-\mu)^2}{2\sigma^2}} \tag{8.1}$$

The distribution function in Equation (8.1) gives the probability that bird *i* has a weight y_i, and μ and σ^2 are the population mean and variance, respectively, in the following formula:

$$E(Y)=\mu \qquad \text{and} \qquad \text{var}(Y)=\sigma^2 \tag{8.2}$$

The probability function is also called a density function. The notation $f(y_i;\mu,\sigma)$ means that the parameters are after the ';' symbol. The variable *y* can take any value between $-\infty$ and ∞. In general, we do not know the population mean μ and variance σ^2, but if we just take the sample mean and variance and substitute these into the distribution function in Equation (8.1), we can calculate the probabilities for various values of *y*; see Fig. 8.1D. Note that the *y*-axis in this panel represents probabilities of certain weight values. So, the probability that we measure a sparrow of weight 20 g is about 0.21, and for 5 g, the probability is very small. According to the Normal distribution, we can even measure a bird with weight –10 g, though with a very small probability.

In linear regression, we model the expected values μ_i (the index *i* refers to observations or cases) as a function of the explanatory variables, and this function contains unknown regression parameters (intercept and slopes).

The following R code was used to create Fig. 8.1.

```
> library(AED); data(Sparrows)
> op <- par(mfrow = c(2, 2))
> hist(Sparrows$wt, nclass = 15, xlab = "Weight",
       main = "Observed data")
> hist(Sparrows$wt, nclass = 15, xlab = "Weight",
       main = "Observed data", freq = FALSE)
> Y <- rnorm(1281, mean = mean(Sparrows$wt),
             sd = sd(Sparrows$wt))
> hist(Y, nclass = 15, main = "Simulated data",
       xlab = "Weight")
> X <- seq(from = 0, to = 30, length = 200)
> Y <- dnorm(X, mean = mean(Sparrows$wt),
             sd = sd(Sparrows$wt))
> plot(X, Y, type = "l", xlab = "Weight",
       ylab = "Probablities", ylim = c(0, 0.25),
       xlim = c(0, 30), main = "Normal density curve")
> par(op)
```

The `freq = FALSE` option in the histogram scales it so that the area inside the histogram equals 1. The function `rnorm` takes random samples from a Normal distribution with a specified mean and standard deviation. The functions `mean` and `sd` calculate the mean and standard deviation of the weight variable `wt`. Similarly, the function `dnorm` calculates the Normal density curve for a given range of values *X* and for given mean and variance.

In this case, the histogram of the observed weight data (Fig. 8.1B) indicates that the Normal distribution may be a reasonable starting point. But what do you do if it is not (or if you do not agree with our statement)? The first option is to apply a data transformation, but this will also change the relationship between the response and explanatory variables. The second option is to do nothing yet and hope that the residuals of the model are normally distributed (and the explanatory variables cause the non-normality). Another option is to choose a different distribution and the type of data determines which distribution is the most appropriate. The best way to get some familiarity with different distributions for the response variable is to plot them. We have already seen the Normal distribution in Fig. 8.1, and also in Chapter 2. The second distribution we now discuss is the Poisson distribution.

8.3 The Poisson Distribution

The Poisson distribution function is given by

$$f(y;\mu) = \frac{\mu^y \times e^{-\mu}}{y!} \qquad y \geq 0,\ y \text{ intger} \tag{8.3}$$

This formula specifies the probability of *Y* with a mean μ. Note that *Y* has to be an integer value or else the $y! = y \times (y-1) \times (y-2) \times \ldots \times 1$ is not defined. Once we know μ, we can calculate the probabilities for different *y* values. For example, if $\mu = 3$, the probability that $y = 1$ is given by $3 \times e^{-3} / (1!) = 0.149$. The same can be done for other values of *y*. Figure 8.2 shows four Poisson probability distributions, and to create these graphs, we used different values for the average μ. For small μ, the density curve is skewed, but for larger μ, it becomes symmetrical. Note that μ can be a non-integer, but the *y*s have to be non-negative and integers. Other characteristics of the Poisson distribution are that $P(Y < 0) = 0$ and the mean is the variance, in formula

$$E(Y) = \mu \qquad \text{and} \qquad \text{var}(Y) = \mu \tag{8.4}$$

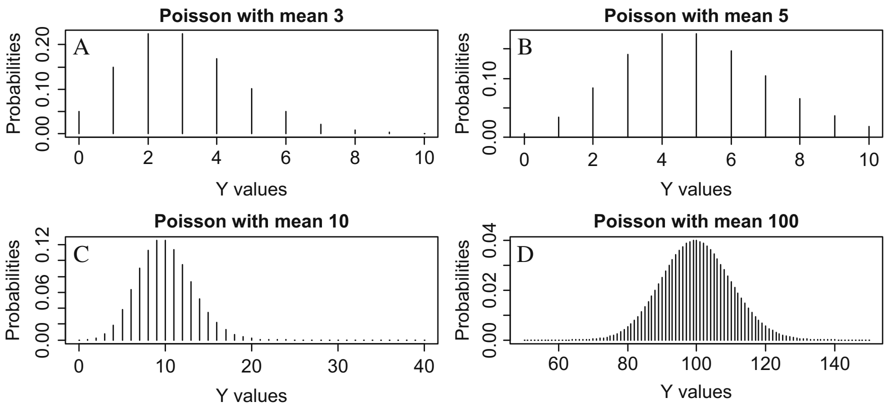

Fig. 8.2 Poisson probabilities for $\mu = 3$ (**A**), $\mu = 5$ (**B**), $\mu = 10$ (**C**), and $\mu = 100$ (**D**). Equation (8.3) is used to calculate the probabilities for certain values. Because the outcome variable y is a count, vertical lines are used instead of a line connecting all the points

This is also the reason that the probability distributions become wider and wider for larger mean values. Note that although the Poisson probability distribution in Fig. 8.2D looks like a normal distribution, it is not *equal* to a Normal distribution; a Normal distribution has two parameters (the mean μ and the variance σ^2), whereas a Poisson distribution only uses one parameter μ (which is the mean and the variance).

The following code was used the make Fig. 8.2.

```
> x1 <- 0:10;   Y1 <- dpois(x1, lambda = 3)
> x2 <- 0:10;   Y2 <- dpois(x2, lambda = 5)
> x3 <- 0:40;   Y3 <- dpois(x3, lambda = 10)
> x4 <- 50:150; Y4 <- dpois(x4, lambda = 100)
> XLab <- "Y values"; YLab <- "Probabilities"
> op <- par(mfrow = c(2, 2))
> plot(x1, Y1, type = "h", xlab = XLab, ylab = YLab,
     main = "Poisson with mean 3")
> plot(x2, Y2, type = "h", xlab = XLab, ylab = YLab,
     main = "Poisson with mean 5")
> plot(x3, Y3, type = "h", xlab = XLab, ylab = YLab,
     main = "Poisson with mean 10")
> plot(x4, Y4, type = "h", xlab = XLab, ylab = YLab,
       main = "Poisson with mean 100")
> par(op)
```

The function `dpois` calculates the Poisson probabilities for a given μ, and it calculates the probability for certain Y-values using Equation (8.3). Note that we use the symbol ';' to print multiple R commands on one line; it saves space. The `type = "h"` part in the `plot` command ensures that vertical lines are used in the graph. The reason for using vertical lines is because the Poisson distribution is for discrete data.

In the graphs in Fig. 8.2, we pretended we knew the value of the mean μ, but in real life, we seldom know its value. A GLM models the value of μ as a function of the explanatory variables; see Chapter 9.

The Poisson distribution is typically used for count data, and its main advantages are that the probability for negative values is 0 and that the mean variance relationship allows for heterogeneity. However, in ecology, it is quite common to have data for which the variance is even larger than the mean, and this is called overdispersion. Depending how much larger the variance is compared to the mean, one option is to use the correction for overdispersion within the Poisson GLM, and this is discussed in Chapter 9. Alternatively, we may have to choose a different distribution, e.g. the negative binomial distribution, which is discussed in the next section.

8.3.1 Preparation for the Offset in GLM

The Poisson distribution in Equation (8.2) is written for only one observation, but in reality we have multiple observations. So, we need to add an index *i* to *y* and μ.

Penston et al. (2008) analysed the number of sea lice at sites around fish farms in the north-west of Scotland as a function of explanatory variables like time, depth, and station. The response variable was the number of sea lice at various sites *i*, denoted by N_i. However, samples were taken from a volume of water, denoted by V_i, that differed per site. One option is to use the density N_i/V_i as the response variable and work with a Gaussian distribution, but if the volumes differ considerably per site, then this is a poor approach as it ignores the differences in volumes.

Alternative scenarios are the number of arrivals Y_i per time unit t_i, numbers Y_i per area of size A_i, and number of bioluminescent flashes per depth range V_i. All these scenarios have in common that the volume V_i, time unit t_i, area of size A_i, may differ per observation *i,* making the ratio of Y_i and V_i a rate or density.

We can still use the Poisson distribution for this type of data. For example, for the sea lice data, we assume that Y_i is Poisson distributed with probability function:

$$f(y_i;\mu_i) = \frac{(V_i \times \mu_i)^{y_i} \times e^{-V_i \times \mu_i}}{y_i!} \tag{8.5}$$

The parameter μ_i is now the expected number of sea lice at site *i* for a 1-unit volume. If all the values V_i are the same, we may as well drop it (for the purpose of a GLM) and work with the Poisson distribution in Equation (8.3).

8.4 The Negative Binomial Distribution

We continue the trail of distribution functions with another discrete one: the negative binomial. There are various ways of presenting the negative binomial distribution and a detailed explanation can be found in Hilbe (2007). Because we are working towards a GLM, we present the negative binomial used in GLMs. It is presented in the literature as a combination of two distributions, giving a combined Poisson-gamma distribution. This means we first assume that the *Y*s are Poisson distributed with the mean μ assumed to follow a gamma distribution. With some mathematical manipulation, we end up with the negative binomial distribution for *Y*. Its density function looks rather more intimidating than that of the Poisson or Normal distributions and is given by

$$f(y;k,\mu) = \frac{\Gamma(y+k)}{\Gamma(k)\times\Gamma(y+1)} \times \left(\frac{k}{\mu+k}\right)^k \times \left(1-\frac{k}{\mu+k}\right)^y \tag{8.6}$$

Nowadays, the negative binomial distribution is considered a stand-alone distribution, and it is not necessary to dig into the Poisson-gamma mixture background. The distribution function has two parameters: μ and *k*. The symbol Γ is defined as: $\Gamma(y + 1) = (y + 1)!$. The mean and variance of *Y* are given by

$$E(Y) = \mu \qquad \text{var}(Y) = \mu + \frac{\mu^2}{k} \tag{8.7}$$

We have overdispersion if the variance is larger than the mean. The second term in the variance of *Y* determines the amount of overdispersion. In fact, it is indirectly determined by *k*, where *k* is also called the dispersion parameter. If *k* is large (relative to μ^2), the term μ^2/k approximates 0, and the variance of *Y* is μ; in such cases the negative binomial converges to the Poisson distribution. In this case, you might as well use the Poisson distribution. The smaller *k*, the larger the overdispersion.

Hilbe (2007) uses a different notation for the variance, namely,

$$\text{var}(Y) = \mu + \alpha \times \mu^2$$

This notation is slightly easier as $\alpha = 0$ means that the quadratic term disappears. However, the R code below uses the notation in Equation (8.7); so we will use it here.

It is important to realise that this distribution is for discrete (integers) and non-negative data. Memorising the complicated formulation of the density function is not needed; the computer can calculate the Γ terms. All you need to remember is that with this distribution, the mean of *Y* is equal to μ and the variance is $\mu + \mu^2/k$.

The probability function in Equation (8.6) looks complicated, but it is used in the same way as we used it in the previous section. We can specify a μ value and a *k* value, and calculate the probability for a certain *y* value. To get a feeling for the shape of the negative binomial probability curves, we drew a couple of density

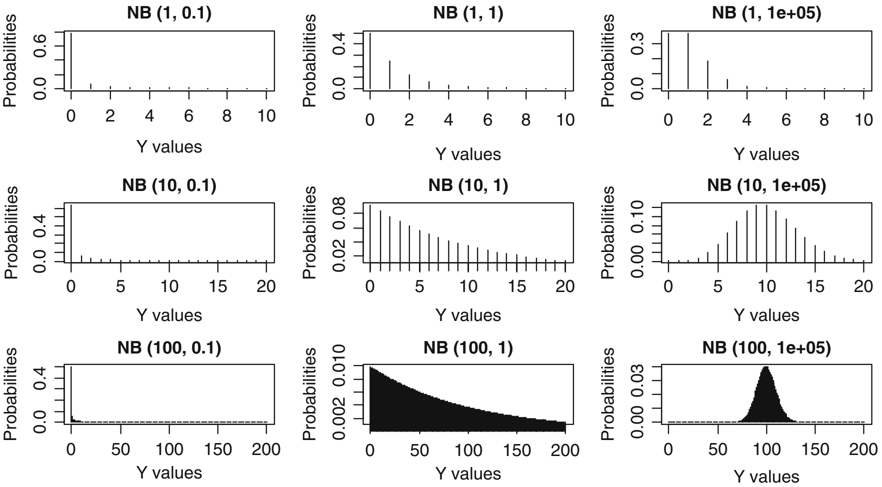

Fig. 8.3 Nine density curves from a negative binomial distribution NB(μ, k), where μ is the mean and k^{-1} is the dispersion parameter. The column of panels on the right have a large k, and these negative binomial curves approximate the Poisson distribution. R code to create this graph is given on the book website. If $k = 1$, the negative binomial distribution is also called the geometric distribution

curves for various values of μ and k, see Fig. 8.3. We arbitrarily choose three values for μ, of 1, 10, and 100. We also choose arbitrarily three values for k, of 0.1, 1, and 100,000. For $k = 100{,}000$, we expect to see a distribution function similar to the Poisson distribution with mean and variance μ, and this is indeed the case: see the panels in the right column. The three panels in the middle column have E(Y) $= \mu$ and var(Y) $= \mu + \mu^2$, because $k = 1$.

If we set $k = 1$ in the negative binomial distribution, then the resulting distribution is called the geometric distribution. Its mean and variance are defined by

$$E(Y) = \mu \qquad \text{var}(Y) = \mu + \mu^2 \tag{8.8}$$

Hence, the variance increases as a quadratic function of the mean. As with the Poisson distribution, observations of the response variables with the value of zero are allowed in the negative binomial and the geometric distribution. Most software will not have a separate function for the geometric distribution; just set the parameter k in the software for a negative binomial equal to 1.

Returning to the negative binomial probability function, note that for a small mean μ and large overdispersion (small k), the value of 0 has by far the highest probability.

In Fig. 8.3 we know the values of μ and k. In reality we do not know these values, and in GLM models, the mean μ is a function of covariates. Estimation of k depends on the software, but can for example be done in a 2-stage iterative approach (Agresti, 2002).

8.5 The Gamma Distribution

The gamma distribution can be used for a continuous response variable Y that has positive values ($Y > 0$), and the distribution function has various forms. Within the context of a GLM, we use (Faraway, 2006)

$$f(y;\mu,\nu) = \frac{1}{\Gamma(\nu)} \times \left(\frac{\nu}{\mu}\right)^{\nu} \times y^{\nu-1} \times e^{\frac{y\times\nu}{\mu}} \qquad y > 0 \tag{8.9}$$

Before starting to memorise the exact mathematical definition of this density function, let us first look at the mean and variance of a variable Y that is gamma distributed and sketch the density curve for various values of μ and v (which is the equivalent of the k in the negative binomial distribution). The mean and variance of Y are

$$E(Y) = \mu \qquad \text{and} \qquad \text{var}(Y) = \frac{\mu^2}{\nu} \tag{8.10}$$

The dispersion is determined by v^{-1}; a small value of v (relative to μ^2) implies that the spread in the data is large. Density curves for difference values of μ and v are given in Fig. 8.4. Note the wide range of shapes between these curves. For a large v, the gamma distribution becomes bell shaped and symmetric. In such cases, the Gaussian distribution can be used as well. Faraway (2006) gives an example of a linear regression model and a gamma GLM with a small (0.0045) dispersion parameter v^{-1}; estimated parameters and standard errors obtained by both methods are nearly identical. However, for larger values of v^{-1}, this is not the necessarily the case.

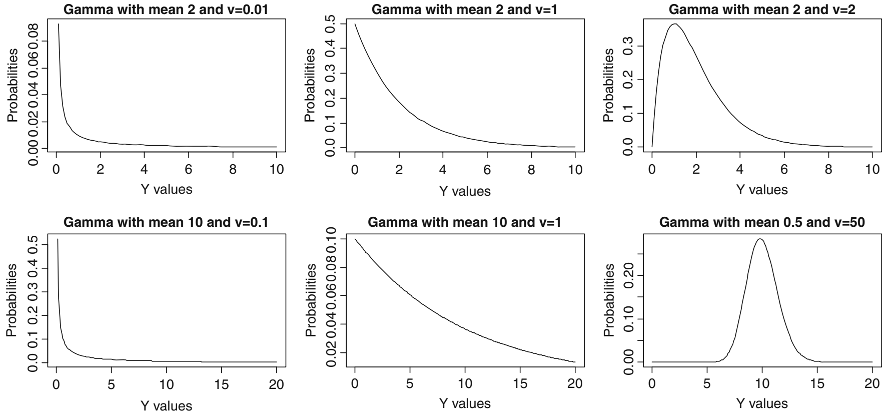

Fig. 8.4 Gamma distributions for different values of μ and v. The R function dgamma was applied, which uses a slightly different parameterisation: E(Y) = $a \times s$ and var(Y) = $a \times s^2$, where a is called the shape and s the scale. In our parameterisation, $v = a$ and $\mu = a \times s$

Note that the allowable range of Y values is larger then 0. So, you cannot use this distribution if your response variable takes negative values or has a value of zero.

8.6 The Bernoulli and Binomial Distributions

The last two distributions we review are the Bernoulli and binomial distributions, and we start with the latter. In a first year statistics course, it is often introduced as the distribution that is used for tossing a coin. Suppose you know that a coin is fair (no one has tampered with it and the probability of getting a head is the same as getting a tail), and you toss it 20 times. The question is how many heads do you expect? The possible values that you can get are from 0 to 20. Obviously, the most likely value is 10 heads. Using the binomial distribution, we can say how likely it is that you get 0, 1, 2, ..., 19 or 20 heads.

A binomial distribution is defined as follows. We have N independent and identical trials, each with probability $P(Y_i = 1) = \pi$ of success, and probability $P(Y_i = 0) = 1 - \pi$ on failure. The labels 'success' and 'failure' are used for the outcomes of 1 and 0 of the experiment. The label 'success' can be thought of $P(Y_i = \text{head})$, and 'failure' can be $P(Y_i = \text{tail})$. The term independent means that all tosses are unrelated. Identical means that each toss has the same probability of success. Under these assumptions, the density function is given by

$$f(y;\pi) = \binom{N}{y} \times \pi^{y} \times (1-\pi)^{N-y} \tag{8.11}$$

The probability for each value of y between 0 and 20 for the tossing example can be calculated with this probability function. For example, if $N = 20$ and $\pi = 0.5$, then the probability of measuring 9 heads is $(20!/(9! \times 11!)) \times 0.5^9 \times (1 - 0.5)^{11}$. The value can either be obtained from a calculator or you can read it from the panel in the middle of Fig. 8.5 ($N = 20$, $\pi = 0.5$). As expected, the value $y = 10$ has the highest probability, but 9 and 11 have very similar probabilities. The probability of getting 20 heads is close to zero; it is too small to read on the vertical axis (it is in fact something that starts with 7 zeros). For some arbitrarily chosen values of π and N, we drew more Binomial probability curves, just to get a feel for the shape of the density curves (Fig. 8.5).

The mean and variance of a Binomial distribution are given by

$$E(Y) = N \times \pi \qquad \text{var}(Y) = N \times \pi \times (1-\pi) \tag{8.12}$$

So, if you know that the probability of tossing a head is 0.5 and toss a coin 20 times, then the answer to the question that we started this section with is $20 \times 0.5 = 10$ heads.

In ecology, we are (hopefully) not tossing with coins, but instead we may go to a deer farm and sample N animals for the presence and absence of a particular disease. In such a research, you want to know the probability π that a particular animal is infected with the disease. Other examples are the presence or absence of koalas at

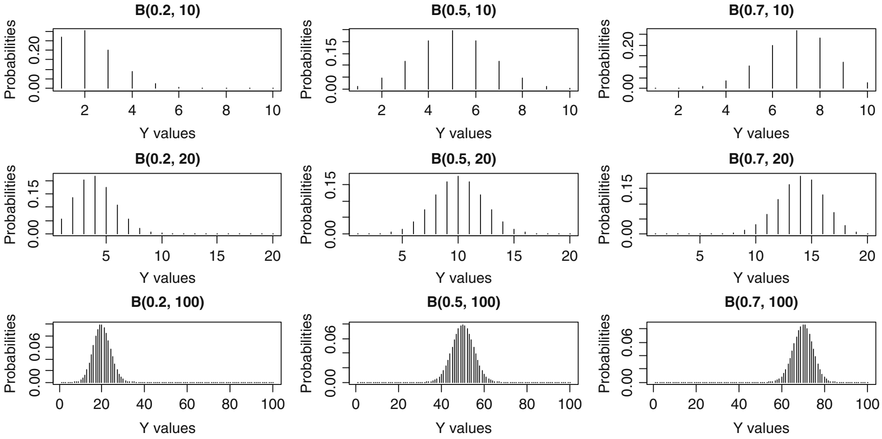

Fig. 8.5 Binomial density curves B(π, N) for various values of π (namely 0.2, 0.5, and 0.7) and N (namely 10, 20, and 100). R code to create this graph is on the book website

particular sites (see Chapter 20 for a detailed example), badger activity (yes or no) around farms (Chapter 21), or the presence and absence of flat fish at 62 sites in an estuary (Chapter 21 in Zuur et al., 2007).

In the example of the N deer at the farm, we do not know the value of π and the GLM is used to model π as a function of covariates. In such a research problem, you can also question if your sample of 20 animals from the same farm is independent. But we leave this problem until Chapters 12 and 13.

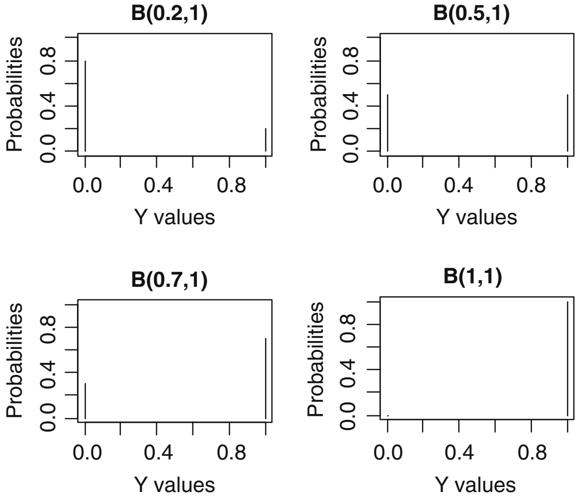

Fig. 8.6 Four Bernoulli distributions B(π,1) for different values of π. R code to create this graph is on the book website

A Bernoulli distribution is obtained if $N = 1$; hence, we only toss once or we only sample one animal on the farm. Four Bernoulli distributions with $\pi = 0.2$, $\pi = 0.5$, $\pi = 0.7$, and $\pi = 1$ are given in Fig. 8.6. Note that we only get a value of the probabilities at 0 (failure) and 1 (success).

In general, we do not make a distinction between a binomial and Bernoulli distribution and use the notation $\mathrm{B}(\pi, N)$ for both, and $N = 1$ automatically implies the Bernoulli distribution.

8.7 The Natural Exponential Family

So far, we have discussed the Normal, Poisson, negative binomial, gamma, binomial, and Bernoulli distributions. There are, however, a lot more distributions around, for example, the multinomial distribution (useful for a response variable that is a categorical variable with more than two levels) and inverse Gaussian distribution (e.g. for lifetime distributions; these can be used for failure time of machines in production processes or lifetime of a product). It is relatively easy to show that all the distributions we have used so far can be written in a general formulation:

$$f(y;\theta,\phi) = e^{\frac{y\times\theta - b(\theta)}{a(\phi)} + c(y,\theta)} \tag{8.13}$$

For example, if we use $\theta = \log(\mu)$, $\phi = 1$, $a(\phi) = 1$, $b(\theta) = \exp(\theta)$, $c(y, \phi) = –\log(y!)$, we get the Poisson distribution function. Similar definitions exist for the binomial, negative binomial, geometric, Normal, and gamma distributions; see McCullagh and Nelder (1989), Dobson (2002), Agresti (2002), or Hardin and Hilbe (2007). The advantage of this general notation is that when we build up a maximum likelihood criterion and optimise this to estimate the regression parameters, we can do this in terms of the general notation. This means that one set of equations can be used for all these distributions.

Using first- and second-order derivatives for the density function specified in Equation (8.13), we can easily derive an expression for the mean and variance of Y. These are as follows:

$$\begin{aligned} E(Y) &= b^{\prime(\theta)} \\ \mathrm{var}(Y) &= b''(\theta) \times a(\phi) \end{aligned} \tag{8.14}$$

The notation $b'(\theta)$ refers to the first-order derivative of the function b with respect to θ, and $b''(\theta)$ the second-order derivative. If you check this for the Poisson distribution, you will see that we get the familiar relationships $\mathrm{E}(Y) = \mu$ and $\mathrm{var}(Y) = \mu$.

The term $a(\phi)$ determines the dispersion. In the Gaussian linear regression model, we have $\theta = \mu$, $\phi = \sigma^2$, $a(\phi) = \phi$, $b(\theta) = \theta^2/2$, and $c(y, \phi) = -(\mathrm{y}^2/\phi + \log(2\pi\phi))/2$, which gives us $\mathrm{E}(Y) = \mu$ and $\mathrm{var}(Y) = \sigma^2$.

Hence, the notation in Equation (8.13) allows us to summarise all distribution functions discussed so far in a general notation, and the mean and variance are specified by the set of equations in (8.14).

8.7.1 Which Distribution to Select?

We have discussed a large number of distributions for the response variable, but which one should we use? This choice should, in first instance, be made a priori based on the available knowledge on the response variable. For example, if you model the presence and absence of animals at M sites as a function of a couple of covariates, then your choice is simple: the binomial distribution should be used because your response variable contains zeros and ones. This is probably the only scenario where the choice is so obvious. Having said that, if we aggregate the response variable into groups, we (may) have a Poisson distribution.

If your data are counts (of animals, plants, etc.) without an upper limit, then the Poisson distribution is an option. This is because counts are always non-negative, and tend to be heterogeneous and both comply with the Poisson distribution. If there is high overdispersion, then the negative binomial distribution is an alternative to the Poisson for count data.

You can also use the Normal distribution for counts (potentially combined with a data transformation), but the Poisson or negative binomial may be more appropriate. However, the Normal distribution does not exclude negative realisations.

You can also have counts with an upper limit. For example, if you count the number of animals on a farm that are infected with a disease, out of a total of N animals. The maximum number of infected animals is then N. If you consider each individual animal as an independent trial and each animal has the same probability of being infected, then we are in the world of a binomial distribution.

But, what do you do with densities? Density is often defined as the numbers (which are counts!) per volume (or area, depth range, etc.). We will see in Chapter 9 that this can be modelled with the Poisson (or NB) distribution and an offset variable.

If the response variable is a continuous variable like weight of the animal, then the Normal distribution is your best option, but the gamma distribution may be an alternative choice.

The important thing to realise is that these distributions are for the response variable, not for explanatory variables. The choice of which distribution to use is an a priori choice. A list of all discussed distributions in this section is given in Table 8.1. If you are hesitating between two competing distributions, e.g. the Normal distribution and the gamma distribution, or the Poisson distribution and the negative binomial distribution, then you could plot the mean versus the variance of the response variable and see what type of mean–variance relationship you have and select a distribution function accordingly. In Chapter 9, we will see that the Poisson distribution is nested in the NB distribution, which opens the possibility for a likelihood ratio test.

Table 8.1 List of distributions for the response variable. Density means numbers per Area (or volume, range, etc), and in this case the offset option is needed in the Poisson or NB GLM

Distribution	Type of data	Mean – variance relationship
Normal	Continuous	Equation (8.2)
Poisson	Counts (integers) and density	Equation (8.4)
Negative binomial	Overdispersed counts and density	Equation (8.7)
Geometric	Overdispersed counts and density	Equation (8.8)
Gamma	Continuous	Equation (8.10)
Binomial	Proportional data	Equation (8.12)
Bernoulli	Presence absence data	Equation (8.12) with $N = 1$

8.8 Zero Truncated Distributions for Count Data

The discussion presented in this section applies to the Poisson, negative binomial, and the geometric distributions. All three distributions can be used for count data. Suppose we sample N sites, and at each site we count the number of birds, denoted by Y_i. The values that we can measure are 0, 1, 2, 3, ..., etc. For a given mean μ, the Poisson, negative binomial, and geometric distributions specify the probability of having a count of 0, 1, 2, etc. For example, if we use the Poisson distribution with $\mu = 3$, Fig. 8.2A shows that the probability of counting 0 animals is 13.5%. So, if you had a sample of size $N = 100$, you would expect to have a zero count approximately 14 times in your resulting data set. But what if you have a response variable that cannot take the value of 0? A typical example from the medical literature is the length of stay of a patient in a hospital. As soon as the patient enters the hospital, the length of stay is at least 1. In ecology, it is more difficult to envisage examples that structurally exclude zeros, but think of the number of plants in a transect and you know that it would be impossible to have transects with zero abundance due to the experimental design, the time that a whale stays at the surface before submerging (it has to breath) or the number of days per year with rain in Scotland. These are all variables that cannot have the value of 0. However, the Poisson, negative binomial, and geometric distributions do not exclude this value, and this can be a problem for small mean values μ.

The solution is to modify the distribution and exclude the possibility of a zero observation, and this is called a zero truncated distribution. We illustrate the process for a Poisson distribution, but the process is similar for the other two distributions. In fact, the same problem exists for continuous distributions. Think of the weight of an animal. The weight is always positive, and if the majority of the observations have small values, a Gaussian distribution may not be appropriate as it allows for negative values and realisations. In this chapter, we focus on discrete distributions (because we need them in Chapter 11), but the Tobit model can be used for the Gaussian distribution. Cameron and Trivedi (1998) is a good reference.

Recall that the Poisson distribution is given by

$$f(y_i;\mu) = \frac{\mu^{y_i} \times e^{-\mu}}{y_i!} \tag{8.15}$$

The probability that $y_i = 0$, is given by

$$f(0;\mu) = \frac{\mu^0 \times e^{-\mu}}{0!} = e^{-\mu}$$

The probability of not measuring a 0 is given by 1 – e$^{-\mu}$. If we use $\mu = 2$, then the probability that $y_i = 0$, is 0.135 and the probability of not measuring a 0 is 0.864. In Fig. 8.7A, we have sketched the Poisson distribution with $\mu = 2$. In panel B, we put a cross through the line that represents the probability of sampling a 0 count. The cross is our pedagogical way of saying that we are changing the Poisson density and setting the probability that $y = 0$ equal to 0. However, this leaves us with the problem that by definition the sum of the probabilities of all outcome should be exactly 1. Removing the probability of $y = 0$ means that the remaining probabilities add up to 0.864. The solution is simple; divide the probability of each outcome larger than 0 by 0.864. The sum of all scaled probabilities will then add up to 1 again. We

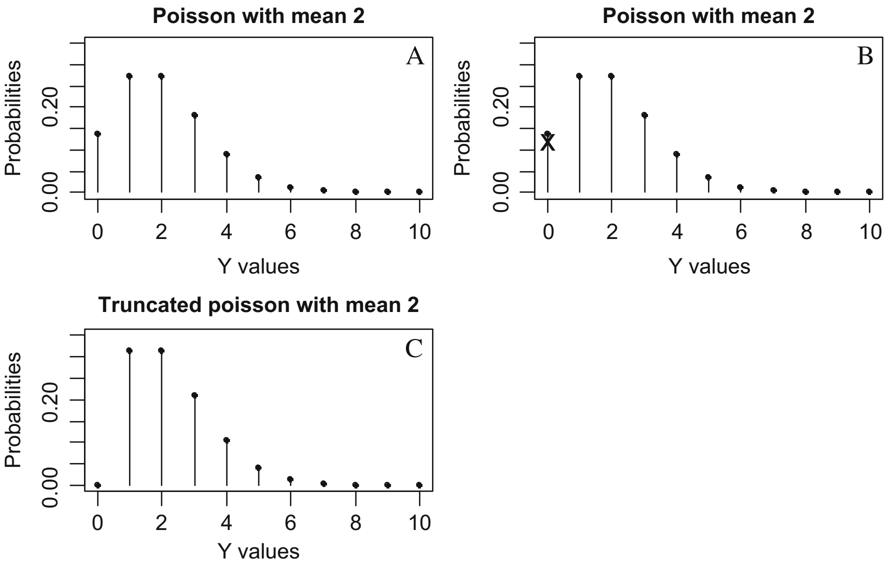

Fig. 8.7 **A**: Poisson distribution with $\mu = 2$. The sum of all probabilities is 1. **B**: The zero outcome is dropped from the possible range of outcomes, as indicated by a cross. The sum of all probabilities is equal to 0.864. **C**: Adjusted probabilities according to Equation (8.15). The vertical lines are slightly higher (because each probability was divided by 0.864), and the probability that $y_i = 0$ is zero. The sum of all scaled probabilities equals 1

therefore need to divide the Poisson probability function by the probability that we have a count larger than 0, and the new probability function is

$$f(y_i;\mu|y_i > 0) = \frac{\mu^{y_i} \times e^{-\mu}}{(1 - e^{-\mu}) \times y_i!} \tag{8.16}$$

The notation '$|\,y_i > 0$' is used to indicate that y_i is larger than 0. This is called the zero-truncated Poisson distribution. The same can be done for the negative binomial distribution.

The distribution function in Equation (8.15) will be used in GLMs and GAMs to model zero-truncated data. The underlying principle will also be applied in models that have too many zeros (zero inflated Poisson). For further details, see Chapter 11 or Hilbe (2007).