2

STATISTICAL MODELS AND SCR

In the previous chapter we described the basics of capture-recapture methods and the advantages that spatial models have over traditional non-spatial models. We avoided statistical terminology like the plague so that we could focus on a few key concepts. Although it is critical to understand the non-technical motivation for this broad class of models, it is impossible to fully appreciate them, and apply them to real data, without a solid grasp of the fundamentals of statistical inference.

In this chapter, we present a brief overview of the basic statistical principals that are referenced throughout the remainder of this book. Emphasis is placed on the definition of a random variable, the common probability distributions used to model random variables, and how hierarchical models can be used to describe conditionally related random variables. For some readers, this material will be familiar, perhaps even elementary, and thus you may want to skip to the next chapter. However, our experience is that many basic statistics courses taken by ecologists do not emphasize the important subjects covered in this chapter. Instead, there seems to be much attention paid to minor details such as computing the number of degrees of freedom in various F-tests, which, although useful in some contexts, do not provide the basis for drawing conclusions from data and evaluating scientific hypotheses.

The material in the beginning of this chapter is explained in numerous other texts. Technical treatments that emphasize ecological problems are given by Williams et al. (2002), Royle and Dorazio (2008) and Link and Barker (2010), to name just a few. A very accessible introduction to some of the topics covered in this chapter is presented in Chapt. 3 of MacKenzie et al. (2006). With all these resources, one might wonder why we bother rehashing these concepts here. Our motivation is two-fold: first, we wish to develop this material using examples relevant to spatial capture-recapture, and second, we find that most introductory texts are not accompanied by code that can be helpful to the novice. We therefore attempt to present

simple **R** code throughout this chapter so that those who struggle with equations and mathematical notation can learn by doing. As mentioned in the Preface, we rely on **R** because it provides tremendous flexibility for analyzing data and because it is free. We do not, however, try to explain how to use **R** because there are so many good references already, including Venables and Ripley (2002); Bolker (2008); Venables et al. (2012).

After covering some basic concepts of hierarchical modeling, we end the chapter by describing spatial capture-recapture models using hierarchical modeling notation. This makes the concepts outlined in the previous chapter more precise, and it highlights the fact that SCR models include explicit models for the ecological processes of interest (e.g. spatial variation in density) and the observation process, which describes how individuals are encountered.

2.1 RANDOM VARIABLES AND PROBABILITY DISTRIBUTIONS

2.1.1 Stochasticity in ecology

Few ecological processes can be described using purely deterministic models, and thus we need a formal method for drawing conclusions from data while acknowledging the stochastic nature of ecological systems. This is the role of statistical inference, which is founded on the laws of probability. For our purposes, it suffices to be familiar with a small number of concepts from probability theory—the most important of which is the concept of a random variable, say X. A random variable is a variable whose realized value is the outcome of some stochastic process. To be more precise, a random variable is characterized by a function that describes the probability of observing the value x. This probability function can be written $\Pr(X = x|\theta)$ where θ is a parameter, or set of parameters of the function. If x is discrete, e.g. binary or integer, then we call the probability function a probability mass function (pmf). If x is continuous, the function is called a probability density function (pdf).

To clarify the concept of a random variable, let X be the number of American shad (*Alosa sapidissima*) caught after $K = 20$ casts at the shad hole on Deerfield River in Massachusetts. Suppose that we had a good day and caught $x = 7$ fish. If there were no random variation at play, we would say that the probability of catching a fish, which we will call p, is $p = 7/20 = 0.35$, and we would always expect to catch 7 shad after 20 casts. In other words, our deterministic model is $x = 0.35 \times K$. In reality, however, we can be pretty sure that this deterministic model would not be very good. Even if we knew for certain that $p \equiv 0.35$, we would expect some variation in the number of fish caught on repeated fishing outings. To describe this variation, we need a model that acknowledges uncertainty (i.e., stochasticity), and specifically we need a model that describes the probability of catching x fish given K and p, $\Pr(X = x|K, p)$. Since x is discrete, not continuous, we need a pmf. Before contemplating which pmf is most appropriate in this case,

we need to first mention a few issues related to notation.

Statisticians make things easier for themselves, and more complicated for everyone else, by using different notation for probability distributions. Sometimes you will see $\Pr(X = x|K, p)$ expressed as $f(X|K, p)$ or $f(X; K, p)$ or $p(X|K, p)$ or $\pi(X|K, p)$ or $\mathbb{P}(X|K, p)$ or $[X|K, p]$ or even just $[X]$! Just remember that these expressions all have the same meaning—they are all probability distributions that tell us the probability of observing any possible realization of the random variable X. In this book, we will almost always use bracket notation (the last two examples above) to represent arbitrary probability distributions. Hence, from here on out, when you see $[X|K, p]$, just remember that this is equivalent to the more traditional expression $\Pr(X = x|K, p)$. In addition, from here on, to achieve a more concise presentation, we will no longer use uppercase letters to denote random variables and lowercase letters for realized values. Rather, we will define a random variable by some symbol (x, N, etc...) and let the context determine whether we are talking about the random variable itself, or realized values of it. In some limited cases, we will want upper- and lower-case letters to represent different variables. For example, we will often let N denote population size and n denote the number of individuals actually detected.

When we wish to be specific about a probability distribution, we will do so in one of two ways, one mathematically precise and one symbolic. Before explaining these two options, let's choose a specific distribution as a model for the data in our example. In this case, the natural choice for $[x|K, p]$ is the binomial distribution, the mathematically precise representation of which is

$$[x|K, p] = \binom{x}{K} p^x (1-p)^{K-x}. \tag{2.1.1}$$

The right-hand side of this equation is the binomial pmf (described in more detail in Sec. 2.2), and plugging in values for the parameters K, and p will return the probability of observing any realized value of the random variable x. This is precise, but it is also cumbersome to write repetitively, and it may make the eyes glaze over when seen too often. Thus, we will often simplify Eq. 2.1.1 using the symbolic notation:

$$x \sim \text{Binomial}(K, p) \tag{2.1.2}$$

The "$\sim$" symbol is meant to represent a stochastic relationship, and can be read "is distributed as." Another reason for using this notation is that it resembles the syntax of the **BUGS** language, which we will frequently use to conduct Bayesian inference.

Note that once we choose a probability distribution, we have chosen a model. In our example, we have specified our model as $x \sim \text{Binomial}(K, p)$, and because we are assuming that the parameters are known, we can make probability statements about future outcomes. Continuing with our fish example, we might want to know the probability of catching $x = 7$ again after $K = 20$ casts on a future fishing

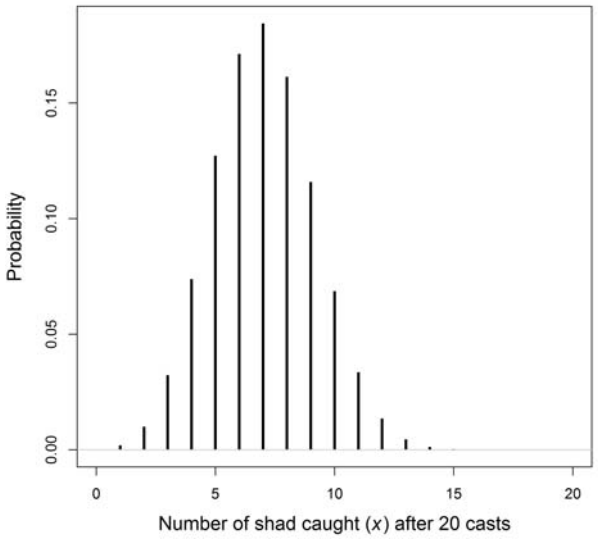

Figure 2.1. The binomial probability mass function with $N = 20$ and $p = 0.35$.

outing, assuming that we know $p = 0.35$. Evaluating the binomial pmf returns a probability of approximately 0.18, as show using this bit of **R** code:

```
> dbinom(7, 20, 0.35)
[1] 0.1844012
```

By definition, the pmf allows us to evaluate the probability of observing any x given $K = 20$ and $p = 0.35$, thus the distribution of the random variable can be visualized by evaluating it for all values of x that have non-negligible probabilities, as can be easily done in **R**:

```
plot(0:20, dbinom(0:20, 20, 0.35), type="h", ylab="Probability",
     xlab="Number of shad caught (X)")
```

the result of which is shown in Fig. 2.1 with some extra details.

The purpose of this little example is to show that once we specify a model for the random variable(s) being studied, we can begin drawing conclusions, i.e. making inferences, about the processes of interest, even in the face of uncertainty. Probability distributions are essential to this process, and thus we need to understand them in more depth.

Table 2.1. Common probability density functions (pdfs) and probability mass functions (pmfs) used throughout this book.

Distribution	Notation	pmf or pmf	Support	Mean $\mathbb{E}(x)$	Variance $\text{Var}(x)$
		Discrete random variables			
Poisson	$x \sim \text{Pois}(\lambda)$	$\exp(-\lambda)\lambda^x/x!$	$x \in \{0, 1, \ldots\}$	λ	λ
Bernoulli	$x \sim \text{Bern}(p)$	$p^x(1-p)^{1-x}$	$x \in \{0, 1\}$	p	$p(1-p)$
Binomial	$x \sim \text{Bin}(N, p)$	$\binom{N}{x} p^x(1-p)^{N-x}$	$x \in \{0, 1, \ldots, N\}$	Np	$Np(1-p)$
Multinomial	$\mathbf{x} \sim \text{Multinom}(N, \boldsymbol{\pi})$	$\binom{N}{x_1 \cdots x_k} \pi_1^{x_1} \cdots \pi_k^{x_k}$	$x_k \in \{0, 1, \ldots, N\}$	$N\pi_k$	$N\pi_k(1-\pi_k)$
		Continuous random variables			
Normal	$x \sim \text{N}(\mu, \sigma^2)$	$\frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{(x-\mu)^2}{2\sigma^2})$	$x \in [-\infty, \infty]$	μ	σ^2
Uniform	$x \sim \text{Unif}(a, b)$	$1/(b-a)$	$x \in [a, b]$	$(a+b)/2$	$(b-a)^2/12$
Beta	$x \sim \text{Beta}(a, b)$	$\frac{\Gamma(a+b)}{\Gamma(a)+\Gamma(b)} x^{a-1}(1-x)^{b-1}$	$x \in [0, 1]$	$a/(a+b)$	$\frac{ab}{(a+b)^2(a+b+1)}$
Gamma	$x \sim \text{Gamma}(a, b)$	$\frac{b^a}{\Gamma(a)} x^{a-1} \exp(-bx)$	$x \in [0, \infty]$	a/b	a/b^2
Multivariate Normal	$\mathbf{x} \sim \text{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	$(2\pi)^{-\frac{k}{2}} \lvert\boldsymbol{\Sigma}\rvert^{-\frac{1}{2}} \exp(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})' \times \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}))$	$x_k \in [-\infty, \infty]$	$\boldsymbol{\mu}$	$\boldsymbol{\Sigma}$

2.1.2 Properties of probability distributions

A pdf or a pmf is a function like any other function in the sense that it has one or more arguments whose values determine the result of the function. However, probability functions have a few properties that distinguish them from other functions. The first is that the function must be non-negative for all possible values of the random variable, i.e. $[x] \geq 0$. The second requirement is that the integral of a pdf must be unity, $\int_{-\infty}^{\infty}[x]\,\mathrm{d}x = 1$, and similarly for a pmf, the summation over all possible values is unity, $\sum_x [x] = 1$. The following **R** code demonstrates this for the normal and binomial distributions:

```
> integrate(dnorm, -Inf, Inf, mean=0, sd=1)$value
[1] 1
> sum(dbinom(0:5, size=5, p=0.1))
[1] 1
```

This requirement is important to remember when one develops a non-standard probability distribution. For example, in Chapt. 11 and 13, we work with resource selection functions whose probability density function is not one that is pre-defined in software packages such as **R** or **BUGS**.

Another feature of probability distributions is that they can be used to compute important summaries of random variables. The two most important summaries are the expected value, $\mathbb{E}(x)$, and the variance $\mathrm{Var}(x)$. The expected value, or mean, can be thought of as the average of a very large sample from the specified distribution. For example, one way of approximating the expected values of a binomial distribution with $K = 20$ trials and $p = 0.35$ can be implemented in **R** using:

```
> mean(rbinom(10000, 20, 0.3))
[1] 6.9865
```

For most probability distributions used in this book, the expected values are known exactly, as shown in Table 2.1, and thus we don't need to resort to such Monte Carlo approximations. For instance, the expected value of the binomial distribution is exactly $\mathbb{E}(x) = Kp = 20 \times 0.35 = 7$. In this case, it happens to take an integer value, but this is not a necessary condition, even for discrete random variables.

A more formal definition of an expected value is the average of all possible values of the random variable, weighted by their probabilities. For continuous random variables, this weighted average is found by integration:

$$\mathbb{E}(x) = \int_{-\infty}^{\infty} x \times [x]\,\mathrm{d}x. \tag{2.1.3}$$

For example, if $[x]$ is normally distributed with mean 3 and unit variance, we could find the expected value using the following code.

```
> integrate(function(x) x*dnorm(x, 3, 1), -Inf, Inf)
3 with absolute error < 0.00033
```

Of course, the mean *is* the expected value of the normal distribution, so we didn't need to compute the integral but, the point is, that Eq. 2.1.3 is generic. For discrete random variables, the expected value is found by summation rather than integration:

$$\mathbb{E}(x) = \sum_x x \times [x] \tag{2.1.4}$$

where the summation is over all possible values of x. Earlier we approximated the expected value of the binomial distribution with $K = 20$ trials and $p = 0.35$ by taking a Monte Carlo average. Eq. 2.1.4 let's us find the exact answer, using this bit of **R** code:

```
> sum(dbinom(0:100, 20, 0.35)*0:100)
[1] 7
```

This is great. But of what use is it? One very important concept to understand is that when we fit models, we are often modeling changes in the expected value of some random variable. For example, in Poisson regression, we model the expected value of the random variable, which may be a function of environmental variables.

The ability to model the expected value of a random variable gets us very far, but we also need a model for the variance of the random variable. The variance describes the amount of variation around the expected value. Specifically, $\mathrm{Var}(x) = \mathbb{E}((x - \mathbb{E}(x))^2)$. Clearly, if the variance is zero, the variable is not random as there is no uncertainty in its outcome. For some distributions, notably the normal distribution, the variance is a parameter to be estimated. Thus, in ordinary linear regression, we estimate both the expected value $\mu = \mathbb{E}(x)$, which may be a function of covariates, and the variance σ^2, or similarly the residual standard error σ. For other distributions, the variance is not an explicit parameter to be estimated, and instead, the mean to variance ratio is fixed. In the case of the Poisson distribution, the mean is equal to the variance, $\mathbb{E}(x) = \mathrm{Var}(x) = \lambda$. A similar situation is true for the binomial distribution—the variance is determined by the two parameters K and p, $\mathrm{Var}(x) = Kp(1-p)$. In our earlier example with $K = 20$ and $p = 0.35$, the variance is 4.55. Toying around with these ideas using random number generators may be helpful. Here is some code to illustrate some of these basic concepts:

```
> 20*0.35*(1-0.35)              # Exact variance, Var(x)
[1] 4.55
> x <- rbinom(100000, 20, 0.35)
> mean((x-mean(x))^2)           # Monte Carlo approximation
[1] 4.545525
```

2.2 COMMON PROBABILITY DISTRIBUTIONS

We got a little ahead of ourselves in the previous sections by using the binomial and Poisson distributions without describing them in detail. A solid understanding of the binomial, Poisson, multinomial, uniform, and normal (or Gaussian) distributions is absolutely essential throughout the remainder of the book. We will occasionally make use of other distributions such as the beta, log-normal, gamma, Dirichlet, etc... that can be helpful when modeling capture-recapture data, but these distributions can be readily understood once you are comfortable with the more commonly used distributions described in this section.

2.2.1 The binomial distribution

The binomial distribution plays a critical role in ecology. It is used for purposes as diverse as modeling count data, survival probability, occurrence probability, and capture probability, just to name a few. To describe the properties of the binomial distribution, and related distributions, we will introduce a new example. Suppose we are conducting a bird survey at a site in which $N = 10$ chestnut-sided warblers (*Stetophaga pensylvanica*) occur, and each of these individuals has a detection probability of $p = 0.5$. The binomial distribution is the natural choice for describing the number of individuals that we would expect to detect (n) in this situation, and using our notation, we can write the model as: $n \sim \text{Binomial}(10, 0.5)$. When $p < 1$, we can expect that we will observe a different number of warblers on each of K replicate survey occasions. To see this, we simulate data under this simple model with $K = 3$.

```
> n <- rbinom(3, size=10, prob=0.5) # Generate 3 binomial outcomes
> n                                 # Display the 3 values
[1] 6 4 8
```

The vector of counts will typically differ each time you issue this command; however, we know the probability of observing any value of n_k because it is defined by the binomial pmf. As we demonstrated earlier, in **R** this probability can be found using the `dbinom` function. For example, the probability of observing $n_k = 5$ is given by:

```
> dbinom(5, 10, 0.5)
```

This simply evaluates the function shown in Table 2.1. We could do the same more transparently, but less efficiently, using any of the following:

```
> n <- 5; N <- 10; p <- 0.5
> factorial(N)/(factorial(n)*factorial(N-n))*p^n*(1-p)^(N-n)
> exp(lgamma(N+1) - (lgamma(n+1) + lgamma(N-n+1)))*p^n*(1-p)^(N-n)
> choose(N, n)*p^n*(1-p)^(N-n)
```

Note that the last three lines of code differ only in how they compute the binomial coefficient $\binom{N}{n}$, which is the number of different ways we could observe $n = 5$ of the $N = 10$ chestnut-sided warblers at the site. The binomial coefficient, which is read "N choose n" is defined as

$$\binom{N}{n} = \frac{N!}{n!(N-n)!}. \tag{2.2.1}$$

Now that we know how to simulate binomial data and compute the probabilities of observing any particular outcome n, conditional on the parameters N and p, we can contemplate the relevance of the binomial distribution in spatial capture-recapture models. One important application of the binomial distribution is as a model encounter frequencies. Indeed, one of the most important encounter models in SCR will be referred to as the "binomial encounter model", in which the number of times individual i is captured at "trap" j after K survey occasions is modeled as $y_{ij} \sim \text{Binomial}(K, p_{ij})$. Here, p_{ij} is the encounter probability determined, in part, by the distance between an animal's activity center and the trap location. This binomial encounter model is described in detail in Sec. 7.1. Another important application of the binomial distribution is as a prior for the population size parameter in Bayesian analyses, as is discussed in Chapt. 4.

2.2.2 The Bernoulli distribution

Above, we showed 3 alternatives to `dbinom` for evaluating the binomial pmf. These three commands differed only in how they computed the binomial coefficient, which we needed because of the numerous ways in which we could observe $n = 5$ given $N = 10$. To conceptualize this, let y_i be a binary variable indicating ~~if~~ individual i was detected ~~or not~~. Hence, given that 5 individuals were detected, the vector of individual detections could be something like $\mathbf{y} = (0, 0, 1, 1, 1, 1, 1, 0, 0, 0)$, indicating that we detected individuals 3-7 but not 1-2 or 8-10. For $N = 10$ and $n = 5$, the binomial coefficient tells us that there are 252 possible vectors $\mathbf{y}$ with 5 ones. However, when $N \equiv 1$, this term drops from the pmf and the result is the pmf for the Bernoulli distribution. That is, the Bernoulli distribution is simply the binomial distribution when $N \equiv 1$. Alternatively, we could say that the binomial distribution is the outcome of N *iid* Bernoulli trials. We use the standard abbreviation "*iid*" to mean *independent, identically distributed*.

The utility of the Bernoulli distribution is evident when we imagine that not all of the chestnut-sided warblers have the same detection probability. Thus, if some individuals can be detected with probability 0.3 and others have a 0.7 detection probability, then the model $n \sim \text{Binomial}(N, p)$ is no longer an accurate description of system since p is no longer constant for all individuals.

To properly account for variation in p, we could redefine our model for the

counts of chestnut-sided warblers as

$$y_{ik} \sim \text{Bernoulli}(p_i)$$
$$n_k = \sum_{i=1}^{N} y_{ik} \tag{2.2.2}$$

This states that individual i is detected with probability p_i, and the observed count is the sum of the N Bernoulli outcomes.

An important point is that the individual-specific data y_{ik} can only be observed if the individuals are uniquely distinguishable, such as when they are marked by biologists with color bands. In such cases, the Bernoulli distribution allows us to model variation in detection probability among individuals and thus would be preferable to the binomial distribution, which assumes that each of the N individuals have the same p. For this reason, the Bernoulli distribution, as simple as it is, is of paramount importance in capture-recapture models, including spatial capture-recapture models in which there is virtually always substantial and important variation in capture probability among individuals. Indeed, it could be said that the Bernoulli model is the canonical model in capture-recapture studies, and most of the different flavors of capture-recapture models differ primarily in how p_i is specified.

The Bernoulli pmf is given by $p^n(1-p)^{1-n}$ and hence we do not need canned functions to facilitate its evaluation. Of course, if you wanted to, you could always use `dbinom` with the `size` argument set to 1. For example, `dbinom(1, 1, 0.3)` returns the Bernoulli probability of observing $n = 1$ given $p = 0.3$.

2.2.3 The multinomial and categorical distributions

The binomial distribution is used when we are accumulating a binary response—that is, one in which there are two possible categories such as success/failure or captured/not-captured. The multinomial distribution is a multivariate extension of the binomial used when there are $G > 2$ categories. The multinomial distribution can be thought of as a model for placing N items in the G categories, which are also called bins or cells. Each bin has its own probability π_g and these probabilities must sum to one. In ecology, N is often population size or the number of individuals detected, but the definition of the G bins varies among applications. For example, in distance sampling, when the distance data are aggregated into intervals, the bins are the distance intervals, and the cell probabilities are functions of detection probability in each interval (Royle et al., 2004).

The multinomial distribution is widely used to model data from traditional, non-spatial capture-recapture studies. Earlier we let y_{ik} denote a binary random variable indicating if warbler i was detected on survey k. The vector of observations for an individual, $\mathbf{y}_i$, is often referred to as the individual's "encounter history".

The number of possible encounter histories depends on K, the number of survey occasions. Specifically, there are 2^K possible encounter histories[1]. If we tabulate the number of individuals with each encounter history, the frequencies can be modeled using the multinomial distribution.

Going back to our chestnut-sided warbler example, suppose the 10 individuals are marked and we make $K = 2$ visits to the site such that there are $2^K = 4$ possible encounter histories: $(11, 10, 01, 00)$, where, for example, "10" is the encounter history for an individual detected on the first visit but not the second. If $p = 1$, then the encounter history for each of the 10 individuals must be "11". That is, we would detect each individual on both occasions. In this case, ~~we~~ the data would be: $\mathbf{h} = (10, 0, 0, 0)$, which indicates that all 10 warblers had the first encounter history. The corresponding cell probabilities would be $\boldsymbol{\pi} = (1, 0, 0, 0)$. What about the situation where $p < 1$, e.g. $p = 0.3$? In this case, the probability of observing the capture history "11" (detected on both occasions) is $p \times p = 0.3 \times 0.3 = 0.09$. The probability of observing "10" is $p \times (1 - p) = 0.21$. Following this logic, the vector of cell probabilities is $\boldsymbol{\pi} = (0.09, 0.21, 0.21, 0.49)$. We can simulate data under this model as follows:

```
> caphist.probs <- c("11"=0.09, "10"=0.21, "01"=0.21, "00"=0.49)
> drop(rmultinom(1, 10, caphist.probs))
11 10 01 00
 0  3  2  5
```

The result of our simulation is that zero individuals were observed with the capture history "11" and 5 individuals were observed with the capture history "00". The other 5 individuals were observed one out of the two occasions. This is not such a surprising outcome given $p = 0.3$.

As in non-spatial capture-recapture studies, the multinomial distribution turns out to be very important in spatial capture-recapture studies. However, N is not defined as population size. Rather, we use the multinomial distribution when an individual can only be captured in a single trap during an occasion. Thus $N = 1$ and the cell probabilities are the probabilities of being captured in each trap. A thorough discussion of this point can be found in Chapt. 9. Another application of the multinomial distribution in SCR models is discussed in Chapt. 11 where we discuss how to model the probability that an individual's activity center is located in one of the cells of a raster defining the spatial region of interest.

Just as the Bernoulli distribution is the elemental form of the binomial distribution (being the case $N = 1$), the categorical distribution is essentially equivalent to the multinomial distribution with size parameter $N \equiv 1$. The only difference is that, rather than returning a vector with a single element equal to 1, it returns the element *location* where the 1 occurs. For example, if $\mathbf{y} = (0, 0, 1, 0)$ is an outcome

[1] When N is unknown, we can never observe the "all-0" encounter history, corresponding to an individual that is not detected, and thus the number of "observable" encounter histories is 2^{K-1}

of a multinomial distribution with $N = 1$, then the categorical outcome would be 3 because the 1 is located in third position in the vector. Thus, in spatial capture-recapture models, we might use either the multinomial distribution with $N = 1$ or the categorical distribution. The various **BUGS** engines describe the categorical distribution by the declaration `dcat` and, in **R**, we can simulate categorical outcomes using the function `sample` or as so:

```
> which(rmultinom(1, 1, c(0.1, 0.7, 0.2)) == 1)
[1] 2
```

2.2.4 The Poisson distribution

The Poisson distribution is the canonical model for count data in ecology. More generally, the Poisson distribution is a model for random variables taking on non-negative, integer values. Although it is a simple model having just one parameter, $\lambda = \mathbb{E}(x) = \text{Var}(x)$, its applications are highly diverse, including as a model of spatial variation in abundance or as a model for the frequency of behaviors over time. Just as logistic regression is the standard generalized linear model (GLM) used to model binary data, Poisson regression is the default GLM for modeling count data and variation in λ.

The Poisson distribution is related to both the binomial and multinomial distributions, and the following three bits of trivia are occasionally worth knowing. First, it is the limit of the binomial distribution as $N \to \infty$ and $p \to 0$, which means that for high values of N and low values of p, Poisson$(N \times p)$ is approximately equal to Binomial(N, p). Second, if $\{n_1 \sim \text{Poisson}(\lambda_1), \ldots, n_K \sim \text{Poisson}(\lambda_K)\}$ then the vector of counts is multinomial, $\{n_1, \ldots, n_K\} \sim \text{Multinomial}(\sum_k n_k, \{\frac{\lambda_1}{\sum_k \lambda_k}, \ldots, \frac{\lambda_K}{\sum_k \lambda_k}\})$. Third, the sum of two Poisson random variables $x_1 \sim \text{Poisson}(\lambda_1)$ and $x_2 \sim \text{Poisson}(\lambda_2)$ is also Poisson: $x_1 + x_2 \sim \text{Poisson}(\lambda_1 + \lambda_2)$.

The Poisson distribution has two important uses in spatial capture-recapture models: (1) as a prior distribution for the population size parameter N, and (2) as a model for the frequency of captures in a trap. In the first context, the Poisson prior for N results in a Poisson point process for the location of the N activity centers in the region of interest. This topic is discussed in Chapt. 5 and Chapt 11. The second use of the Poisson distribution in spatial capture-recapture is to describe data from sampling methods in which an individual can be detected multiple times at a trap during a single occasion. For example, in camera trapping studies we might obtain multiple pictures of the same individual at a trap during a single sampling occasion. Thus, λ in this case would be defined as the expected number of detections or captures per occasion.

2.2.5 The uniform distribution

The lowly uniform distribution is a continuous distribution whose only two parameters are the lower and upper bounds that restrict the possible values of the

random variable x. These bounds are almost always known, so there is typically nothing to estimate. Nonetheless, the uniform distribution is one of the most widely used distributions, especially among Bayesians who frequently use it to as a "non-informative" prior distribution for a parameter. For example, if we have a capture probability parameter p that we wish to estimate, but we have no prior knowledge of what value it may take in the range [0,1], we will often use the prior $p \sim \text{Uniform}(0, 1)$. This states that p is equally likely to take on any value between zero and one. Prior distributions are described in more detail in the next chapter.

Another common usage of the uniform distribution is as a prior for the coordinates of points in the real plane, i.e. in two-dimensional space. Such a use of the uniform distribution implies that a point process is "homogeneous", meaning that the location of one point does not affect the location of another point and that the expected density of points is constant throughout the region. Thus, to simulate a realization from a homogeneous Poisson point process in the unit square $[0, 1] \times [0, 1]$, we could use the following **R** code:

```
D <- 100      # points per unit area
A <- 1        # Area of unit square
N <- rpois(1, D*A)
plot(s <- cbind(runif(N), runif(N)))
```

where **s** is a matrix of coordinates with N rows and 2 columns. We will often represent the uniform point process using the following notation:

$$\mathbf{s} \sim \text{Uniform}(\mathcal{S}) \tag{2.2.3}$$

where $\mathcal{S}$ is some specific unit of space called the state-space of the random variable **s**. It would be more correct to ~~somehow~~ distinguish this two-dimensional uniform distribution ~~for~~ the univariate one. That is, it might be more clear to use notation such as $\mathbf{s} \sim \text{Uniform}_2(\mathcal{S})$ instead, but this is somewhat cumbersome, so we will opt for the former expression.

2.2.6 Other distributions

The other continuous distributions that are regularly encountered in SCR models are primarily used as priors in Bayesian analyses, and thus we will avoid a lengthy discussion of their properties. The normal distribution, also called the Gaussian distribution, is perhaps the most widely recognized and applied probability model in statistics, but it plays only a minor role in SCR models other than as a model for signal strength in acoustic SCR models (Efford et al., 2009b; Dawson and Efford, 2009~~)~~, and see Sec. 9.4. Nonetheless, it is the canonical prior for any continuous random variable with infinite support, and thus it is often used as a prior when applying Bayesian methods. One common usage is as a prior for the β coefficients of a linear model defining some parameter as a function of covariates (usually on

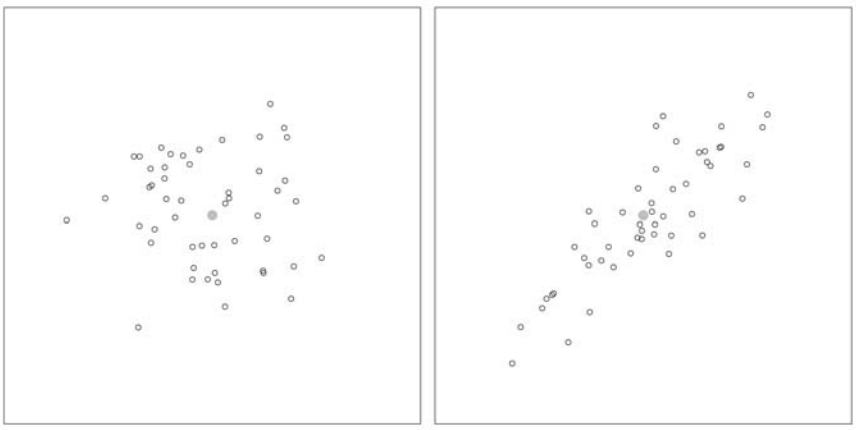

Figure 2.2. Two realized point patterns from the bivariate normal distribution.

a transformed scale). An example, including a cautionary note, is provided in Sec. 3.5.1. Be aware that although the normal distribution is typically parameterized in terms of the variance parameter σ^2, in the **BUGS** language, the inverse of the variance, or precision, is used instead, $\tau = 1/\sigma^2$. In **R**, the `dnorm` function requires the standard deviation σ, rather than the variance σ^2.

The bivariate normal distribution is a generalization of the normal distribution and a special case of the multivariate normal distribution whose pdf is shown in Table 2.1. The bivariate normal distribution is used to model two (possibly) dependent continuous variables whose symmetric variance-covariance matrix is denoted $\mathbf{\Sigma}$. In SCR models, we most often use this model as a rudimentary description of movement outcomes about a home range center. If there is no correlation, then the model reduces to two independent normal draws along the coordinate axes. The following code generates bivariate normal outcomes with no correlation ($\rho = 0$), as well as outcomes in which the correlation is $\rho = 0.9$.

```
library(mvtnorm)
set.seed(3)
mu <- c(0,0)
Sigma <- matrix(c(1, .9, .9, 1), 2, 2)
X1 <- cbind(rnorm(50, mu[1], Sigma[1,1]), # No correlation (rho=0)
            rnorm(50, mu[2], Sigma[2,2]))
X2 <- rmvnorm(50, mu, Sigma)              # rho=0.9
```

Fig. 2.2 shows the simulated points.

Several of the parameters in capture-recapture models do not have infinite support, but instead are probabilities restricted to the range $[0, 1]$, or are positive

valued living between zero and ∞. The beta distribution is the standard prior used for probabilities because it can be used to express either a lack of knowledge or very precise knowledge about a parameter. For example, a Beta$(1, 1)$ distribution is equivalent to a Uniform$(0, 1)$ distribution. However, unlike the the uniform distribution, the beta distribution can be used as an informative prior; for example if published estimates of detection probability exist we can choose parameters of the beta distribution to reflect that. To gain some familiarity with the beta distribution, execute the following **R** commands:

```
curve(dbeta(x, 1, 1), col="black", ylim=c(0,5))
curve(dbeta(x, 10, 10), col="blue", add=TRUE)
curve(dbeta(x, 10, 20), col="darkgreen", add=TRUE)
```

Other parameters in SCR models are continuous but positive-valued and can be modeled using the gamma distribution. As with the beta distribution, the gamma distribution is typically favored over the uniform distribution when one is interested in using an informative prior. It is also frequently used as a vague prior for the inverse of variance parameters, but it is wise to compare this prior to a uniform to assess its influence on the posterior.

2.3 STATISTICAL INFERENCE AND PARAMETER ESTIMATION

If the parameters of a statistical model were known with absolute certainty, then it would be possible to use pdfs and pmfs to make direct probability statements about unknowns such as future outcomes. However, we almost never know the actual values of parameters, and instead we have to estimate them from observations (i.e., data). Our inferences must then acknowledge the uncertainty associated with our imperfect knowledge of the parameters. Doing so is most often accomplished using one of two approaches: classical (frequentist) inference or Bayesian inference. These two modes of inference regard the uncertainty about parameters in entirely different ways. In the next chapter, we will review some of the important concepts in Bayesian inference, so here, we will focus on the frequentist perspective.

Suppose we count oak trees at J sites, and the resulting data $\{y_1, \ldots, y_J\}$ can be assumed to be *iid* outcomes from some distribution, such as the Poisson with unknown parameter λ. We want to estimate this parameter. In classical inference, the only uncertainty about λ is that attributable to sampling. For instance, we can imagine repeatedly sampling the population (sites in this example) and obtaining sample-specific estimates of λ. Typically, we entertain the idea that there are an infinite number of possible samples and so we could obtain an infinite number of estimates: $\{\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_\infty\}$. If these estimates are produced using the method of maximum likelihood, and as n tends to infinity, the distribution of estimates, called the sampling distribution, will be normally distributed with $\mathbb{E}(\hat{\lambda}) = \lambda$. The standard deviation of the sampling distribution is called the standard error, which can also be estimated as part of the maximum likelihood procedure. Of course, we

almost always have just a single sample of data, and hence a single $\hat{\lambda}$ and a single estimate of the standard error. However, under the assumption of a normally distributed sampling distribution, we can construct a confidence interval that will include the true value of λ with coverage probability $1-\alpha$, where α is a prescribed value like 0.05. An important point is that there is no uncertainty associated with the actual parameter—it is regarded as a fixed value, and hence probability is only used to characterize the estimator via its sampling distribution.

Maximum likelihood is heuristically a method of finding the most "likely" value of λ, given the observed data, and of characterizing the variance of the sampling distribution. Of course, it also applies to cases where the observations are multivariate, or the probability distribution is a function of multiple parameters. Endless numbers of textbooks and online resources are available for those interested in a detailed explanation of maximum likelihood. For our purposes, we wish to keep it simple and focus on *how* to do it. The first step is to define the likelihood function, which is is the joint distribution of the data regarded as a function of the parameter(s). If the joint distribution of the observations is denoted by $[y_1, y_2, \ldots, y_n|\lambda]$, we usually denote the likelihood by flipping the arguments: $\mathcal{L}(\lambda|\mathbf{y}) = [\lambda|y_1, y_2, \ldots, y_n]$.

If the observations are *iid*, the likelihood simplifies to

$$\mathcal{L}(\lambda|\mathbf{y}) = \prod_{i=1}^{n} [y_i|\lambda]. \tag{2.3.1}$$

where $[y_i|\lambda]$ is a probability distribution, like those discussed in the previous sections. For example, if y_i is Poisson distributed, then $[y_i|\lambda] = \text{Poisson}(\lambda) = \frac{\lambda^{y_i} e^{-\lambda}}{y_i!}$. Although likelihoods are typically shown on the natural scale, we almost always maximize the logarithm of the likelihood to avoid computational problems that arise when multiplying very small probabilities. Thus, we rewrite Eq. 2.3.1 as

$$\ell(\lambda|\mathbf{y}) = \sum_{i=1}^{n} \log(f(y_i|\lambda)) \tag{2.3.2}$$

Here is some simple **R** code to simulate independent Poisson outcomes and estimate λ (as though we did not know it) using the method of maximum likelihood. Actually, we will minimize the negative log-likelihood because it is equivalent and is the default for **R**'s optimizers like `optim` and `nlm`.

```
> lambda <- 3                  # Actual parameter value
> y1 <- rpois(100, lambda)     # Realized values (data)
> negLogLike1 <- function(par) -sum(dpois(y1, par, log=TRUE))
> starting.value <- c('lambda'=1)
> optim(starting.value, negLogLike1)$par # MLE
  lambda
3.039844
```

Explicitly maximizing the likelihood, numerically, isn't actually necessary here because the MLE of λ is given by the mean of the observations. A more interesting example is when there are covariates of λ. For example, suppose λ is a function of elevation and vegetation height according to: $\log(\lambda_i) = \beta_0 + \beta_1 \texttt{ELEV}_i + \beta_2 \texttt{VEGHT}_i$. This is a standard Poisson regression problem, with likelihood:

$$\mathcal{L}(\boldsymbol{\beta}|\mathbf{y}) = \prod_i \text{Poisson}(y_i|\lambda_i) \tag{2.3.3}$$

This likelihood is almost identical to the previous one except that λ is now a function, and so we need to estimate the parameters of the function, i.e. the β's. Some code to fit this model to simulated data is shown here:

```
> nsites <- 100
> elevation <- rnorm(100)
> veght <- rnorm(100)
> beta0 <- 1
> beta1 <- -1
> beta2 <- 0
> lambda <- exp(beta0 + beta1*elevation + beta2*veght)
> y2 <- rpois(nsites, lambda)
> negLogLike2 <- function(pars) {
+     beta0 <- pars[1]
+     beta1 <- pars[2]
+     beta2 <- pars[3]
+     lambda <- exp(beta0 + beta1*elevation + beta2*veght)
+     -sum(dpois(y2, lambda, log=TRUE))
+ }
> starting.values <- c('beta0'=0, 'beta1'=0, 'beta2'=0)
> optim(starting.values, negLogLike2)$par
      beta0       beta1       beta2
 0.98457756 -1.03025173 -0.01218292
```

We see that the maximum likelihood estimates (MLEs) are very close to the true parameter values.

In these examples, the parameters we estimated are called fixed effects by frequentists. Fixed effects are parameters that are not regarded as being random variables. A random effect, in contrast, is a parameter that can be regarded as the outcome of a random variable. For instance, we could entertain the idea that the intercept of our GLM differs among locations, and that its actual value is an outcome of a normal distribution with parameters μ and σ^2. In this case, β_i would

be a random effect, and our model could be written:

$$y_i \sim \text{Poisson}(\lambda_i)$$
$$\log(\lambda_i) = \beta_i + \beta_1 \text{ELEV}_i + \beta_2 \text{VEGHT}_i$$
$$\beta_i \sim \text{Normal}(\mu, \sigma^2)$$

This is an example of a mixed effects model or a hierarchical model. How do we estimate the parameters of a model that includes random effects? Earlier the likelihood function was written as the product of probabilities determined by a single pmf or pdf, $[y|\lambda]$, but now we have an additional random variable, and we are forced to think about conditional relationships, because y depends upon β_i and β_i depends upon other parameters, specifically μ and σ^2. This type of conditional dependence among parameters is the essence of hierarchical models, and statistical analysis of hierarchical models requires that we discuss joint distributions, marginal distributions and conditional distributions. These concepts will be used extensively in Chapt. 6 where we demonstrate how to estimate parameters of hierarchical models using maximum likelihood.

2.4 JOINT, MARGINAL, AND CONDITIONAL DISTRIBUTIONS

So far we have restricted our attention to situations in which we wish to make inference about a single random variable. However, in ecology, we often are interested in multiple random variables and how they are related. Let Y be a random variable that may or may not be independent of X (here again we will distinguish between random variables and realized values for conceptual clarity). Inference about these two random variables can be made using the joint, marginal, or conditional distributions—or, we may make use of all of them depending on the question being asked. In the case of discrete random variables, the joint distribution is the probability that X takes on the value x *and* that Y takes on the value y, which is written $[X = x, Y = y]$. To clarify this concept, let's go back to our original example where X was the number of fish caught after 20 casts, which we said was an *iid* binomial random variable. Now, let's suppose that X depends on the random variable Y, which is the number of other fisherman at the hole. Specifically, let's say that the probability of catching a fish p is related to Y according to $\text{logit}(p) = -0.6 + -2y$. Furthermore, let's make the intuitive assumption that the number of fishermen at the hole is a Poisson random variable with mean 0.6, i.e. $Y \sim \text{Poisson}(0.6)$. Our model is now fully specified, and so we can answer the question: "what is the probability of catching x fish *and* of there being y fishermen at the hole". This joint distribution is given by the product of the binomial pmf (with p determined by y) and the Poisson pmf with $\lambda = 0.6$. The following **R** code creates the joint distribution.

```
> X <- 0:20  # All possible values of X
> Y <- 0:10  # All possible values of Y
> lambda <- 0.6
```

```
> p <- plogis(-0.62 + -2*Y) # p as function of Y
> round(p,2)
 [1] 0.35 0.07 0.01 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00
> joint <- matrix(NA, length(X), length(Y))
> rownames(joint) <- paste("X=", X, sep="")
> colnames(joint) <- paste("Y=", Y, sep="")
>
> # Joint distribution [X,Y]
> for(i in 1:length(Y)) {
+     joint[,i] <- dbinom(X, 20, p[i]) * dpois(Y[i], lambda)
+ }
> round(joint,2)
      Y=0  Y=1  Y=2  Y=3 Y=4 Y=5 Y=6 Y=7 Y=8 Y=9 Y=10
X=0  0.00 0.08 0.08 0.02   0   0   0   0   0   0    0
X=1  0.00 0.12 0.02 0.00   0   0   0   0   0   0    0
X=2  0.01 0.08 0.00 0.00   0   0   0   0   0   0    0
X=3  0.02 0.04 0.00 0.00   0   0   0   0   0   0    0
X=4  0.04 0.01 0.00 0.00   0   0   0   0   0   0    0
X=5  0.07 0.00 0.00 0.00   0   0   0   0   0   0    0
X=6  0.09 0.00 0.00 0.00   0   0   0   0   0   0    0
X=7  0.10 0.00 0.00 0.00   0   0   0   0   0   0    0
X=8  0.09 0.00 0.00 0.00   0   0   0   0   0   0    0
X=9  0.06 0.00 0.00 0.00   0   0   0   0   0   0    0
X=10 0.04 0.00 0.00 0.00   0   0   0   0   0   0    0
X=11 0.02 0.00 0.00 0.00   0   0   0   0   0   0    0
X=12 0.01 0.00 0.00 0.00   0   0   0   0   0   0    0
X=13 0.00 0.00 0.00 0.00   0   0   0   0   0   0    0
X=14 0.00 0.00 0.00 0.00   0   0   0   0   0   0    0
X=15 0.00 0.00 0.00 0.00   0   0   0   0   0   0    0
X=16 0.00 0.00 0.00 0.00   0   0   0   0   0   0    0
X=17 0.00 0.00 0.00 0.00   0   0   0   0   0   0    0
X=18 0.00 0.00 0.00 0.00   0   0   0   0   0   0    0
X=19 0.00 0.00 0.00 0.00   0   0   0   0   0   0    0
X=20 0.00 0.00 0.00 0.00   0   0   0   0   0   0    0
```

This matrix tells us the probability of all possible combinations of x and y, and we see that the most likely value is $(X = 1, Y = 1)$, i.e. we will catch 1 fish and there will be 1 other fisherman. This matrix also demonstrates the law of total probability, which dictates that the sum of of these probabilities must equal 1.

Perhaps most fisherman don't care about joint distributions, but a question that might be asked is "what is the probability of catching 1 fish today?" We know that this depends on the number of fisherman, but we don't know how many will show up today, so this is a different question than "what is most likely value of X and Y". This brings us to the marginal distribution, which is defined by

$$[X] = \sum_Y [X, Y] \qquad [Y] = \sum_X [Y, X]$$

for discrete random variables, and

$$[X] = \int_{-\infty}^{\infty} [X, Y] \, \mathrm{d}Y \qquad [Y] = \int_{-\infty}^{\infty} [Y, X] \, \mathrm{d}X$$

for continuous random variables. The key idea here is that to get the marginal distribution of X, we have to contemplate all possible values of Y. Computing marginal distributions is a key step in maximizing likelihoods involving random effects, as will be demonstrated in Chapt.6. Here is some **R** code to compute the marginal distribution of X, i.e. the probability of catching $X = x$ fish:

```
> margX <- rowSums(joint)
> round(margX, 2)
 X=0  X=1  X=2  X=3  X=4  X=5  X=6  X=7  X=8  X=9 X=10 X=11 X=12 X=13 X=14
0.18 0.14 0.09 0.05 0.05 0.07 0.09 0.10 0.09 0.06 0.04 0.02 0.01 0.00 0.00
X=15 X=16 X=17 X=18 X=19 X=20
0.00 0.00 0.00 0.00 0.00 0.00
```

Bad news—the most likely value is $X = 0$. However, the chances of catching 1 fish is pretty similar.

The last type of question we can ask about these two random variables relates to their conditional distributions. The conditional probability distribution is the distribution of one variable, given a realized value of the other. In the case of two discrete random variables, the conditional distribution may be written as $[X = x|Y = y]$, i.e. the probability of X taking on the value x given the realized value of Y being y. For simplicity, we will write this as $[X|Y]$. Conditional distributions are defined as follows:

$$[X|Y] = \frac{[X, Y]}{[Y]} \qquad [Y|X] = \frac{[X, Y]}{[X]}.$$

That is, the conditional distribution of X given Y is the joint distribution divided by the marginal distribution of Y.

```
> XgivenY <- joint/matrix(margY, nrow(joint), ncol(joint), byrow=TRUE)
> round(XgivenY, 2)
      Y=0  Y=1  Y=2  Y=3 Y=4 Y=5 Y=6 Y=7 Y=8 Y=9 Y=10
X=0  0.00 0.25 0.82 0.97   1   1   1   1   1   1    1
X=1  0.00 0.36 0.16 0.03   0   0   0   0   0   0    0
X=2  0.01 0.25 0.02 0.00   0   0   0   0   0   0    0
X=3  0.03 0.11 0.00 0.00   0   0   0   0   0   0    0
X=4  0.07 0.03 0.00 0.00   0   0   0   0   0   0    0
X=5  0.13 0.01 0.00 0.00   0   0   0   0   0   0    0
X=6  0.17 0.00 0.00 0.00   0   0   0   0   0   0    0
X=7  0.18 0.00 0.00 0.00   0   0   0   0   0   0    0
```

```
X=8  0.16 0.00 0.00 0.00   0   0   0   0   0   0   0
X=9  0.12 0.00 0.00 0.00   0   0   0   0   0   0   0
X=10 0.07 0.00 0.00 0.00   0   0   0   0   0   0   0
X=11 0.03 0.00 0.00 0.00   0   0   0   0   0   0   0
X=12 0.01 0.00 0.00 0.00   0   0   0   0   0   0   0
X=13 0.00 0.00 0.00 0.00   0   0   0   0   0   0   0
X=14 0.00 0.00 0.00 0.00   0   0   0   0   0   0   0
X=15 0.00 0.00 0.00 0.00   0   0   0   0   0   0   0
X=16 0.00 0.00 0.00 0.00   0   0   0   0   0   0   0
X=17 0.00 0.00 0.00 0.00   0   0   0   0   0   0   0
X=18 0.00 0.00 0.00 0.00   0   0   0   0   0   0   0
X=19 0.00 0.00 0.00 0.00   0   0   0   0   0   0   0
X=20 0.00 0.00 0.00 0.00   0   0   0   0   0   0   0
```

Note that we have 11 probability distributions for X, one for each possible value of Y, and each pmf sums to unity as it should. Note also that if you show up at the hole and there are > 2 fisherman, your chance of catching a fish is very low. Go home. These concepts are explained in more detail in other texts such as Casella and Berger (2002), Royle and Dorazio (2008), and Link and Barker (2010), but hopefully, the code shown here complements the equations and makes it easier for non-statisticians to understand these concepts.

The last point we wish to make in the section is that this simple example *is* a hierarchical model, and we can put the pieces together using the following notation:

$$Y \sim \text{Poisson}(0.6) \tag{2.4.1}$$

$$\text{logit}(p) = -0.6 + -2Y \tag{2.4.2}$$

$$X|Y \sim \text{Binomial}(20, p) \tag{2.4.3}$$

From here on out, when you see such notation, you should immediately grasp the fact that Y is a random variable independent of X, but X depends upon Y through p. Now you have the tools to make probability statements about the random variables in this system. The one caveat faced in reality is that we typically do not know the values of the parameters, and instead we have to estimate them. Maximum likelihood methods for hierarchical models are covered in Chapt. 6.

2.5 HIERARCHICAL MODELS AND INFERENCE

The term hierarchical modeling (or hierarchical model) has become something of a buzzword over the last decade with hundreds of papers published in ecological journals using that term. So then, what exactly is a hierarchical model, anyhow? Obviously, this term stems from the root "hierarchy" which means:

Definition: *hierarchy* (noun) – a series of ordered groupings of people or things within a system;

In the case of a hierarchical model (hierarchical being the adjective form of hierarchy), the "things" are probability distributions, and they are ordered according to their conditional probability structure. Thus, a hierarchical model is *an ordered series of models, ordered by their conditional probability structure.*

A canonical hierarchical model in ecology is this elemental model of species occurrence or distribution (MacKenzie et al., 2002; Tyre et al., 2003; Kéry, 2011):

$$y_i|z_i \sim \text{Binomial}(K, z_i\, p)$$

$$z_i \sim \text{Bernoulli}(\psi)$$

where y_i = observation of presence/absence at a site i and z_i = occurrence status ($z_i = 1$ if a species occurs at site i and $z_i = 0$ if not). Note that if $p = 1$, then we would perfectly observe z and the model would no longer be hierarchical—it would be a simple logistic regression model. Note also that this hierarchical model has an important conceptual distinction between other types of classical multi-level models such as repeated measures on subjects, in that z_i is an actual state of nature. In that sense, z is a random variable that is the outcome of a "real" process. Royle and Dorazio (2008) used the term *explicit* hierarchical model to describe this type of model to distinguish from hierarchical models (*implicit* hierarchical models) where the latent variables don't correspond to an actual state of nature—but rather just soak up variation that is unmodeled by explicit elements of the model. At best, latent variables in such models are surrogates for something of ecological relevance ("time effects", "space effects" etc.).

With these examples, we expand on our definition of a hierarchical model as we will use it in this book:

Definition: *Hierarchical Model*: A model with explicit component models that describe variation in the data due to (spatial/temporal) variation in *ecological process*, and due to *imperfect observation* of the process.

Most models considered in this book describe the encounter of individuals conditional on the "activity center" of the individual, which is a latent variable (i.e., unobserved random effect). The definition of an activity center will be context-dependent as discussed in Chapt. 5, but often it can be thought of as an individual's home range center. The collection of these latent variables represents the outcome of an ecological process describing how individuals distribute themselves over the landscape. Moreover, how individuals are encountered in traps is, in some cases, the result of a model governing movement. As such, these models are examples of hierarchical models that contain formal model components representing both ecological process and also the observation of that process. That is, they are explicit hierarchical models (Royle and Dorazio, 2008) as opposed to implicit hierarchical models.

2.6 CHARACTERIZATION OF SCR MODELS

For the purposes of this book, an SCR model is any "individual encounter model" (not just "capture-recapture"!) where auxiliary spatial information is also obtained. To be more precise we could as well use the term "spatial capture and/or recapture" but that is slightly unwieldy and, besides, it also abbreviates to SCR. The class of SCR models includes traditional capture-recapture models with auxiliary spatial information and even some models that do not even require "recapture" (e.g., distance sampling). There is even a class of models (Chapt. 18) which don't require capture or unique identification of individuals.

Conceptually, SCR models involve a collection of random variables, $\mathbf{s}$, $\mathbf{u}$ and y where $\mathbf{s}$ is the activity center, or home range center, $\mathbf{u}$ is the location of the individual at the time of sampling, which we may think of as a realization from some movement model, and y is the "response variable"—what the observer records. For example, $y = 1$ means "detected" and $y = 0$ means "not detected", but many other types of responses are possible (Chapt 9). A broad class of models for estimating density are unified by a hierarchical model involving explicit models for animal activity centers $\mathbf{s}$, movement outcomes $\mathbf{u}$, and encounter data y. In some cases, we don't observe y but rather summaries of y, say $n(y)$, yet it might be convenient in such cases to retain an explicit focus on y in terms of model construction. We thus introduce a sequence of models—a hierarchical model—to relate these random variables, which can be written as

$$[n(y)|y][y|\mathbf{u}][\mathbf{u}|\mathbf{s}][\mathbf{s}]. \tag{2.6.1}$$

Every model we talk about in this book has a subset of these components although we never fit the full model because we have not encountered a situation requiring that we do so. However, a detailed description of this model and its various components is the subject of this book, and we will not pretend to condense hundreds of pages of material into the next few paragraphs. However, we give a cursory overview here to whet the appetite and provide some indication of where we are going. Don't worry if some of this material doesn't sink in just yet—we will walk through it slowly in the subsequent chapters.

Let's begin with the model $[\mathbf{s}]$ that describes the distribution of the activity centers of each animal in the spatial region $\mathcal{S}$ (the state-space as we called it previously). As will be explained in Chapt. 5 and Chapt. 11, $[\mathbf{s}]$ defines a spatial point process, which may be inhomogeneous if there exists spatial variation in density, or it may be homogeneous if density is constant throughout $\mathcal{S}$. In the later case, we can write $[\mathbf{s}] = \text{Uniform}(\mathcal{S})$, which is to say that the N activity centers are uniformly distributed in the polygon $\mathcal{S}$. A point process is also a model for the number of individuals in the population N. So we could write $[\mathbf{s}|\mu]$ where μ is an intensity parameter defined as the number of points per unit area. In other words, μ is population density, and we often model population size as either $N \sim \text{Poisson}(\mu A(\mathcal{S}))$, where $A(\mathcal{S})$ is the area of the state-space; or, $N \sim \text{Binomial}(M, \psi)$ where $\psi = \mu A(\mathcal{S})/M$

and M is some large integer used simply as a convenience measure when conducting Bayesian analysis. As it turns out, there is very little practical difference in the Poisson prior versus a binomial models for N (Chapt. 11).

The model $[\mathbf{u}|\mathbf{s}]$ describes the locations of animals conditional on their activity center. In the original formulation of SCR models (Efford, 2004), this model component was intentionally ignored. Indeed when movement is not of direct interest, or when $\mathbf{s}$ is defined in a way not related to a home range center, it may be preferable to ignore this model component (Borchers, 2012). In other cases, we might use an explicit model, such as the bivariate normal model (Royle and Young, 2008).

The third component of the model, $[y|\mathbf{u}]$, describes how the observed data—the so-called capture-histories—arise conditional on the locations of animals. However, as mentioned previously, most SCR models do not contain a movement model, and thus, we typically entertain the model $[y|\mathbf{s}]$ instead of $[y|\mathbf{u}]$. This encounter model generally has at least two parameters, say p_0 and σ, describing the probability of capturing or detecting an individual given the distance between $\mathbf{s}$ and the trap. The most basic model is often called the half-normal model, although we typically refer to it as the Gaussian model since, in two-dimensional space, it is the kernel of a bivariate normal distribution. The model is $p_{ij} = p_0 \exp(-\|\mathbf{x}_j - \mathbf{s}_i\|/(2\sigma^2)$ where p_0 is the capture probability when the activity center occurs at the trap location $\mathbf{x}_j$, and σ is a spatial scale parameter determining how rapidly capture probability declines with distance. One common design leads to the model $[y_{ij}|\mathbf{s_i}] = \text{Bernoulli}(p_{ij})$. Chapt. 5 and Chapt. 9 describe many other possible encounter models.

When individuals are marked by biologists or have natural markings permitting individual recognition, y_{ij} is the observed data. However, some or all of the individuals cannot be uniquely identified, then we cannot record this individual-specific encounter history data. Instead, the data might be simply the number of detections at a trap or perhaps binary detection/non-detection data at each trap on each survey occasion. We call this reduced information data $n(y)$, and Chapt. 18 and Chapt. 19 describe models for $[n(y)|y]$ that still allow for density estimation. The basic strategy is to view y as "missing data" and to use the spatial correlation in the counts, or other sources of information, to provide information about these latent encounter histories.

Eq. 2.6.1 is a compact description of the the basic components of a SCR model, but it is also rather vague. The previous four paragraphs added enough extra detail so that we can now describe a specific SCR model. Perhaps the simplest SCR model is this:

$$\begin{aligned} N &\sim \text{Poisson}(\mu A(\mathcal{S})) \\ \mathbf{s}_i &\sim \text{Uniform}(\mathcal{S}) \\ y_{ijk}|\mathbf{s}_i &\sim \text{Bernoulli}(p(\|\mathbf{x}_j - \mathbf{s}_i\|) \end{aligned} \tag{2.6.2}$$

These "assumptions" are statistical statements of three basic hypotheses that (1)

population size N is Poisson distributed (2) activity centers are uniformly distributed in two-dimensional space, and (3) capture probability is a function of the distance between the activity and the trap. Each of these model components can be modified as needed to match specific hypotheses, study designs, and data structures. For example, spatial variation in abundance or density can be easily modeled as a function of habitat covariates (Chapt. 11).

We realize that many the model description in Eq. 2.6.2 may not be self-evident to some ecologists. However, it is absolutely essential that one can understand such a model description—not just for being able to read this book, but also for understanding any statistical model in ecology. One of the best ways of familiarizing oneself with this notation is to translate it into **R** code that simulates outcomes from the model. The following code is an example.

```
set.seed(36372)
Area <- 1                                  # area of state-space (unit square)
x <- cbind(rep(seq(.1,.9,.2), each=5),     # trap locations
           rep(seq(.1,.9,.2), times=5))
p0 <- 0.3                                  # baseline capture probability
sigma <- 0.05                              # Gaussian scale parameter
mu <- 50                                   # population density
N <- rpois(1, mu*Area)                     # population size
s <- cbind(runif(N, 0, 1),                 # activity centers in unit square
           runif(N, 0, 1))
K <- 5
y <- matrix(NA, N, nrow(x))                # capture data
for(i in 1:N) {
  d.ij <- sqrt((x[,1] - s[i,1])^2 +        # distance between x and s[i]
               (x[,2] - s[i,2])^2)
  p.ij <- p0*exp(-d.ij^2 / (2*sigma^2)) # capture probability
  y[i,] <- rbinom(nrow(x), K, p.ij)        # capture history for animal i
}
```

Fig. 2.3 shows the results of this simulation from a basic, yet very useful, SCR model.

Having briefly explained each of the model components in Eq. 2.6.1, and having shown how a subset of these components results in a basic SCR model, we can now discuss other relevant arrangements. Examples include: (1) Classical distance sampling (Buckland et al., 2001; Borchers et al., 2002), (2) Spatial capture-recapture models with fixed arrays of traps (Efford, 2004; Borchers and Efford, 2008; Royle et al., 2009a,b; Gardner et al., 2010a; Royle et al., 2011b), and (3) Search-encounter models (Royle and Young, 2008; Royle et al., 2011a). We will now elaborate on some of these distinctions.

1. **Distance sampling**. The last 2 stages of the hierarchy are confounded (implicitly) and so analysis is based on the model $[y|\mathbf{u}][\mathbf{u}]$. The "process model" is that of "uniformity": $\mathbf{u} \sim \text{Uniform}(\mathcal{S})$.

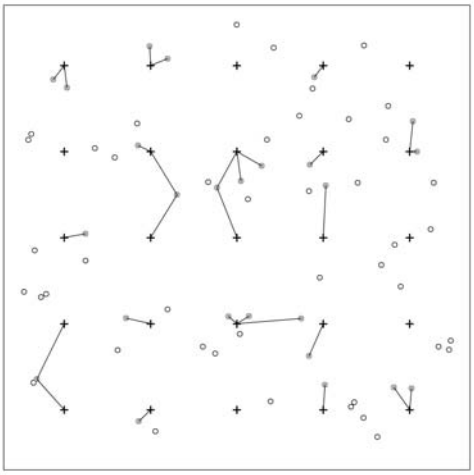

Figure 2.3. Population of $N = 69$ home-range centers ($\mathbf{s}$, circles) and 25 trap locations ($\mathbf{x}$, crosses). Lines connect activity centers to the traps where the individuals were detected. As in many SCR models, movement outcomes ($\mathbf{u}$) are ignored.

2. **Spatial capture-recapture model with a fixed array of traps**. SCR models appear to have little in common with distance sampling because observations are made only at a pre-defined set of discrete locations—where traps are placed. However, the models are closely related in terms of our hierarchical representation above In SCR models based on fixed arrays, we cannot estimate both $\Pr(y = 1|\mathbf{u})$ and $\Pr(\mathbf{u}|\mathbf{s})$—the probability that an individual "moves to $\mathbf{u}$" cannot be separated from the probability that it is detected given that it moves to $\mathbf{u}$, because of the fact that the observation locations are fixed by design. Formally, such SCR models confound $[y|\mathbf{u}]$ with $[\mathbf{u}|\mathbf{s}]$ so that the observation model arises as:

$$[y|\mathbf{s}] = \int_{\mathbf{u}} [y|\mathbf{u}][\mathbf{u}|\mathbf{s}]\mathrm{d}\mathbf{u}$$

   This confounding happens because SCR sampling is spatially biased—restricted to a fixed pre-determined set of locations. Conversely, distance sampling confounds $[\mathbf{u}|\mathbf{s}][\mathbf{s}]$ because, essentially, there is only a single realization of the encounter process. It is probably reasonable to assume that $\Pr(y = 1|\mathbf{u}) = 1$ or at least it is locally constant for most devices (e.g., cameras, etc..), and thus the detection model will have the interpretation in terms of movement (see Chapt. 13 and 12).
3. **Search-encounter models**. What we call "search-encounter" models (Royle

and Young, 2008; Royle et al., 2011a) are kind of a hybrid model combining features of SCR models and features of distance sampling. Like distance sampling they allow for encounters in continuous space which provide direct observations from [**u**|**s**]. Thus, the hierarchical model is fully identified. These models are described in Chapt. chapt.search-encounter.

2.7 SUMMARY AND OUTLOOK

Spatial capture-recapture models are hierarchical models, and hierarchical models are models of multiple random variables that are conditionally related. It is therefore important that the basic rules of modeling random variables are understood, and we hope that this chapter has made some of the basic concepts accessible to ecologists with rudimentary background in statistics. If some of this material still seems difficult to grasp, we recommend working with the provided **R** code, which is perhaps the best way of making the equations more tangible.

In some respects, it is possible to understand the jist of SCR without knowing anything about marginal and conditional relationships. One can always fit models using canned software and interpret the output without understanding the guts of the model or the details of the estimation process. For some applied ecologists, this may be perfectly fine, and this book is meant to be useful for both statistical novices and ecologists with more advanced quantitative skills. In most chapters, we begin with a basic conceptual discussion, then we explain the technical details that require an understanding of the concepts in this chapter, and finally we end with one or more worked examples. For those not interested in the technical details, we recommend focusing on the chapter introductions and the examples. However, taking the time to understand the concepts presented in this chapter can only increase one's ability to tackle the unique and complex problems that often present themselves when modeling spatial and temporal aspects of population dynamics.