CHAPTER

17

Binomial "t-Test"

OUTLINE

17.1 INTRODUCTION

The next three chapters deal with another common kind of count data, where we want to estimate a binomial proportion. The associated models are often called logistic regressions. The crucial difference between binomial and Poisson random variables is the presence of a ceiling in the former: binomial counts cannot be greater than some upper limit. Therefore, we model bounded counts, where the bound is provided by trial size N. It makes sense to express a count relative to that bound, and this yields a proportion. In contrast, Poisson counts lack an upper limit, at least in principle.

Modeling a binomial random variable in essence means modeling a series of coin flips. We count the number of heads among the total number of coin flips (N) and from this want to estimate the general tendency of the coin to show heads. That is, we want to estimate Pr(heads). Data coming from coin flip-like processes are ubiquitous in nature and include survival or the occurrence of an organism. In coin flips, the binomial distribution describes the number of times r a coin shows heads among a number of N flips, where the coin has Pr(heads) = p. We also write $r \sim$ Binomial (N, p). A special case of the binomial distribution with $N = 1$, corresponding to

Introduction to WinBUGS for Ecologists
DOI: 10.1016/B978-0-12-378605-0.00017-X

FIGURE 17.1 Cross-leaved gentian (*Gentiana cruciata*), Spanish Pyrenees, 2006. (*Photo M. Kéry*)

a single flip of the coin, is called the Bernoulli distribution. It has just a single parameter p.

As our inferential setting of this chapter, we consider a plant inventory on calcareous grasslands in the Jura mountains. A total of 50 sites were visited by experienced botanists who recorded whether they saw a species or not. The Cross-leaved gentian (Fig. 17.1) was found at 13 sites and the Chiltern gentian (see Chapter 20) at 29 sites. We wonder whether this is enough evidence, given the variation inherent in binomial sampling, to claim that the Cross-leaved gentian has a more restricted distribution in the Jura mountains.

This type of data is often called "presence–absence data." It is more accurate to call it "detection–nondetection data," since the number of sites at which a species is detected depends on two entirely different things: first, the number of sites where a species is actually present and second, the ease with which a species is detected at an occupied site. Without a special kind of data (see Chapter 20), we have no way of distinguishing between the two components of "presence–absence." All we can do is hope, pray, or claim either that both gentian species are found at every occupied site or else that their probability to be overlooked is identical. For now, we assume that every individual present is detected, i.e., that every occupied site is observed as such. Given such data, our question can be framed

statistically by what can be called a binomial version of the t-test, i.e. a logistic regression that contrasts two groups. For gentian species i, let C_i be the number of sites it was detected. A simple model for C_i is this:

1. (Statistical) Distribution: $C_i \sim \text{Binomial}(N, p_i)$
2. Link function: logit, i.e., $\text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \text{linear predictor}$
3. Linear predictor: $\alpha + \beta * x_i$

If x is an indicator for the Chiltern gentian, then α can be interpreted as a logit-scale parameter for the probability of occurrence of the Cross-leaved gentian in the Jura mountains and β is the difference, on a logit-scale, between the probability of occurrence of the Chiltern gentian and that of the Cross-leaved gentian.

17.2 DATA GENERATION

We simulate the data from a binomial process whose parameters were defined such that the sample data approximately match those in our example. Note that our modeled response simply consists of two numbers.

```
N <- 50                          # Binomial total (Number of coin flips)
p.cr <- 13/50                    # Success probability Cross-leaved
p.ch <- 29/50                    # Success probability Chiltern gentian

C.cr <- rbinom(1, 50, prob = p.cr)    ;  C.cr          # Add Binomial noise
C.ch <- rbinom(1, 50, prob = p.ch)    ;  C.ch          # Add Binomial noise
C <- c(C.cr, C.ch)
species <- factor(c(0,1), labels = c("Cross-leaved gentian","Chiltern gentian"))
```

17.3 ANALYSIS USING R

A binomial t-test in R suggests a significant difference in the perceived distribution of the Cross-leaved gentian and the Chiltern gentian.

```
summary(glm(cbind(C, N - C) ~ species, family = binomial))
predict(glm(cbind(C, N - C) ~ species, family = binomial), type = "response")
```

```
> summary(glm(cbind(C, N - C) ~ species, family = binomial))

Call:
glm(formula = cbind(C, N - C) ~ species, family = binomial)
[ ... ]
```

```
Coefficients:
                              Estimate   Std. Error    z value     Pr(>|z|)
(Intercept)                    -1.0460       0.3224     -3.244      0.00118  **
speciesChiltern gentian         1.3687       0.4313      3.173      0.00151  **
- - -
[ ... ]

> predict(glm(cbind(C, N - C) ~ species, family = binomial), type = "response")
   1    2
0.26 0.58
```

17.4 ANALYSIS USING WinBUGS

```
# Define model
sink("Binomial.t.test.txt")
cat("
model {

# Priors
 alpha ~ dnorm(0,0.01)
 beta ~ dnorm(0,0.01)

# Likelihood
 for (i in 1:n) {
     C[i] ~ dbin(p[i], N)        # Note p before N
     logit(p[i]) <- alpha + beta *species[i]
 }
# Derived quantities
 Occ.cross <- exp(alpha) / (1 + exp(alpha))
 Occ.chiltern <- exp(alpha + beta) / (1 + exp(alpha + beta))
 Occ.Diff <- Occ.chiltern - Occ.cross          # Test quantity
}
",fill=TRUE)
sink()

# Bundle data
win.data <- list(C = C, N = 50, species = c(0,1), n = length(C))

# Inits function
inits <- function(){ list(alpha=rlnorm(1), beta=rlnorm(1))}

# Parameters to estimate
params <- c("alpha", "beta", "Occ.cross", "Occ.chiltern", "Occ.Diff")
```

```
# MCMC settings
nc <- 3
ni <- 1200
nb <- 200
nt <- 2

# Start Gibbs sampling
out <- bugs(data=win.data, inits=inits, parameters.to.save=params,
model.file="Binomial.t.test.txt", n.thin=nt, n.chains=nc, n.burnin=nb, n.iter=ni,
debug = TRUE)

print(out, dig = 3)
> print(out, dig = 3)
Inference for Bugs model at "Binomial.t.test.txt", fit using WinBUGS,
 3 chains, each with 1200 iterations (first 200 discarded), n.thin = 2
 n.sims = 1500 iterations saved
                mean     sd    2.5%     25%     50%     75%   97.5%  Rhat  n.eff
alpha         -1.047  0.333  -1.710  -1.273  -1.034  -0.810  -0.441 1.007    290
beta           1.362  0.449   0.520   1.059   1.375   1.657   2.278 1.007    300
Occ.cross      0.265  0.063   0.153   0.219   0.262   0.308   0.392 1.007    290
Occ.chiltern   0.577  0.069   0.438   0.530   0.580   0.626   0.701 1.003    960
Occ.Diff       0.312  0.095   0.124   0.251   0.318   0.376   0.485 1.006    340
[ ... ]
DIC info (using the rule, pD = Dbar-Dhat)
pD = 2.1 and DIC = 12.6
```

Next, we plot the posterior distribution of the (biological) distributions of the two species and their difference, as perceived from the detection–nondetection data (Fig. 17.2):

```
par(mfrow = c(3,1))
hist(out$sims.list$Occ.cross, col = "grey", xlim = c(0,1), main = "", xlab =
"Occupancy Cross-leaved", breaks = 30)
abline(v = out$mean$Occ.cross, lwd = 3, col = "red")
hist(out$sims.list$Occ.chiltern, col = "grey", xlim = c(0,1), main = "", xlab =
"Occupancy Chiltern", breaks = 30)
abline(v = out$mean$Occ.chiltern, lwd = 3, col = "red")
hist(out$sims.list$Occ.Diff, col = "grey", xlim = c(0,1), main = "", xlab =
"Difference in occupancy", breaks = 30)
abline(v = 0, lwd = 3, col = "red")
```

The posterior distribution of the difference barely overlaps zero (Fig. 17.2, bottom), and the 95% credible interval is (0.015, 0.485). Therefore, we can say that the Chiltern gentian is found at significantly more sites than the Cross-leaved gentian. This is not equivalent to saying that the Chiltern gentian was more widespread in the Jura mountains than is

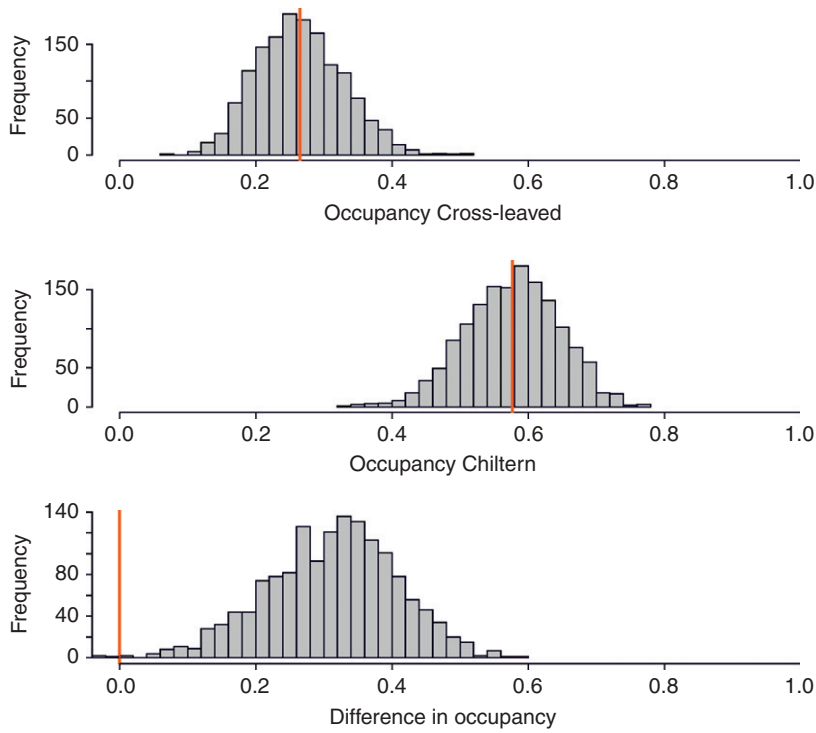

FIGURE 17.2 Posterior distributions of the occupancy probability of Cross-leaved (top panel) and Chiltern gentian (middle panel); vertical red lines indicate posterior means. Bottom panel shows the posterior distribution of the difference of the two species' occupancy probability; vertical red line indicates zero.

the Cross-leaved gentian, only that it is *detected* at more sites. Whether it also means the former depends on whether the assumption of perfect or at least of constant detection probability holds (MacKenzie and Kendall, 2002; Kéry and Schmidt, 2008).

17.5 SUMMARY

As for all generalized linear models, a binomial t-test in WinBUGS is a fairly trivial generalization of the corresponding normal response model. We have seen another example of the ease with which, in Markov chain Monte Carlo-based statistical inference, we can compute derived parameters exactly, i.e., without any approximations, and with full error propagation.

EXERCISES

1. *Binomial and Bernoulli*: Try to formulate the same problem using a Bernoulli distribution. You will need to simulate, not the aggregate number of sites at which each gentian species was found, but rather the detected occupancy status of each individual site. Adapt the code for analysis. In WinBUGS, you can directly specify a Bernoulli by `dbern(p)`. (As usual, when you're unsure, go to the WinBUGS manual: `Help > User Manual` and then scroll down.)
2. *Swiss hare data*: Model the probability of occurrence of a "large" count in arable and grassland areas. Select a useful threshold to call a count "large." Are large counts more common in arable areas?