Occupancy models

Zero-inflated models for hidden processes

Flow of ideas

- Mixture models in general
- Zero-inflation as a useful example of mixture models
- Occupancy as an example of zero-inflation

Mixture models

Given a finite set of probability distributions for the random variable z,

$[z]_i = [z]_1, \ldots, [z]_n$

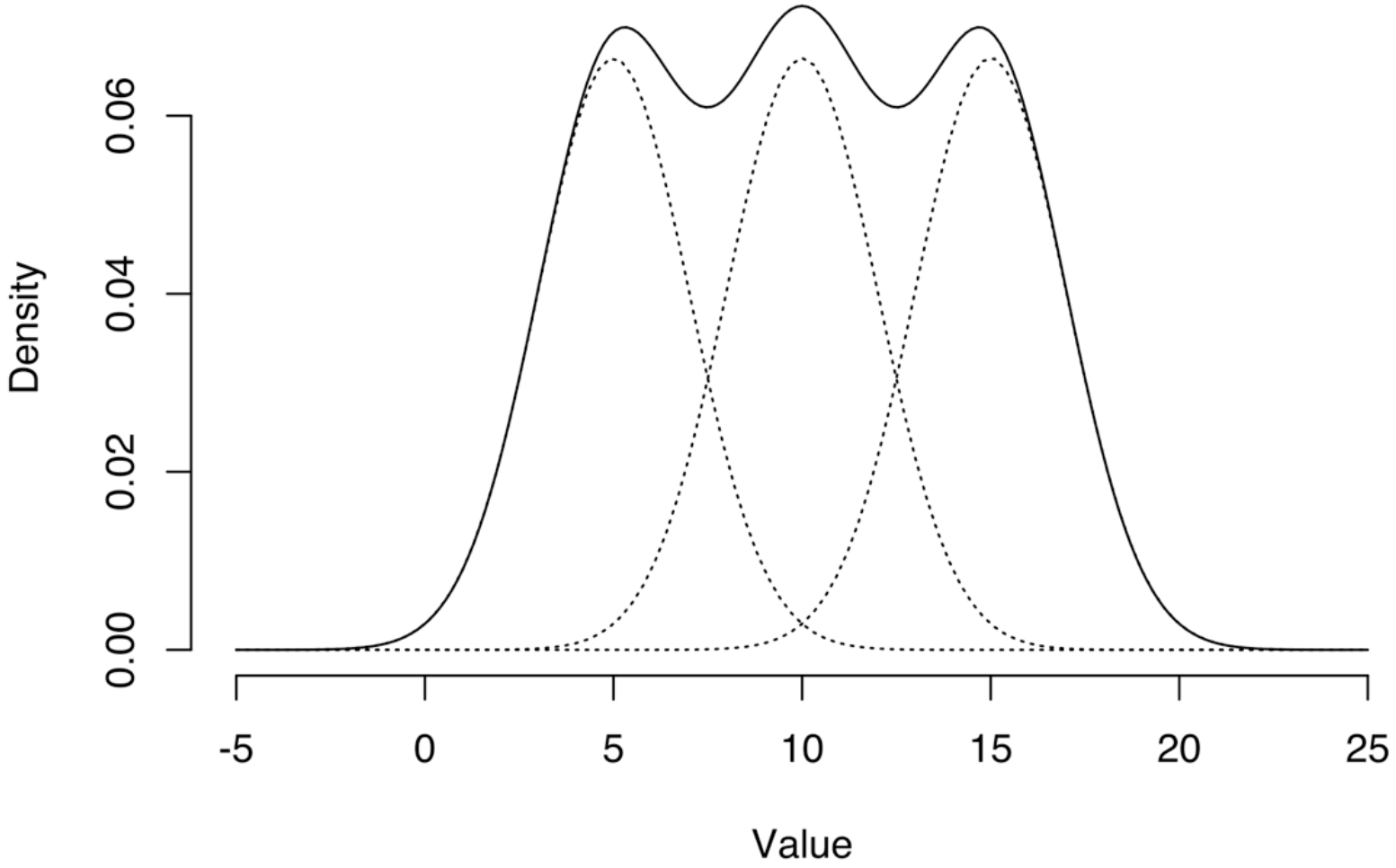
and weights

$$w_1, \ldots, w_n, w_i \geq 0, \sum_{i=1}^{n} w_i = 1,$$

the finite mixture distribution of z is

$$[z] = \sum_i w_i [z]_i .$$

$$z \sim .333\,\text{normal}(z \mid \mu_1, \sigma^2) + .333\,\text{normal}(z \mid \mu_2, \sigma^2) + .333\,\text{normal}(z \mid \mu_3, \sigma^2)$$



Mixture model example

Suppose your study a sexually dimorphic species and you want to represent the distribution of the random variable, body mass of an individual, z . Because body mass is strictly positive, a gamma distribution is a logical choice, but a single gamma probability density function is not up to the task of representing the two sources of variation in body mass arising from males and females. Instead, we might use:

$$z \sim \phi \cdot \text{gamma}\left(z \mid \alpha_m, \beta_m\right) + \left(1-\phi\right) \cdot \text{gamma}\left(z \mid \alpha_f, \beta_f\right) \cdot \text{beta}\left(\phi \mid \eta, \rho\right),$$

α_m, β_m parameters for male body mass

α_f, β_f parameters for female body mass

ϕ proportion of population that is male

η, ρ parameters of distribution of proportion male

Zero-inflation as mixture model

- Often encounter data where 0's are more frequent than would be predicted using a single probability mass function (Poisson, binomial, multinomial, etc.)
- This is because 0's arise from more than one process.
- We use mixture models to represent the operation of the two processes.

Zero inflation as a mixture model

Imagine that you sampled many plots along a coastline, counting the number of species of mussels within each plot. In essence there are two sources of zeros. Some zeros arise because the plot was placed areas that are not mussel habitat, while other zeros occur in plots placed in mussel habitat but that contain no mussels as a result of sampling variation. The Poisson distribution offers a logical choice for modeling the distribution of counts in mussel habitat, but it cannot portray the zeros that arise because plots were placed in areas where mussels never live.

Let z be a random variable, the number of mussel species in a m^2 plot, and w be a random variable describing mussel habitat, $w=1$ if a plot is located outside of mussel habitat and $w=0$ if it is located outside habitat. The probability distribution of number of mussel species in a plot is given by

$$z \sim \begin{cases} 0 & w = 0 \\ \text{Poisson}(\lambda) & w = 1 \end{cases},$$

$$z \sim \phi \text{Poisson}(z \mid 0) + (1-\phi)\text{Poisson}(z \mid \lambda)$$

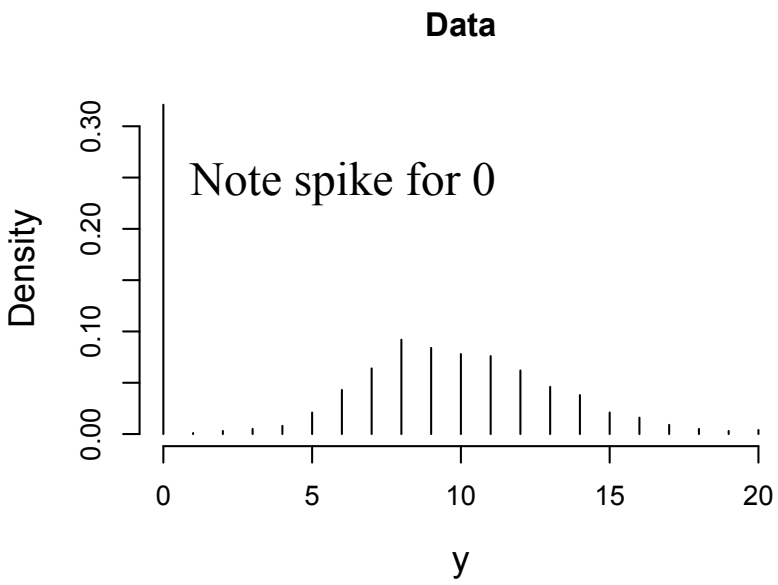
$$z \sim \text{Poisson}\left(z \mid \lambda(1-w)\right) \cdot \text{Bern}\left(w \mid \phi\right) \cdot \text{beta}\left(\phi \mid \alpha, \beta\right)$$

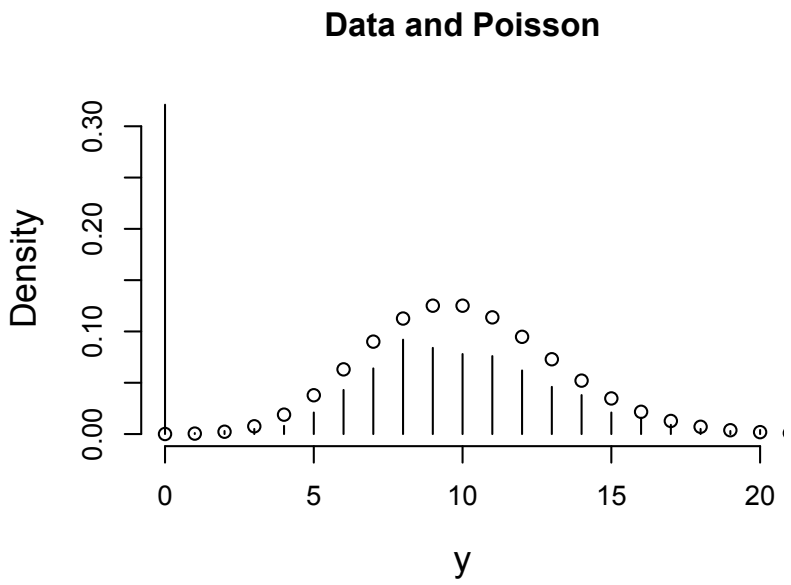
where λ is the averge number of mussels per area of habitat

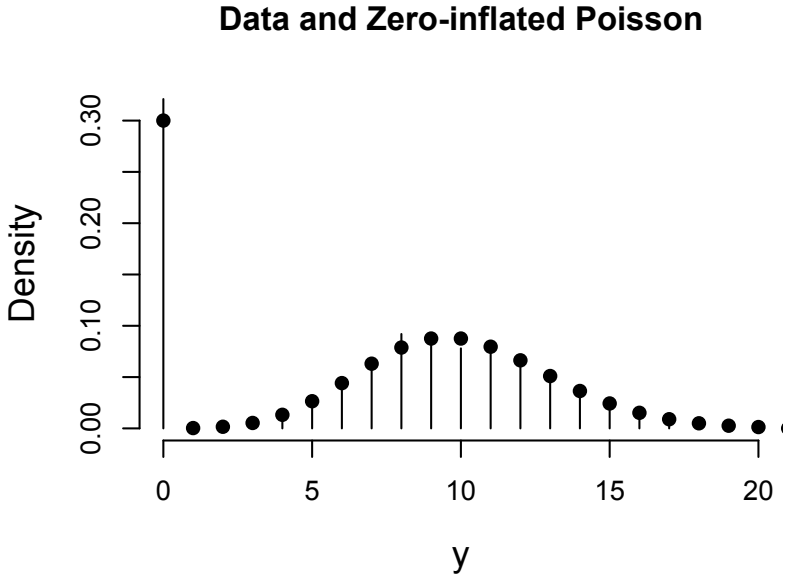
ϕ is the proportion of the total area that is not mussel habitat

α, β are shape parameters contolling the distrirbution of ϕ.

Draw the Bayesian network for this example.







Zero-inflation example

Zero inflated models

Poisson

$$[\lambda,\alpha,\beta,w,\phi \mid y] \propto \text{Poisson}\big(y \mid \lambda(1-w)\big)\cdot \text{Bernoulli}\big(w \mid \phi\big)\cdot \text{beta}\big(\phi \mid \alpha,\beta\big)[\lambda][\alpha][\beta]$$

binomial

$$[p,\alpha,\beta,w,\phi \mid y] \propto \text{binomial}(y \mid p(1-w),n)\cdot \text{Bernoulli}(w \mid \phi)\cdot \text{beta}(\phi \mid \alpha,\beta)[p][\alpha][\beta]$$

In both cases ϕ is the probability of a 0 that is not accounted for by sampling variation alone.

Can also use negative binomial, multinomial.

The problem of occupancy, an example of zero inflation

- We want to understand what controls *presence or absence* of individuals or traits within discrete categories, e.g.,
  - invasive species within different plant communities
  - seedlings within distance bins from forest edge
  - nitrogen fixers within soil types
- We have a problem with interpreting absence:
  - The individual is truly absent.
  - The individual is present but we fail to observe it (i.e, a false negative)
- We often want to explain spatial and temporal variation in the observations using a model.
- We need to estimate uncertainty due to
  - The failure of our model to truly portray spatial and temporal variation in occupancy.
  - The uncertainty in our observations arising because we fail to observe occupancy without error.

Exercise: Courtesy of McCarthy 2007: Box 5.9

- Kristen Parris studied controls on the distribution of tree frogs in the riparian zone of streams on the east coast of sub-tropical Australia.
- Multiple surveys were conducted at 64 sites using 2 observation methods, visual searches at night and auditory searches for responses to taped calls. We assume the presence /absence of frogs at a site does not change during the multiple surveys. (Frogs don't fly.)
- Presence / absence of frogs was modeled as a function of 1) stream size (measured as annual volume of rainfall in the catchment above the site) 2) the presence or absence of palms at the site (an indicator of mesic or xeric conditions) and 3) the interaction between palms and stream size.
- Data are the total number of times frogs were detected using each method for each site and the number of surveys for each site.
- Develop a Bayesian hierarchical model representing the effect of covariates on habitat occupancy. Account for uncertainty in detection in both methods of observation. Sketch a network diagram and write out the posterior distribution and joint distribution.

Tree Frog Example

$\mathbf{x}1$ = vector of observations of stream size

$\mathbf{x}2$ = vector of observations of palms (0=absent,1=present)

$\mathbf{y}^{visual}$ = vector of number of observations at each site by eye

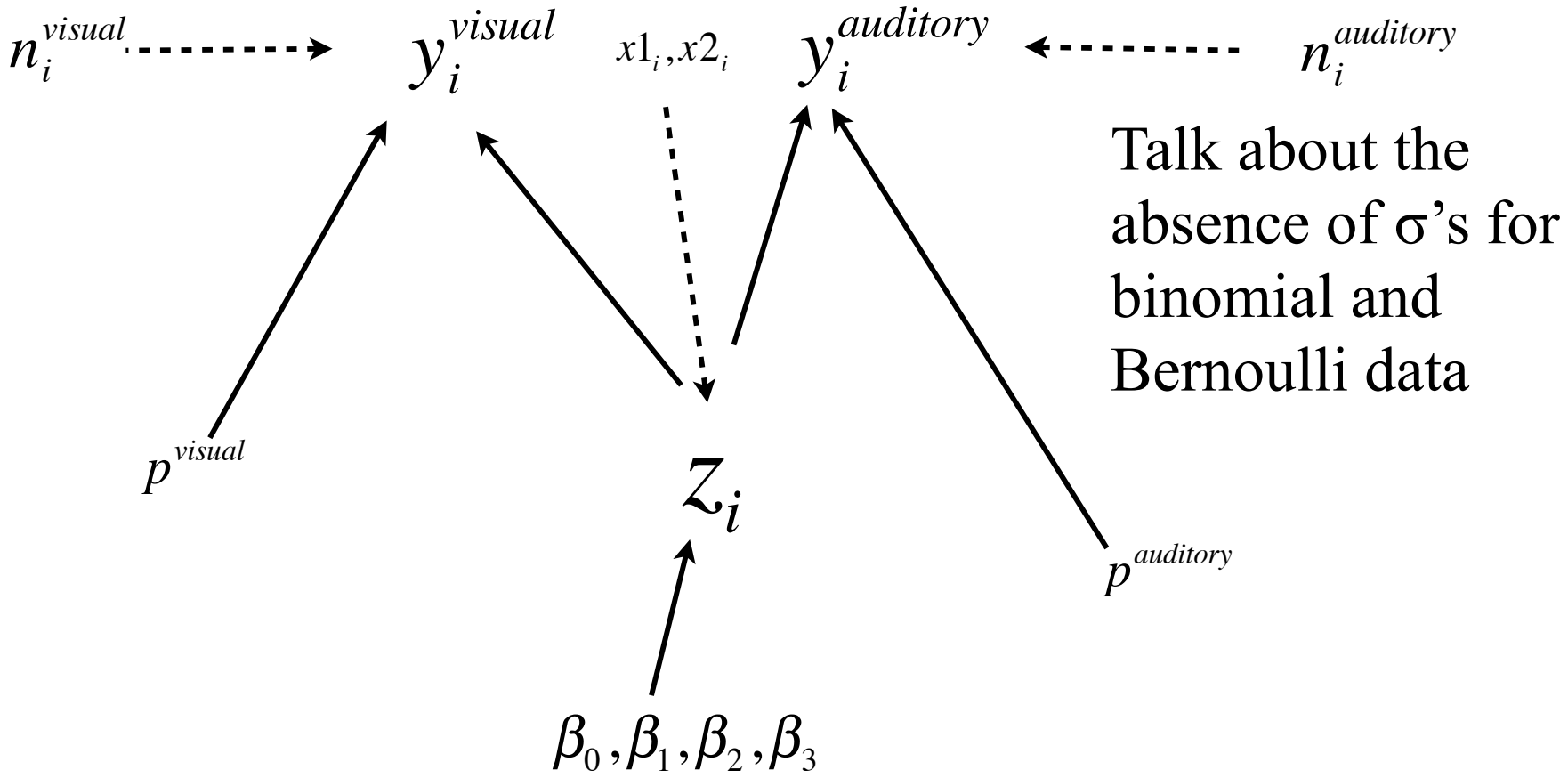
$\mathbf{y}^{auditory}$ = vector of number of observations at each site by ear

$\mathbf{n}^{visual}$ = vector of number of visual surveys

$\mathbf{n}^{auditory}$ = vector of number of auditory surveys

z_i=the true state of patch i. z_i = 1 if occupied and 0 if not occupied

Draw a Bayesian network and write out the posterior and joint distributions.


n_i^{visual}
y_i^{visual}
$x1_i, x2_i$
$y_i^{auditory}$
$n_i^{auditory}$
Talk about the absence of σ's for binomial and Bernoulli data
p^{visual}
z_i
$p^{auditory}$
$\beta_0, \beta_1, \beta_2, \beta_3$

For binomial and Bernoulli:

$$\mu = np \quad \sigma^2 = np(1-p)$$

where σ^2 is the observation variance

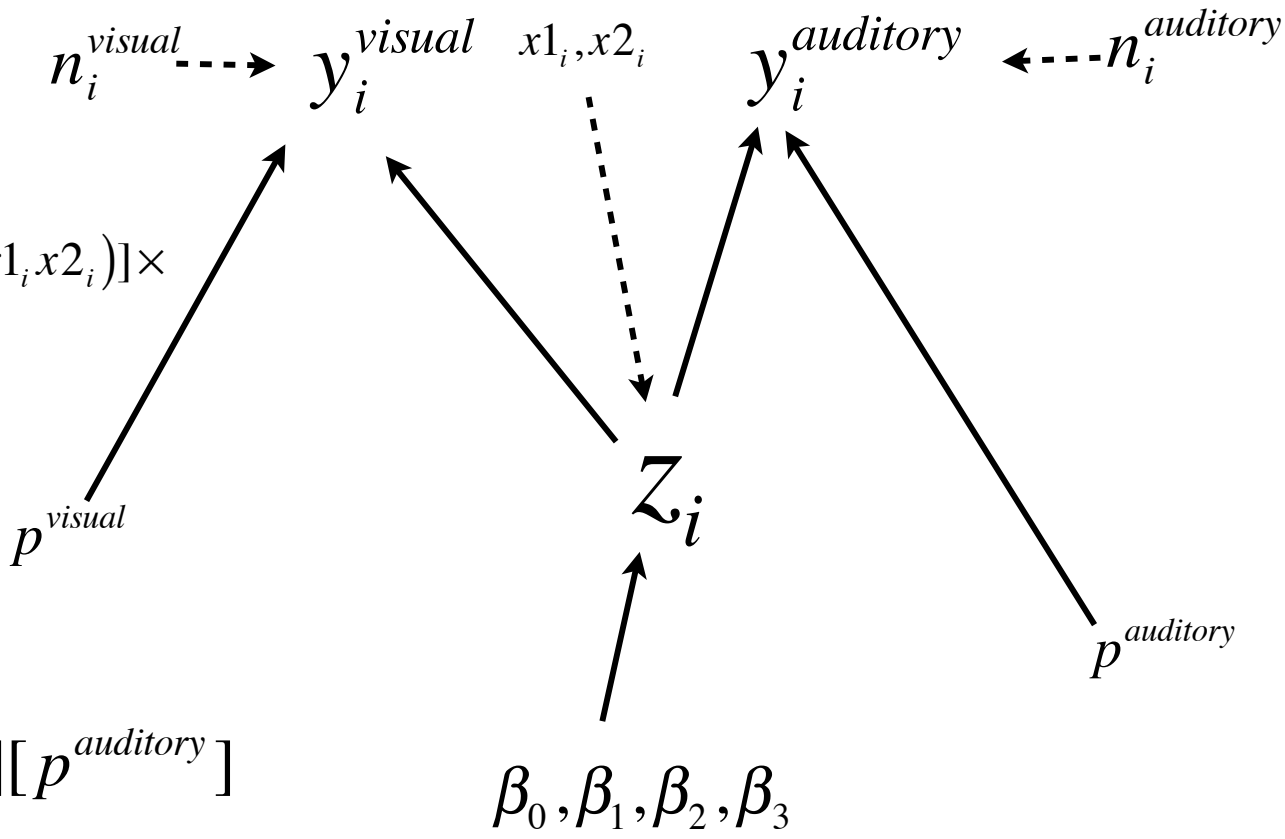
and μ is the mean number of successes on n trials

$$\left[z_i, \boldsymbol{\beta}, p^{auditory}, p^{visual} \mid y_i{}^{auditory}, y_i{}^{visual}\right] \propto$$

$$[y_i^{auditory} \mid z_i \cdot p^{auditory}, n_i^{auditory}] \times$$

$$[y_i^{visual} \mid z_i \cdot p^{visual}, n_i^{visual}] \times$$

$$[z_i \mid \text{invlogit}(\beta_0 + \beta_1 x1_i + \beta_2 x2_i + \beta_3 x1_i x2_i)] \times$$



$$[\beta_0][\beta_1][\beta_2][\beta_3][p^{visual}][p^{auditory}]$$

$$\left[\mathbf{z}, \boldsymbol{\beta}, p^{auditory}, p^{visual} \mid y^{auditory}, y^{visual}\right] \propto$$

$$\prod_{i=1}^{64} \text{binomal}\left(y_i^{auditory} \mid z_i \cdot p^{auditory}, n_i^{auditory}\right) \text{binomial}\left(y_i^{visual} \mid z_i \cdot p^{visual}, n_i^{visual}\right) \times$$

$$\text{Bernoulli}\left(z_i \mid \text{invlogit}\left(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2\right)\right) \times$$

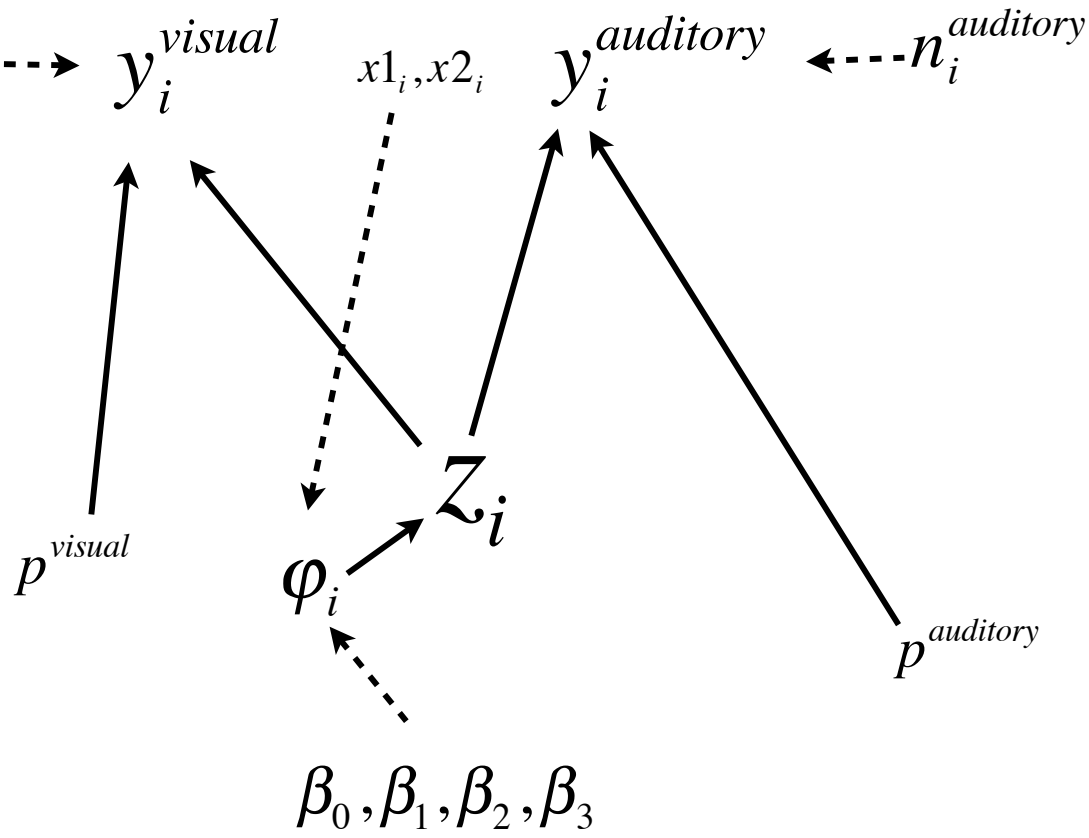
$$\prod_{j=0}^{3} \text{normal}\left(\beta_j \mid 0, .0001\right) \text{uniform}(p^{visual} \mid 0,1) \text{uniform}(p^{auditiory} \mid 0,1)$$

Talk about multiple sources of data

```
model
{
  b0 ~ dnorm(0, 1.0E-6)  # uninformative priors for the variables
  b[1] ~ dnorm(0, 1.0E-6)
  b[2] ~ dnorm(0, 1.0E-6)
  b[3] ~ dnorm(0, 1.0E-6)
  p.visual ~ dunif(0, 1)  # detection probabilities when the species is present
  p.auditory ~ dunif(0, 1)

  mLnCV <- mean(LnCV[])  # average catchment volume

  for (i in 1:64)  # for each of the 64 sites
  {
    phi[i] <- ilogit(b0 + b[1]*(LnCV[i] - mLnCV) + b[2]*palms[i] + b[3]*(LnCV[i] -
mLnCV))*palms[i]  # probability of presence using centered data
    z[i] ~ dbern(phi[i])  # actual, latent presence at the site, 0 or 1
    y.visual[i] ~ dbin(p.visual*z[i], n.visual[i])  # number eye detections with
    y.auditory[i] ~ dbin(p.auditory*z[i], n.auditory[i])  # number of ear detections
  }

  # predicted relationships--derived quantities; note that predictions are done using
centered data. There is no back-transform.
  for (i in 1:20)
  {
    LVol[i] <- 2 + 3*i/20  # covers the range of stream sizes
    logit(predpalms[i]) <- b0 + (b[1] + b[3])*(LVol[i] - mLnCV) + b[2]
    logit(prednopalms[i]) <- b0 + b[1]*(LVol[i] - mLnCV)
  }

}
```

Important

Here is where you can run into trouble. (I did.) Notice that ϕ_i is simply a calculation, a function of the β's and the x's. It is not a stochastic quantity. The relationship between z_i and the β's and the x's form the stochastic quantity. You are certainly free to estimate the ϕ_i as a function of random variables, but putting them in the diagram doesn't make sense in terms of the heads and tails of arrows helping us see the conditioning. If you want them in the diagram, use a different type of arrow indicating a calculated quantity, as I did here.



- You may need to bound phi[1], i.e.:

```
phi[i] <- max(min(ilogit(a + b[1]*(LnCV[i] - mLnCV) +
b[2]*palms[i] + b[3]*(LnCV[i] - mLnCV))*palms[i],.0000001),.
999999)
```

- A better way to specify priors on the intercept (in this case *a*). (Quite a cool trick, actually)

```
phi0 ~ dunif(0,1)
b0 <- logit(phi0)
```

Explain what is going on here and why this coding is superior to

```
b0 ~ dnorm(0,.000001)
```