Modeling zero-truncated and zero-inflated data

Analysis of Ecological and Environmental Data

QERM 514

Mark Scheuerell

20 May 2020

Goals for today

- Understand the difference between zero-truncated and zero-inflated data
- Understand how to model zero-truncated data
- Understand the differences between zero-altered and zero-inflated models

Zero-truncated data

Zero-truncated data cannot take a value of 0

Although somewhat rare in ecological studies, examples include

- time a whale is at the surface before diving
- herd size in elk
- number of fin rays on a fish

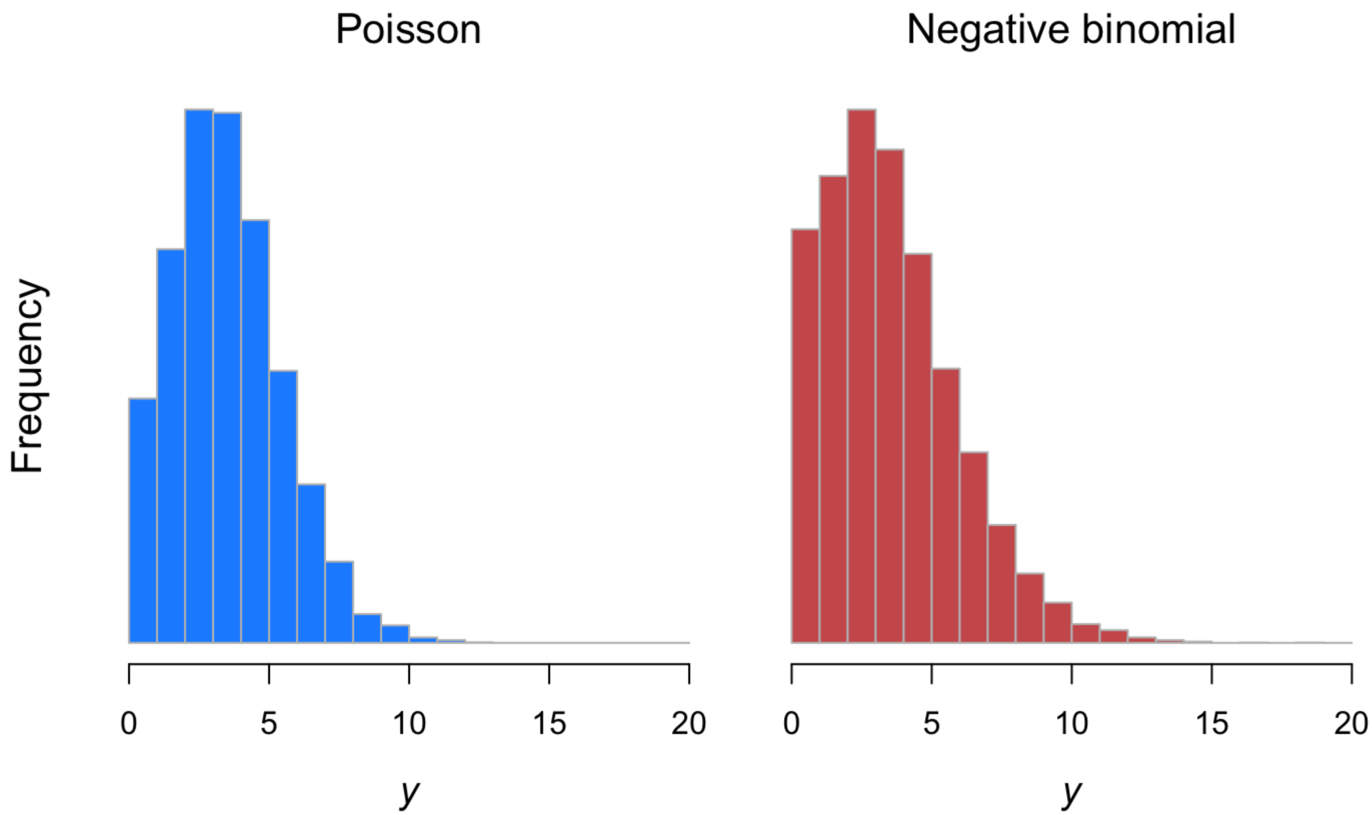
Zero-truncated data

Zero-truncated data are not necessarily a problem

Rather, an underlying assumption of Poisson or neg binomial may be the problem

Zero-truncated data

Both of these examples contain zeros



Poisson distribution

Recall that for $y_i \sim \text{Poisson}(\lambda)$

its probability mass function is

$$f(y_i; \lambda_i) = \frac{\exp(\text{-}\lambda_i)\lambda_i^{y_i}}{y_i!}$$

$f(y_i; \lambda_i)$ gives the probability of $y_i \geq 0$

Poisson for zero-truncated data

The probability that $y_i = 0$ is

$$f(y_i; \lambda_i) = \frac{\exp(\text{-}\lambda_i)\lambda_i^{y_i}}{y_i!}$$

$$\Downarrow$$

$$\begin{aligned} f(y_i = 0; \lambda_i) &= \frac{\exp(\text{-}\lambda_i)\lambda_i^{0}}{0!} \\ &= \exp(\text{-}\lambda_i) \end{aligned}$$

Poisson for zero-truncated data

The probability that $y_i \neq 0$ is therefore

$$f(y_i = 0; \lambda_i) = \exp(\text{-}\lambda_i)$$
$$\Downarrow$$
$$f(y_i \neq 0; \lambda_i) = 1 - \exp(\text{-}\lambda_i)$$

Poisson for zero-truncated data

We can now exclude the probability that $y_i = 0$ by dividing the pmf by the probability that $y_i \neq 0$

$$f(y_i; \lambda_i) = \frac{\exp(\text{-}\lambda_i)\lambda_i^{y_i}}{y_i!}$$

$$\Downarrow$$

$$f(y_i; \lambda_i | y_i > 0) = \frac{\exp(\text{-}\lambda_i)\lambda_i^{y_i}}{y_i!} \cdot \frac{1}{1 - \exp(\text{-}\lambda_i)}$$

$$\Downarrow$$

$$\log \mathcal{L} = (y_i \log \lambda_i - \lambda_i) - (1 - \exp(\text{-}\lambda_i))$$

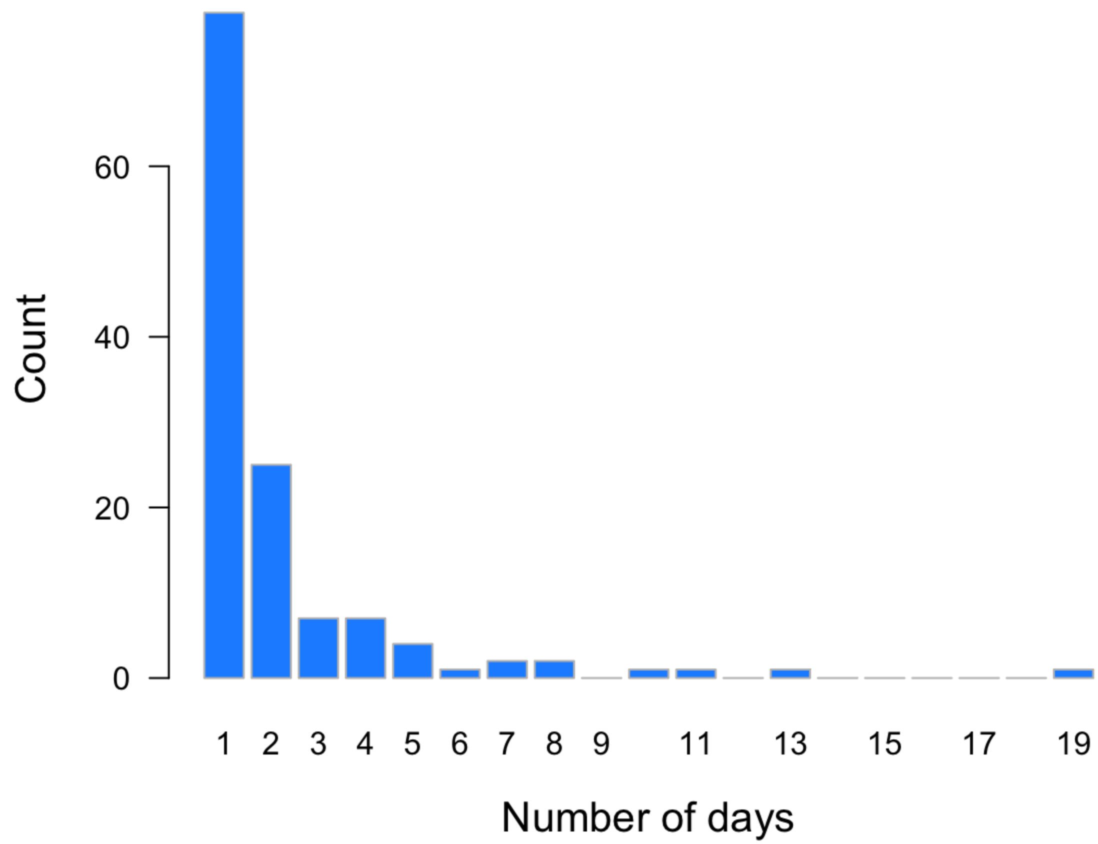
Zero-truncated data

Example

Let’s consider some data presented in Zuur et al. (2009), which detail the number of days that carcasses from road-killed snakes stay on roads

The predictors are the total rainfall (mm) and an indicator of where on the pavement the snake was killed (lane or shoulder)

Longevity of road-killed snakes



Longevity of road-killed snakes

Let's first consider a regular Poisson regression model

```r
## Poisson regression
smod_pois <- glm(n_days ~ location + rain, data = snakes,
                 family = poisson(link = "log"))
```

Longevity of road-killed snakes

```
##
## Call:
## glm(formula = n_days ~ location + rain, family = poisson(link = "log"),
##     data = snakes)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0560  -0.7981  -0.4738   0.3410   6.6474
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 0.439195   0.088539   4.960 7.03e-07 ***
## locationV   0.462663   0.119060   3.886 0.000102 ***
## rain        0.021707   0.003092   7.021 2.21e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 226.38  on 129  degrees of freedom
## Residual deviance: 175.80  on 127  degrees of freedom
## AIC: 497.63
##
## Number of Fisher Scoring iterations: 5
```

Longevity of road-killed snakes

Now let's fit a zero-truncated Poisson regression model with `vglm()` from **VGAM**

```
library(VGAM)
## zero-truncated Poisson regression
smod_ztpois <- vglm(n_days ~ location + rain, data = snakes,
                    family = pospoisson)
```

Longevity of road-killed snakes

```
##
## Call:
## vglm(formula = n_days ~ location + rain, family = pospoisson,
##     data = snakes)
##
## Pearson residuals:
##                    Min      1Q  Median     3Q   Max
## loglink(lambda) -2.086 -0.8361 -0.7313 0.4636 12.54
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 0.041426   0.125111   0.331    0.741
## locationV   0.711320   0.149272   4.765 1.89e-06 ***
## rain        0.027329   0.003326   8.217  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Name of linear predictor: loglink(lambda)
##
## Log-likelihood: -218.6267 on 127 degrees of freedom
##
## Number of Fisher scoring iterations: 5
##
## No Hauck-Donner effect found in any of the estimates
```

Longevity of road-killed snakes

Here are the parameter estimates and SE's for both models

```
##               Poisson  Poisson SE  +Poisson  +Poisson SE
## (Intercept)     0.439       0.089     0.041        0.125
## locationV       0.463       0.119     0.711        0.149
## rain            0.022       0.003     0.027        0.003
```

Negative binomial distribution

Recall that for $y_i \sim \text{negBinom}(r, \mu)$

its probability mass function is

$$f(y; \mu, r) = \frac{(y + r - 1)!}{(r-1)!y!} \left(\frac{r}{\mu + r} \right)^r \left(\frac{\mu}{\mu + r} \right)^y$$

$f(y_i; \mu, r)$ gives the probability of $y_i \geq 0$

Neg binomial for zero-truncated data

The probability that $y_i = 0$ is

$$f(y; r, \mu) = \frac{(y + r - 1)!}{(r - 1)! y!} \left(\frac{r}{\mu + r} \right)^r \left(\frac{\mu}{\mu + r} \right)^y$$

$$\Downarrow$$

$$\begin{aligned} f(y_i = 0; r, \mu) &= \frac{(0 + r - 1)!}{(r - 1)! 0!} \left(\frac{r}{\mu + r} \right)^r \left(\frac{\mu}{\mu + r} \right)^0 \\ &= \left(\frac{r}{\mu + r} \right)^r \end{aligned}$$

Neg binomial for zero-truncated data

The probability that $y_i \neq 0$ is therefore

$$f(y_i = 0; r, \mu_i) = \left(\frac{r}{\mu + r}\right)^r$$

$$\Downarrow$$

$$f(y_i \neq 0; r, \mu_i) = 1 - \left(\frac{r}{\mu + r}\right)^r$$

Neg binomial for zero-truncated data

We can now exclude the probability that $y_i = 0$ by dividing the pmf by the probability that $y_i \neq 0$

$$f(y; r, \mu) = \frac{(y+r-1)!}{(r-1)!y!}\left(\frac{r}{\mu+r}\right)^r\left(\frac{\mu}{\mu+r}\right)^y$$

$$\Downarrow$$

$$f(y_i; \lambda_i | y_i > 0) = \frac{\frac{(y+r-1)!}{(r-1)!y!}\left(\frac{r}{\mu+r}\right)^r\left(\frac{\mu}{\mu+r}\right)^y}{1-\left(\frac{r}{\mu+r}\right)^r}$$

$$\Downarrow$$

$$\log \mathcal{L} = \log \mathcal{L}(\text{NB}) - \log\left(1-\left(\frac{r}{\mu+r}\right)^r\right)$$

Longevity of road-killed snakes

Let’s first consider a regular negative binomial regression model

```
## load MASS pkg
library(MASS)
## neg binomial regression
smod_nb <- glm.nb(n_days ~ location + rain, data = snakes,
                  link = "log")
```

Longevity of road-killed snakes

```
##
## Call:
## glm.nb(formula = n_days ~ location + rain, data = snakes, link = "log",
##     init.theta = 4.153875871)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6698  -0.7268  -0.3911   0.3106   4.4713
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 0.418101   0.106859   3.913 9.13e-05 ***
## locationV   0.453524   0.151708   2.989  0.00279 **
## rain        0.025127   0.004529   5.548 2.88e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(4.1539) family taken to be 1)
##
##     Null deviance: 128.298  on 129  degrees of freedom
## Residual deviance:  94.918  on 127  degrees of freedom
## AIC: 469.28
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  4.15
##           Std. Err.:  1.17
##
##  2 x log-likelihood:  -461.279
```

Longevity of road-killed snakes

Now let's fit a zero-truncated neg binomial regression model with `vglm()` from **VGAM**

```
library(VGAM)
## zero-truncated neg binomial regression
smod_ztnb <- vglm(n_days ~ location + rain, data = snakes,
                  family = posnegbinomial)
```

Longevity of road-killed snakes

```
##
## Call:
## vglm(formula = n_days ~ location + rain, family = posnegbinomial,
##     data = snakes)
##
## Pearson residuals:
##                     Min     1Q Median     3Q     Max
## loglink(munb)  0.05515 1.4856 3.7166 4.5336 28.4735
## loglink(size) -0.82119 0.7007 0.7756 0.8972  0.9741
##
## Coefficients:
##                Estimate Std. Error  z value Pr(>|z|)
## (Intercept):1 -17.97164    1.94802   -9.226   <2e-16 ***
## (Intercept):2 -19.22278    0.08787 -218.753   <2e-16 ***
## locationV       0.96528    2.93334    0.329    0.742
## rain            0.06482    0.10346    0.627    0.531
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Names of linear predictors: loglink(munb), loglink(size)
##
## Log-likelihood: -186.8024 on 256 degrees of freedom
##
## Number of Fisher scoring iterations: 3
##
## Warning: Hauck-Donner effect detected in the following estimate(s):
## '(Intercept):1'
```

Longevity of road-killed snakes

Here are the parameter estimates and SE's for both models

```
##                NB   NB SE      +NB  +NB SE
## (Intercept) 0.418   0.107  -17.972   1.948
## locationV   0.454   0.152    0.965   2.933
## rain        0.025   0.005    0.065   0.103
```

QUESTIONS?

Zeros in ecological data

Lots of count data are *zero-inflated*

The data contain more zeros than would be expected under a Poisson or negative binomial distribution

Sources of zeros

In general, there are 4 different types of errors that cause zeros

1. Structural (an animal is absent because the habitat is unsuitable)

Sources of zeros

In general, there are 4 different types of errors that cause zeros

1. Structural (an animal is absent because the habitat is unsuitable)

2. Design (sampling is limited temporally or spatially)

Sources of zeros

In general, there are 4 different types of errors that cause zeros

1. Structural (an animal is absent because the habitat is unsuitable)
2. Design (sampling is limited temporally or spatially)
3. Observer error (inexperience or difficult circumstances)

Sources of zeros

In general, there are 4 different types of errors that cause zeros

1. Structural (an animal is absent because the habitat is unsuitable)
2. Design (sampling is limited temporally or spatially)
3. Observer error (inexperience or difficult circumstances)
4. Process error (habitat is suitable but unused)

Sources of zeros

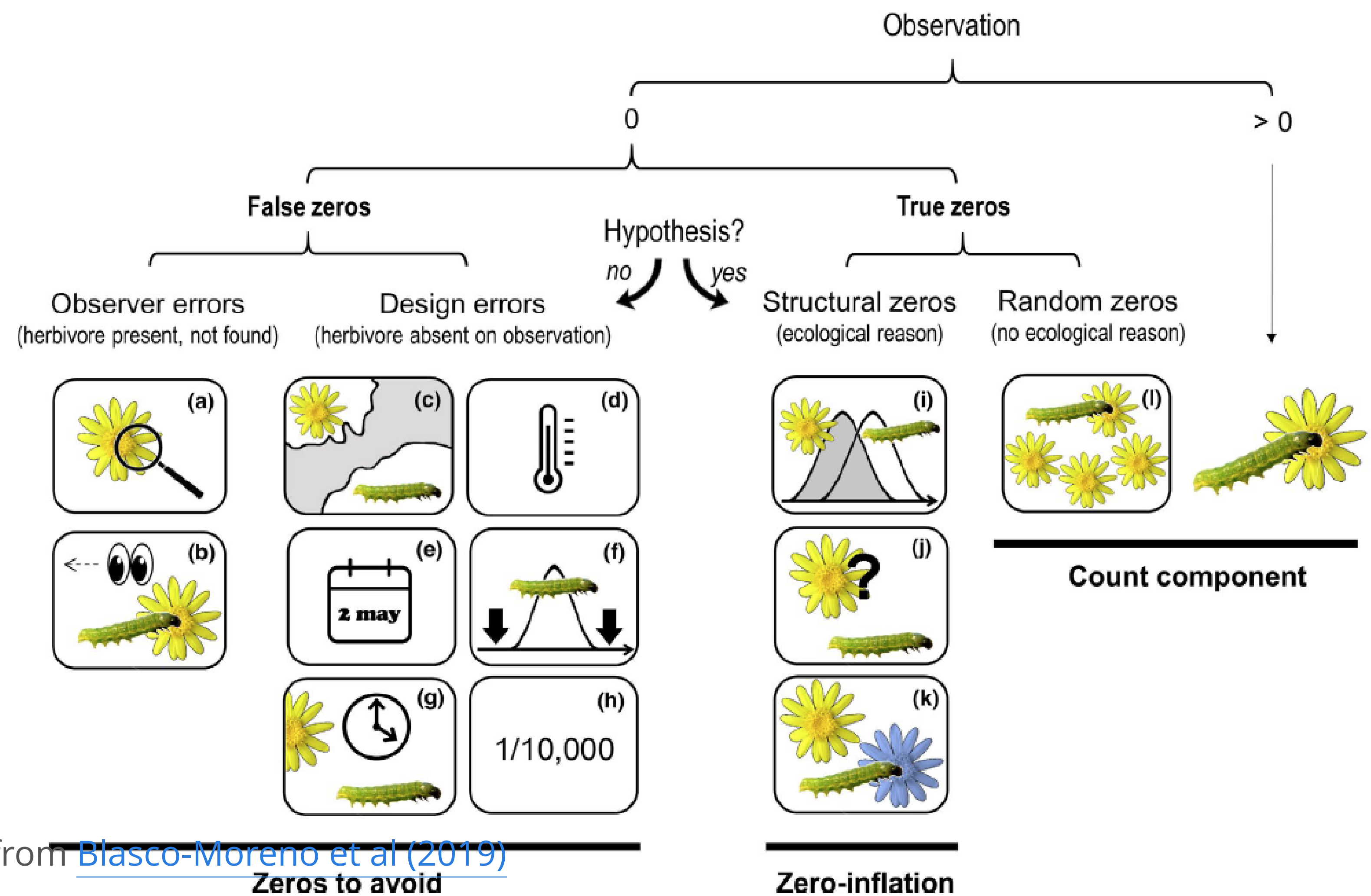


Image from [Blasco-Moreno et al (2019)]

Approaches to zero-inflated data

There are 2 general approaches for dealing with zero-inflated data

1. Zero-altered ("hurdle") models

2. Zero-inflated ("mixture") models

Hurdle models

Hurdle models do not discriminate among the 4 types of zeros

The data are treated as 2 distinct groups:

1. Zeros
2. Non-zero counts

Hurdle models

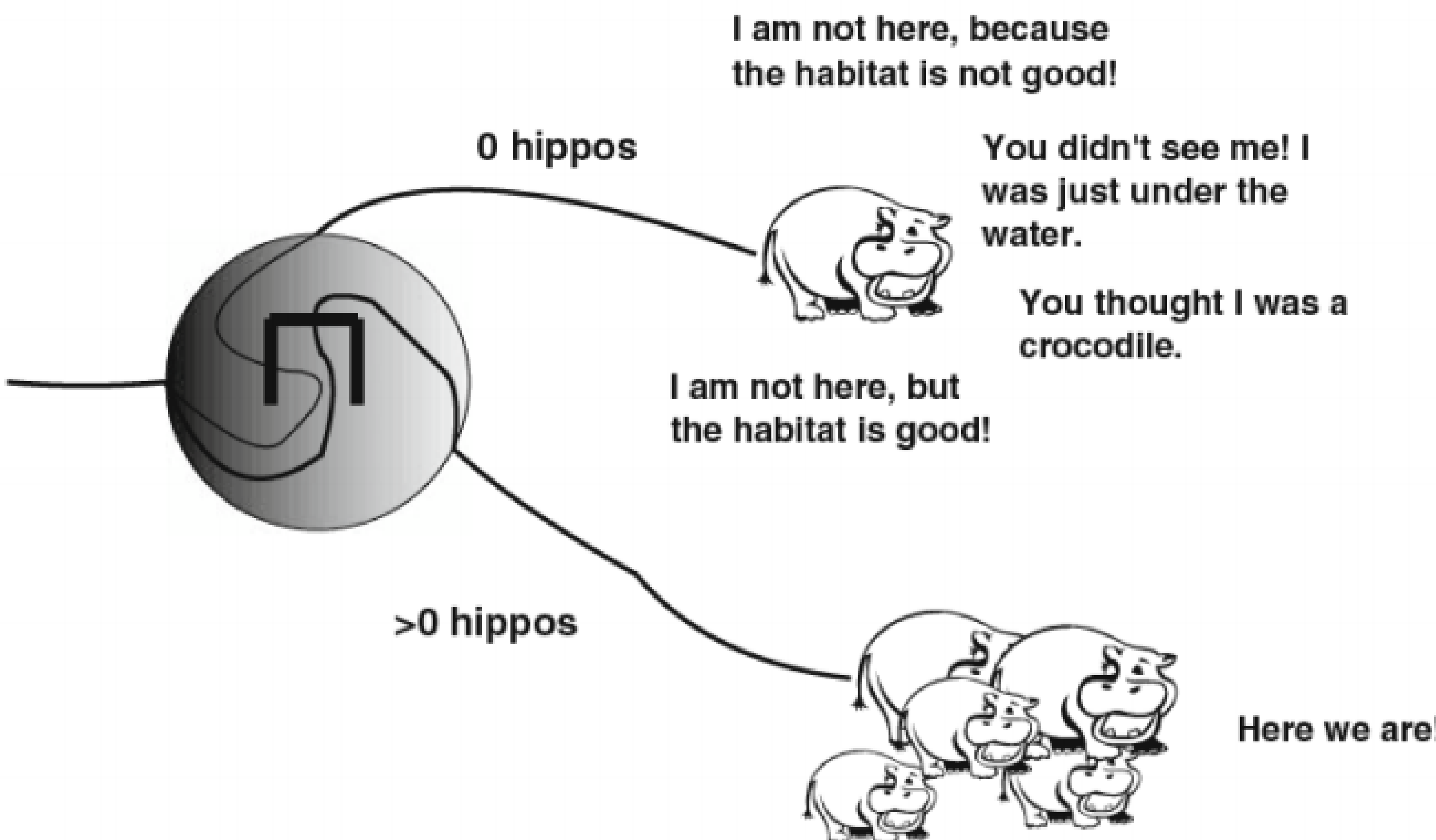


Image from Zuur et al (2009)

Hurdle models

Hurdle models consist of 2 parts

1. Use a binomial model to determine the probability of a zero
2. If non-zero ("over the hurdle"), use a truncated Poisson or negative binomial to model the positive counts

Zero-altered Poisson (ZAP) models

A zero-altered Poisson (ZAP) model is given by

$$f_{\mathrm{ZAP}}(y;\pi,\lambda) = \begin{cases} f_{\mathrm{binomial}}(y=0;\pi) \\ [1 - f_{\mathrm{binomial}}(y=0;\pi)] \times \left(\frac{f_{\mathrm{Poisson}}(y=0;\lambda)}{1 - f_{\mathrm{Poisson}}(y=0;\lambda)} \right) \end{cases}$$

π is the probability of finding *any* individuals

λ is the mean (and variance) of the *positive counts*

Zero-altered Poisson (ZAP) models

We can model both parameters as functions of covariates

Probability of detection

$$\text{logit}(\pi) = \mathbf{X}_d \boldsymbol{\beta}_d$$

Mean and variance of the positive counts

$$\log(\lambda) = \mathbf{X}_c \boldsymbol{\beta}_c$$

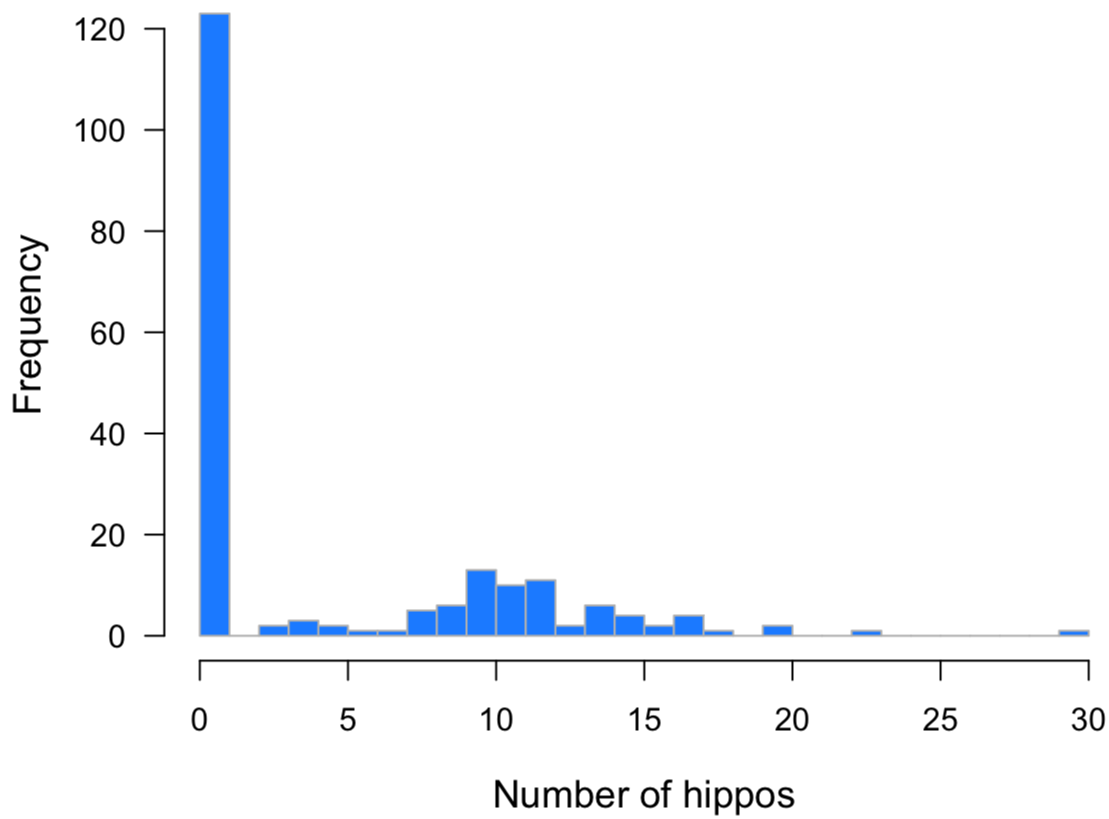
Counts of hippos

Let's apply a ZAP model to survey data for hippos

We'll assume the following

- the probability of finding hippos increases with water availability
- the number of hippos increases with tree density

Counts of hippos



ZAP model for hippos

Detection as a function of water availability W

$$z_i \sim \text{Bernoulli}(\pi_i)$$
$$\text{logit}(\pi) = \gamma_0 + \gamma_1 W_i$$

Positive counts as a function of tree density T

$$c_i \sim \text{Poisson}^+(\lambda_i)$$
$$\log(\lambda) = \beta_0 + \beta_1 T_i$$

Total counts as a function of detections and positive counts

$$y_i = z_i c_i$$

ZAP model for hippos

We can fit ZAP models in R with `hurdle()` from the **pscl** package

The formula for ZAP models is specified as

`y ~ predictors_of_counts | predictors_for_detection`

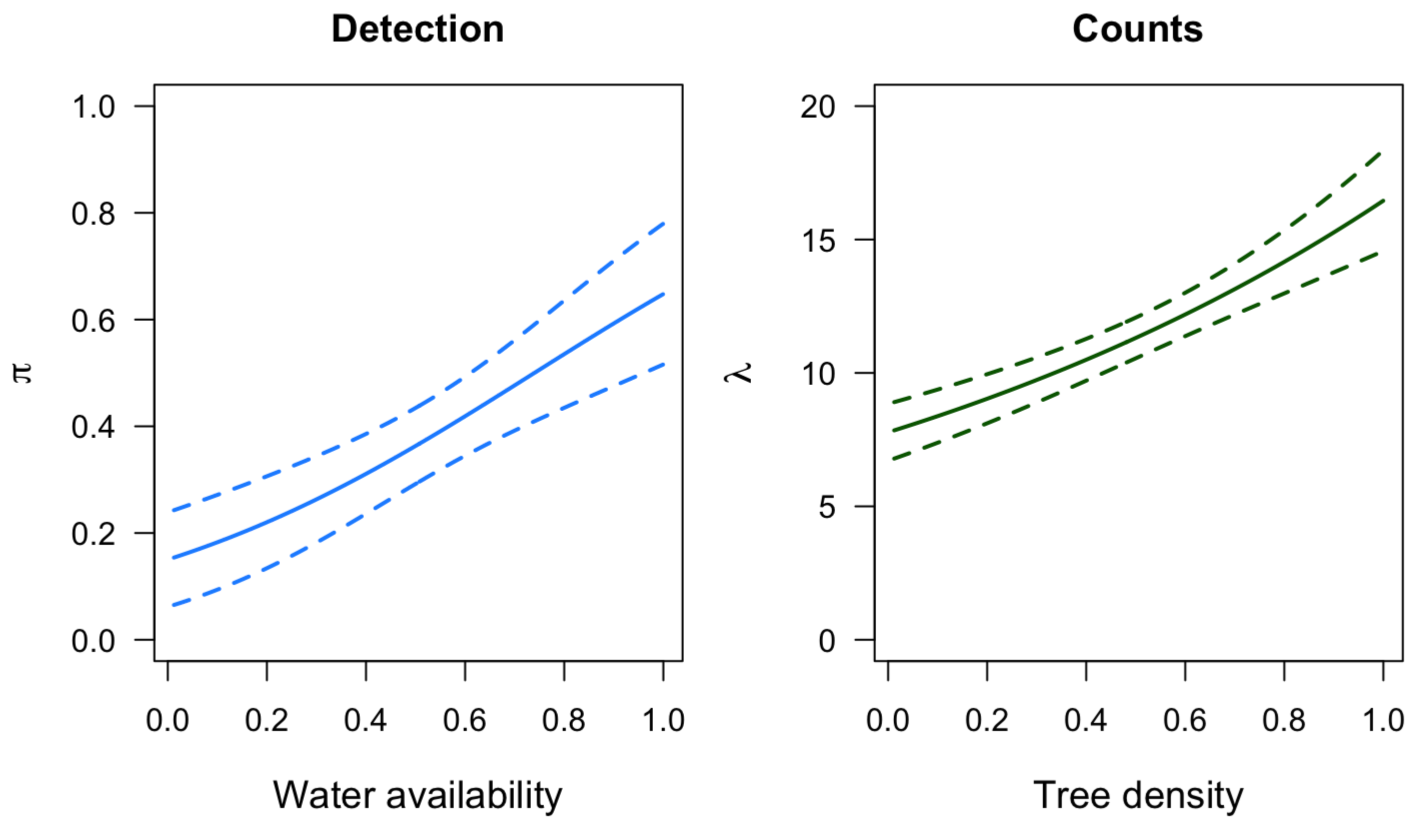
```r
## load pscl
library(pscl)
## fit hurdle model
hippo_zap <- hurdle(y ~ trees | water)
```

ZAP model for hippos

```
summary(hippo_zap)
```

```
##
## Call:
## hurdle(formula = y ~ trees | water)
##
## Pearson residuals:
##     Min      1Q  Median      3Q     Max
## -1.2165 -0.7104 -0.4803  0.9193  2.6988
##
## Count model coefficients (truncated poisson with log link):
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.05104    0.06955  29.492  < 2e-16 ***
## trees        0.74967    0.10843   6.914 4.71e-12 ***
## Zero hurdle model coefficients (binomial with logit link):
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.7326     0.3519  -4.924 8.48e-07 ***
## water         2.3422     0.5676   4.126 3.69e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of iterations in BFGS optimization: 11
## Log-likelihood: -326.1 on 4 Df
```

ZAP model for hippos



Zero-altered neg binomial (ZANB)

A zero-altered negative binomial (ZANB) model is given by

$$f_{\text{ZANB}}(y;\pi,\mu,r) = \begin{cases} f_{\text{binomial}}(y=0;\pi) \\ [1-f_{\text{binomial}}(y=0;\pi)] \times \left(\frac{f_{\text{NB}}(y=0;\mu,r)}{1-f_{\text{NB}}(y=0;\mu,r)} \right) \end{cases}$$

π is the probability of finding *any* individuals

μ is the mean the *positive counts*

r is the scale for the *positive counts*

QUESTIONS?

Zero-inflated (mixture) models

Zero-inflated (mixture) models treat the zeros as coming from 2 sources

1. observation errors (missed detections)
2. ecological (function of environment)

Zero-inflated models

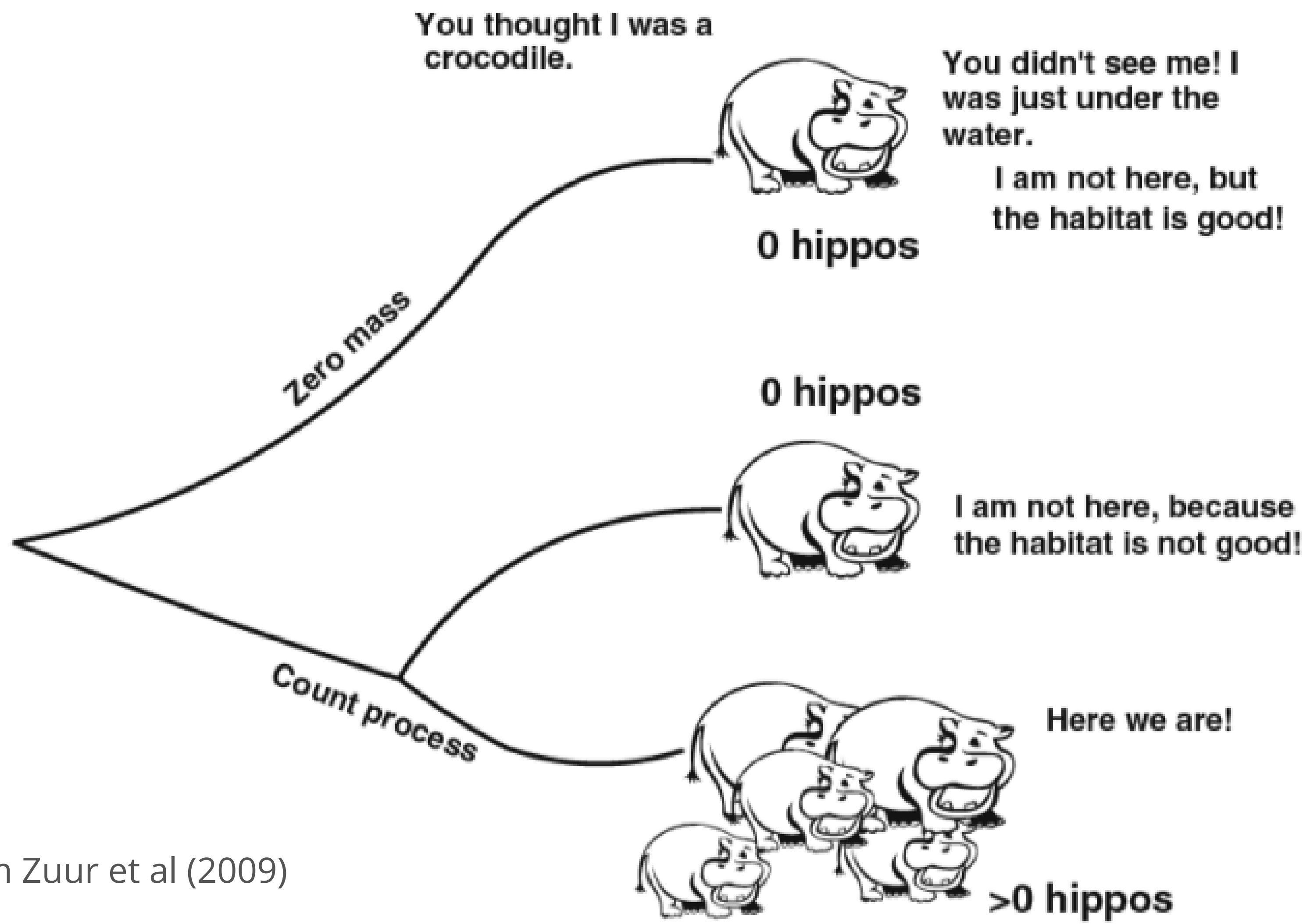


Image from Zuur et al (2009)

Mixture models

Zero-inflated (mixture) models consist of 2 parts

1. Use a binomial model to determine the probability of a zero

2. Use a Poisson or negative binomial to model counts, which can include zeros

Zero-inflated Poisson (ZIP) models

Probability of a zero count comes from 2 sources:

1. false zeros (missed detections)
2. true zeros (ecological reasons)

Pr(zero) = Pr(false zero) + Pr(true zero) $\times$ Pr(count = 0)

Zero-inflated Poisson (ZIP) models

A zero-inflated Poisson (ZIP) model is given by

$$f_{\text{ZIP}}(y = 0) = f_{\text{Binomial}}(\pi) + [1 - f_{\text{Binomial}}(\pi)]f_{\text{Poisson}}(y = 0; \lambda)$$

$$f_{\text{ZIP}}(y | y > 0) = [1 - f_{\text{Binomial}}(\pi)]f_{\text{Poisson}}(y; \lambda)$$

π is the probability of *false zeros* (missed detections)

λ is the mean (and variance) of *all counts* (including zeros)

Zero-inflated Poisson (ZIP) models

We can model both parameters as functions of covariates

Probability of detection

$$\text{logit}(\pi) = \mathbf{X}_d \boldsymbol{\beta}_d$$

Mean and variance of the counts

$$\log(\lambda) = \mathbf{X}_c \boldsymbol{\beta}_c$$

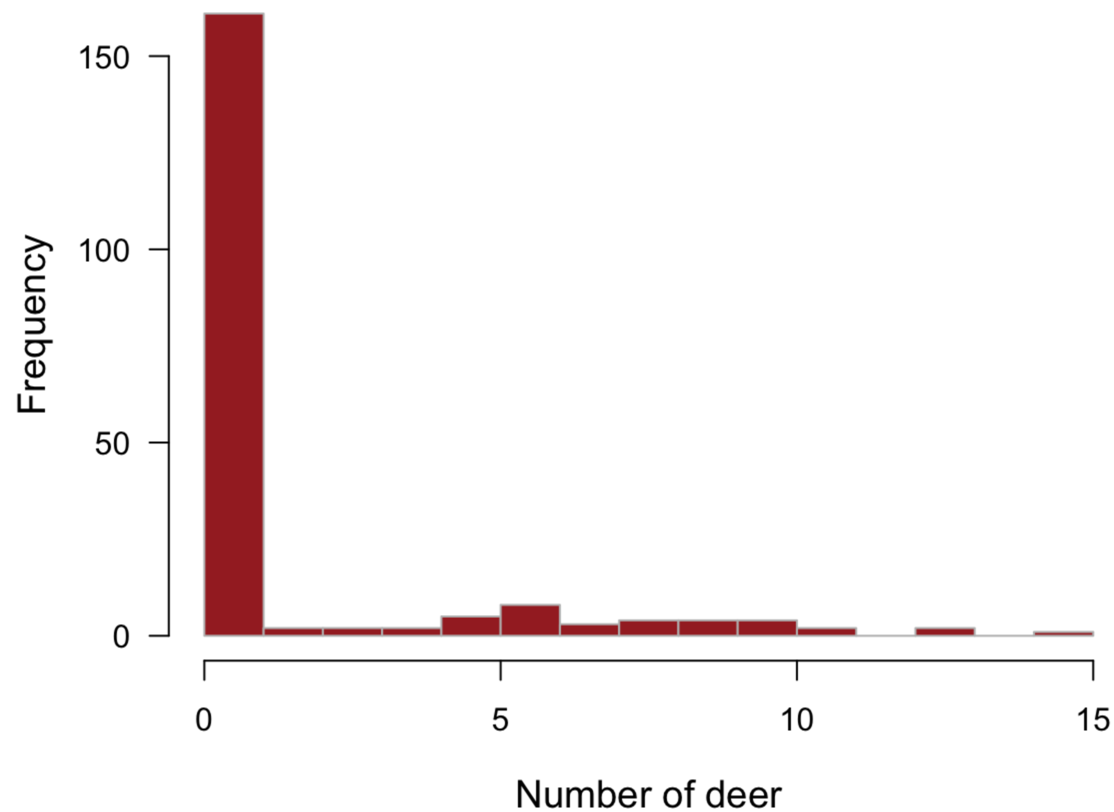
Counts of deer

Let’s apply a ZIP model to survey data for white tailed deer

We’ll assume the following

- the probability of detecting deer decreases with tree density
- the number of deer increases with tree density

Counts of deer



ZIP model for deer

Non-detection as a function of tree density T

$$z_i \sim \text{Bernoulli}(\pi_i)$$
$$\text{logit}(\pi) = \gamma_0 + \gamma_1 T_i$$

Counts as a function of tree density T

$$c_i \sim \text{Poisson}(\lambda_i)$$
$$\log(\lambda) = \beta_0 + \beta_1 T_i$$

Total counts as a function of detections and positive counts

$$y_i = (1 - z_i)c_i$$

ZIP model for deer

We can fit ZIP models in R with `zeroinfl()` from the **pscl** package

The formula for ZIP models is specified as

`y ~ predictors_of_counts | predictors_for_detection`

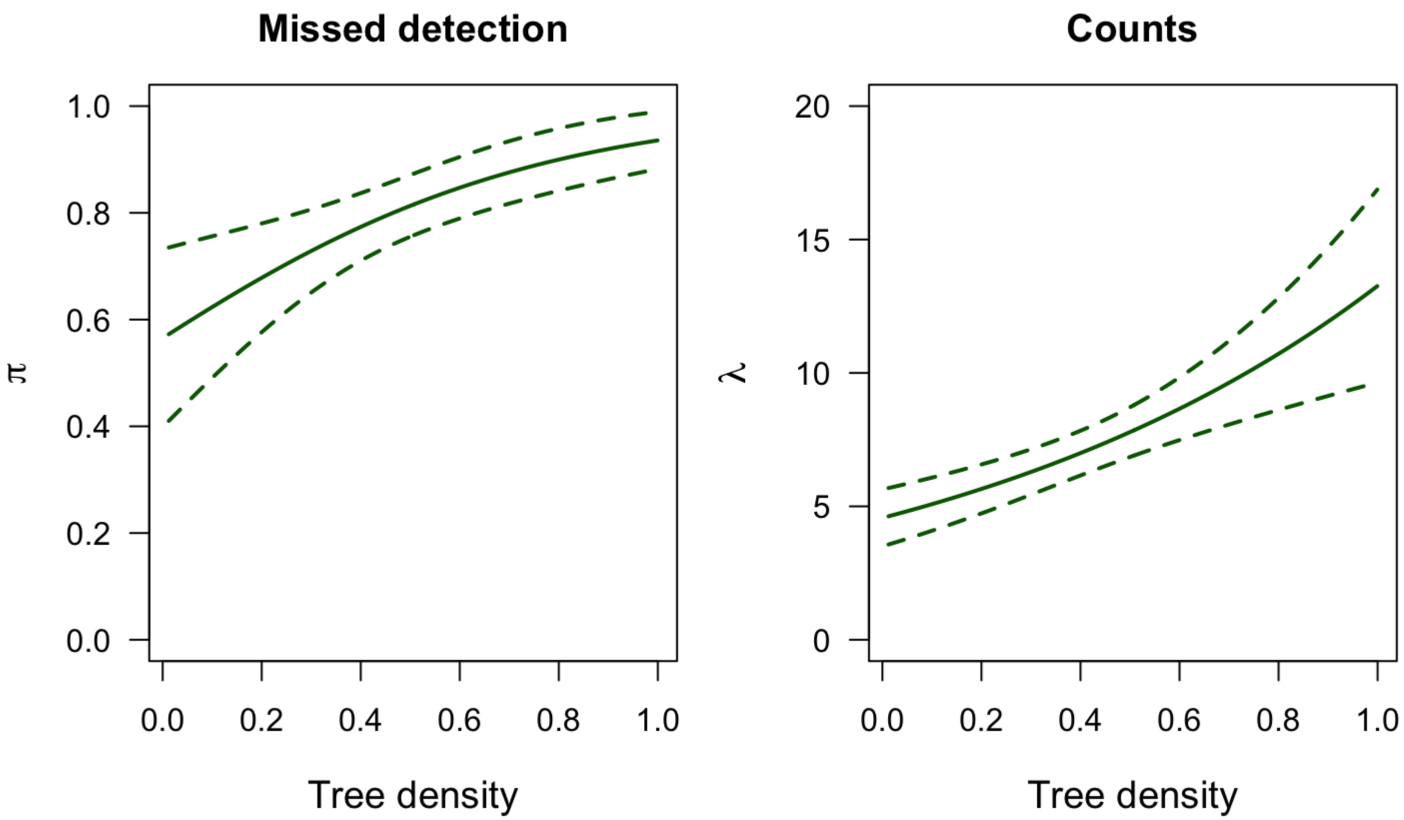
```
## fit hurdle model
deer_zip <- zeroinfl(y ~ trees | trees)
```

ZIP model for deer

```
summary(deer_zip)
```

```
##
## Call:
## zeroinfl(formula = y ~ trees | trees)
##
## Pearson residuals:
##     Min      1Q  Median      3Q     Max
## -0.7312 -0.5103 -0.3674 -0.2611  4.1697
##
## Count model coefficients (poisson with log link):
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.5190     0.1183  12.841  < 2e-16 ***
## trees         1.0660     0.2272   4.693 2.69e-06 ***
##
## Zero-inflation model coefficients (binomial with logit link):
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.2625     0.3438   0.764 0.445104
## trees         2.4158     0.7048   3.428 0.000609 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of iterations in BFGS optimization: 12
## Log-likelihood: -185.8 on 4 Df
```

ZIP model for deer



Zero-inflated neg binomial (ZINB)

A zero-inflated negative binomial (ZINB) model is given by

$$f_{\text{ZIP}}(y = 0) = f_{\text{Binomial}}(\pi) + [1 - f_{\text{Binomial}}(\pi)]f_{\text{NB}}(y = 0; \mu, r)$$

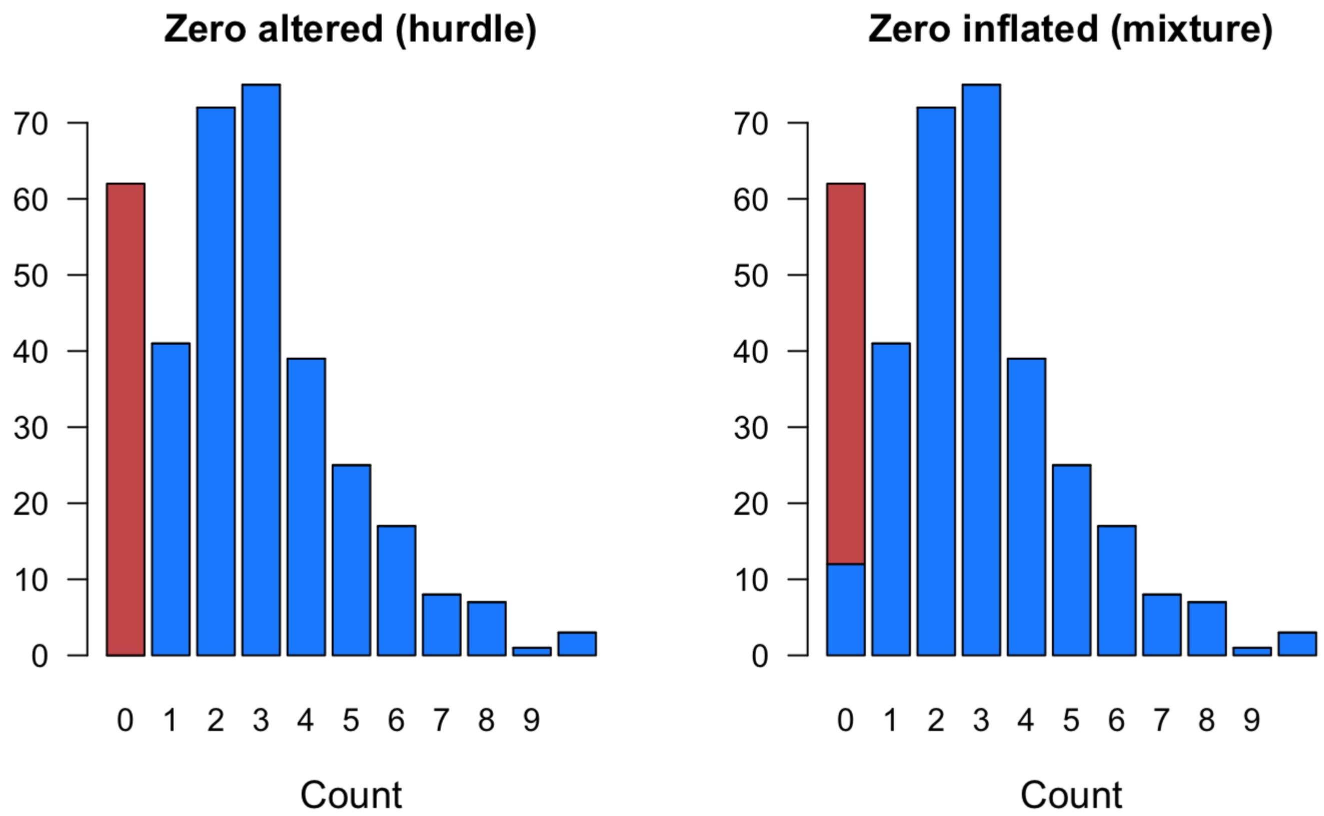
$$f_{\text{ZIP}}(y | y > 0) = [1 - f_{\text{Binomial}}(\pi)]f_{\text{NB}}(y; \mu, r)$$

π is the probability of *false zeros* (missed detections)

μ is the mean of *all counts* (including zeros)

r is the scale of the counts

ZA versus ZI models for counts



Steps for modeling counts

1. Understand the system of interest

Formulate good hypotheses and create a robust study design

Steps for modeling counts

1. Understand the system of interest
2. Detect and classify zeros

Remove false zeros due to design or observer errors

Steps for modeling counts

1. Understand the system of interest
2. Detect and classify zeros
3. Identify suitable covariates for zeros & non-zeros

What are the causes of zeros (non-zeros)

Steps for modeling counts

1. Understand the system of interest
2. Detect and classify zeros
3. Identify suitable covariates for zeros & non-zeros
4. Test for overdispersion

Steps for modeling counts

1. Understand the system of interest
2. Detect and classify zeros
3. Identify suitable covariates for zeros & non-zeros
4. Test for overdispersion
5. Choose appropriate model

Sources of zeros and approaches

Source	Reason	Over-dispersion	Zero inflation	Approach
Random	Sampling variability	No	No	Poisson
		Yes	No	Neg binomial
Structural	Outside count process	No	Yes	ZAP or ZIP
		Yes	Yes	ZANB or ZINB