Bio Stats II : Lecture 14, Zero-inflated models

Acknowledgments: Sean Anderson

Gavin Fay & Amanda Hart

03/25/2025

This Week. . .

3/25: Zero-Inflated Models
3/26: UMass Marine Science Symposium
3/27: Spatial GLMMs

Objectives

- Review generalized linear modeling
- Rationale for modeling zero-inflation
- Zero inflated models for count data
- Zero-altered (hurdle) models for count data
- Hurdle models for continuous response

Generalized linear modeling

Recall:

$$\eta = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

$$f_Y(y; \mu, \varphi) = \exp\left[\frac{A}{\varphi} y\lambda(\mu) - \gamma(\lambda(\mu)) + \tau(y, \varphi)\right]$$

$$\mu = m(\eta),\ \eta = m^{-1}(\mu) = l(\mu)$$

The combination of a response distribution, a link function and other information needed to carry out the modeling exercise is called the *family* of the generalized linear model.

The `glm()` function

The R function to fit a generalized linear model is `glm()` which uses the form:
`fitted.model <- glm(formula, family=family.generator, data=data.frame)`

Only new piece is the call to 'family.generator'
Where there is a choice of link, link can be supplied with the family name as a parameter.

Simple (inefficient) use: The following are equivalent.

```
> RIKZ.lm1 <- lm(Richness ~ NAP, data = RIKZ)
> RIKZ.glm1 <- glm(Richness ~ NAP, family = gaussian,
+                  data = RIKZ)
```

How do we represent count data?

Poisson regression

$$P(X = x) = \frac{e^{-\mu}\mu^x}{x!},\ \mu_i = e^{\alpha+\beta_1 x_{1,i}+\cdots+\beta_j x_{j,i}}$$

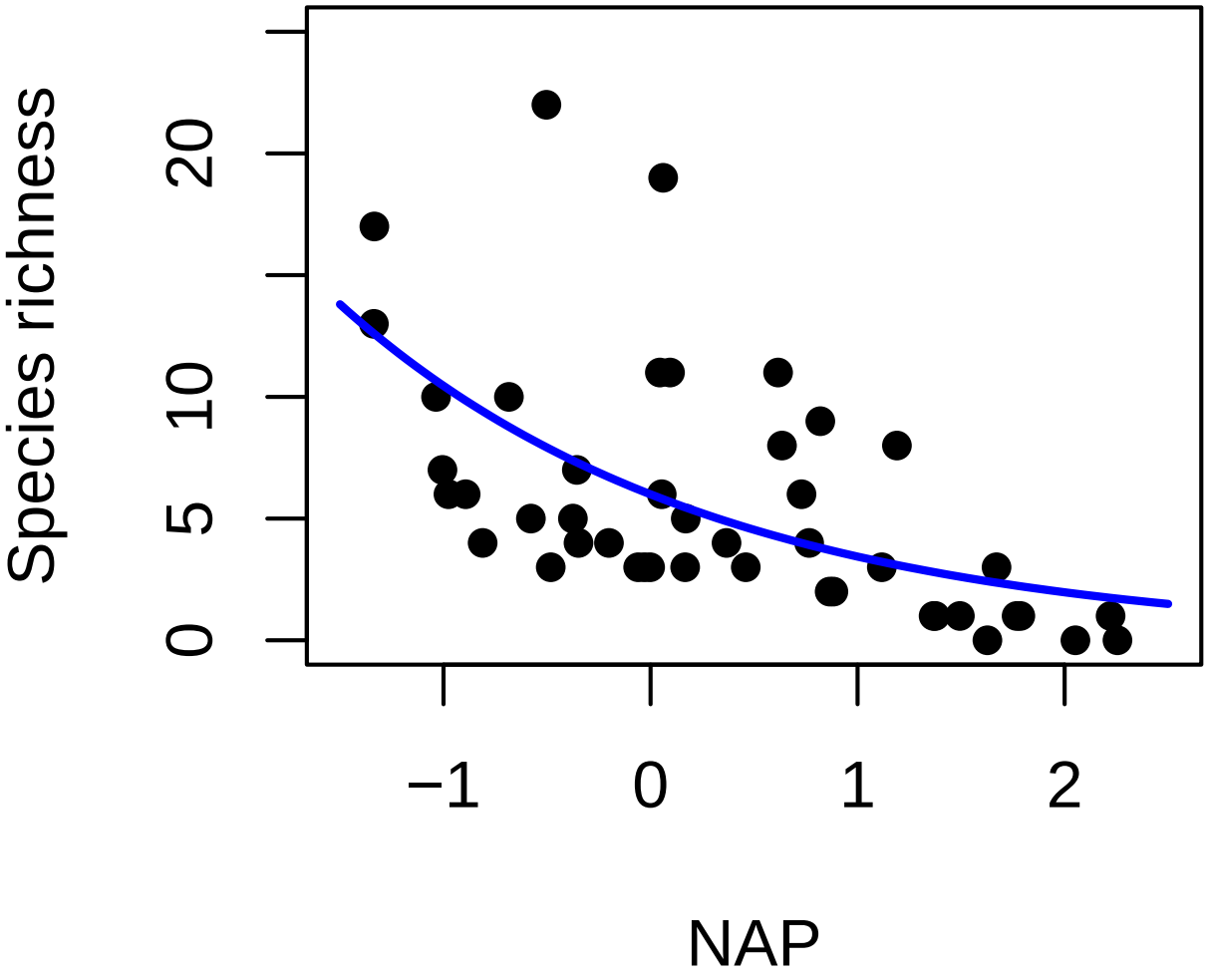
```
> RIKZ_poisson <- glm(Richness ~ NAP, data = RIKZ,
+                     family = poisson)
```

Note that the default link for the poisson is `log` so we don't have to specify here (see `?family`).

summary(RIKZ_poisson)

```
Call:
glm(formula = Richness ~ NAP, family = poisson, data = RIKZ)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.79100    0.06329  28.297  < 2e-16 ***
NAP         -0.55597    0.07163  -7.762 8.39e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 179.75  on 44  degrees of freedom
Residual deviance: 113.18  on 43  degrees of freedom
AIC: 259.18

Number of Fisher Scoring iterations: 5
```

Observed and fitted values for Poisson RIKZ

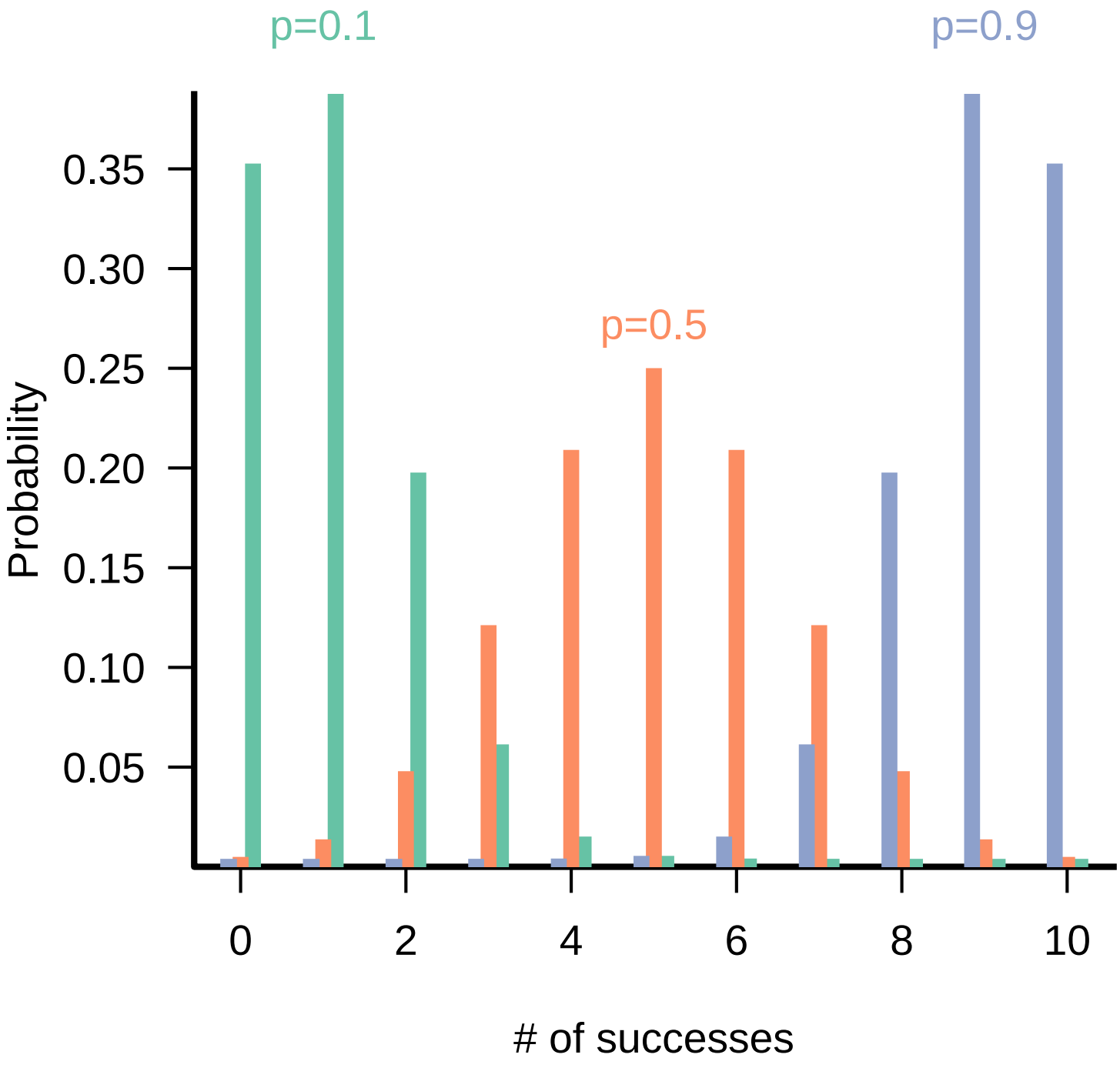


Binomial

Number of successes from a fixed number of Bernoulli trials.
Two possible outcomes on each trial, success or failure.
Probability of success is the same in each trial.

Range: discrete, $0 \leq x \leq N$
Distribution:

$$\binom{N}{x} p^x (1-p)^{N-x}$$


p=0.1
p=0.9
p=0.5
0.35
0.30
0.25
0.20
0.15
0.10
0.05
Probability
0
2
4
6
8
10
# of successes

Poisson

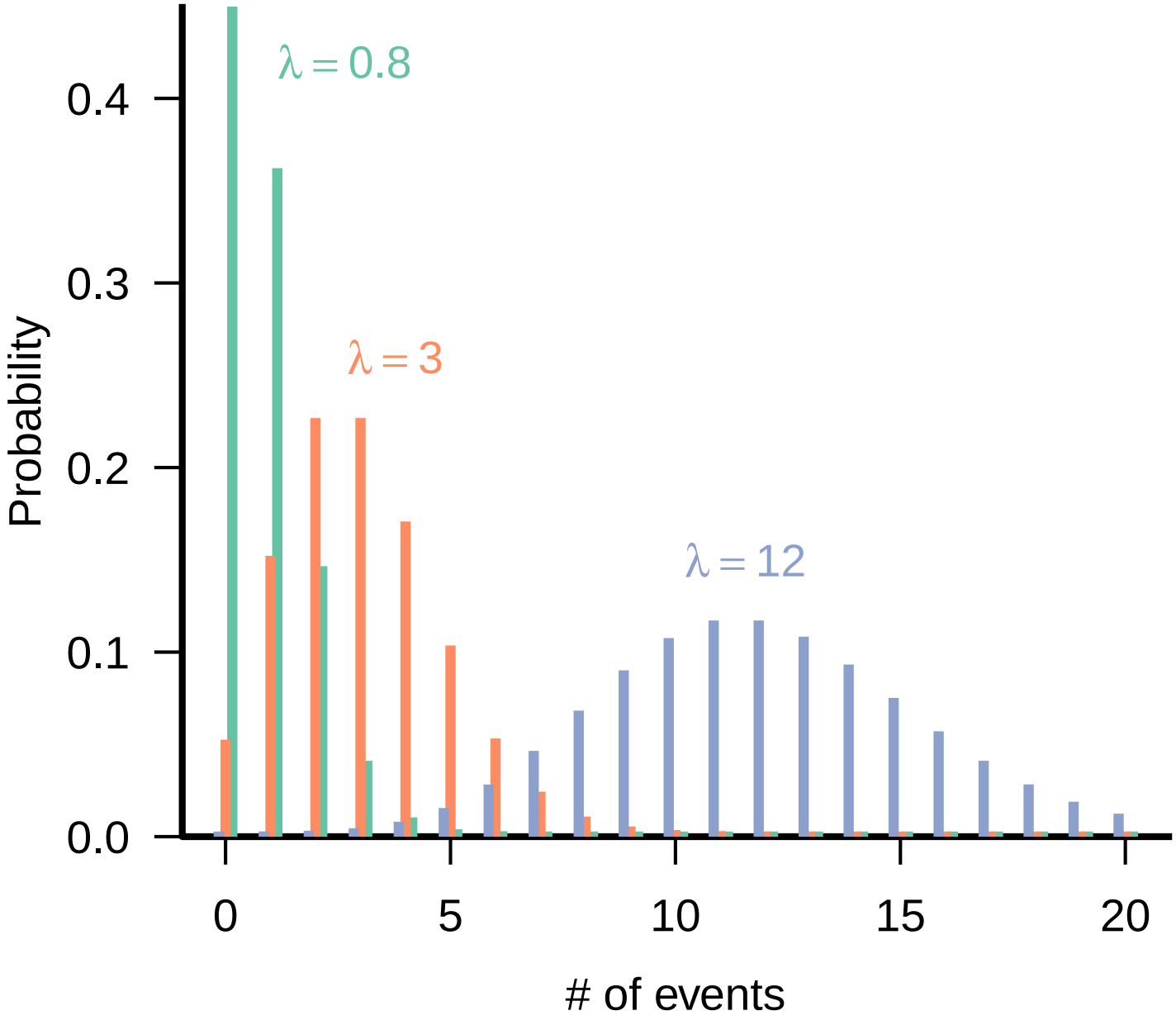
Describes events which occur randomly and independently in time.

Range: discrete ($0 \leq x$)
Distribution:

$$\frac{e^{-\lambda}\lambda^n}{n!} \textbf{ or } \frac{e^{-rt}(rt)^n}{n!}$$

Parameters: λ (real, positive), expected number per sample [`lambda`] **or** r (real, positive), expected number per unit effort, area, time, etc. (*arrival rate*)

Poisson



Negative Binomial

For binomial trials, the number of failures before n successes.

In ecology, most often used because it is discrete like the Poisson but the variance can be greater than the mean (*overdispersed*).

Range: discrete, $x \geq 0$
Distribution:

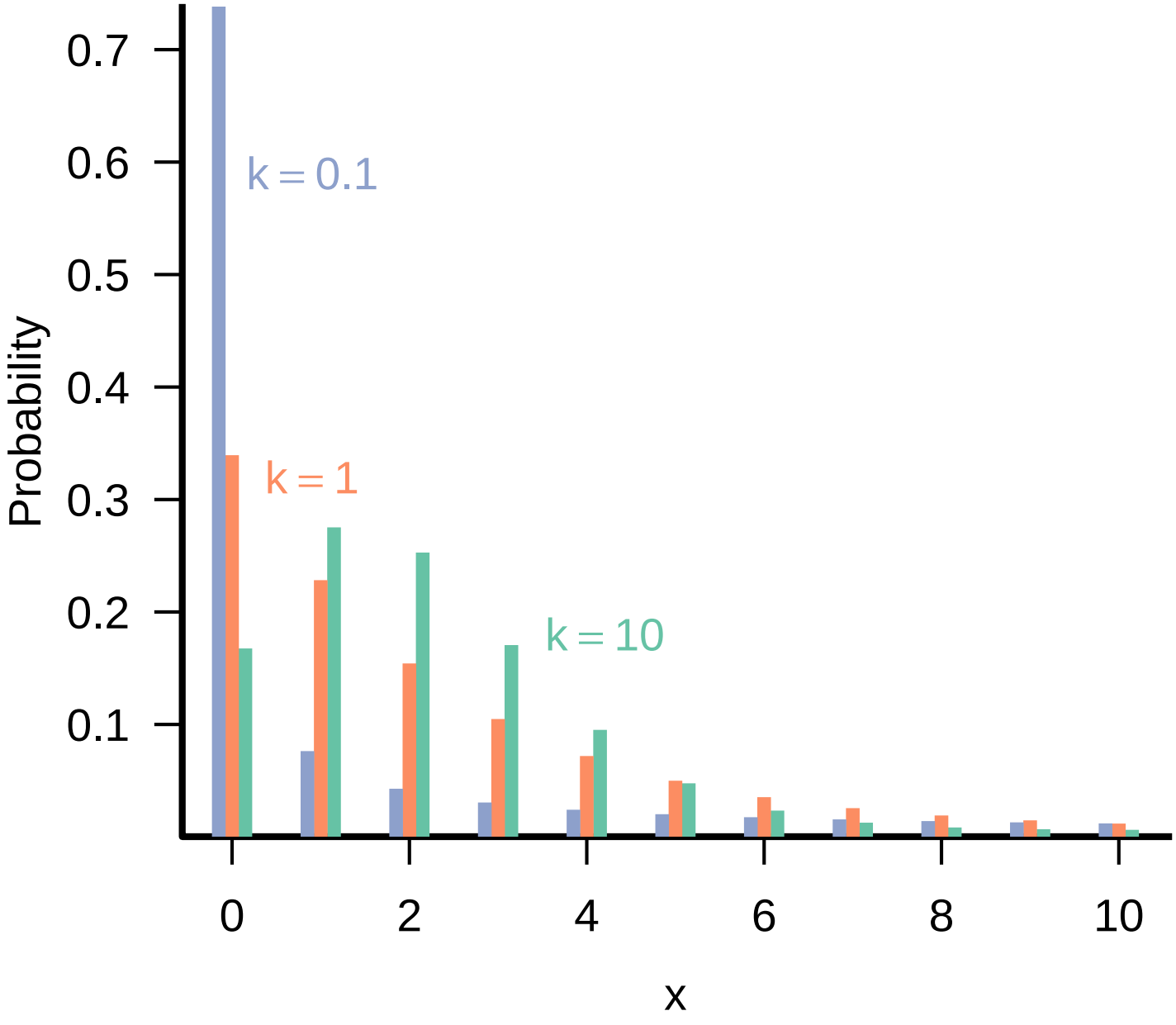
$$P(X = x) = \frac{(n + x - 1)!}{(n - 1!)x!} p^n (1 - p)^x$$

$$\textbf{or } \frac{\Gamma(k + x)}{\Gamma(k)x!} (k/(k + \mu))^k (\mu/(k + \mu))^x$$

Parameters:
μ (real, positive) expected number of counts [`mu`]
k (real, positive), overdispersion parameter [`size`]

The negative binomial is also the result of a Poisson sampling process where λ is Gamma-distributed.

Negative Binomial ($\mu = 2$ all cases)



Quasi-Poisson Regression

We introduced a *dispersion* parameter ρ to allow for deviations from the Poisson (so $\text{Var}(Y_i) = \rho\mu_i$)

$\rho > 1$, more spread, overdispersion
$\rho < 1$, underdispersion

```
Call:
glm(formula = Richness ~ NAP, family = quasipoisson, data = RIKZ

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.7910     0.1104  16.218  < 2e-16 ***
NAP          -0.5560     0.1250  -4.448 6.02e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasipoisson family taken to be 3.0441

    Null deviance: 179.75  on 44  degrees of freedom
Residual deviance: 113.18  on 43  degrees of freedom
AIC: NA
```

Negative Binomial GLM

```
> RIKZ_negbin <- glm.nb(Richness ~ NAP, data = RIKZ)
> summary(RIKZ_negbin)

Call:
glm.nb(formula = Richness ~ NAP, data = RIKZ, init.theta = 3.712
    link = log)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.7985     0.1037  17.336  < 2e-16 ***
NAP          -0.6230     0.1122  -5.552 2.83e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(3.7122) family taken

    Null deviance: 76.231  on 44  degrees of freedom
Residual deviance: 46.275  on 43  degrees of freedom
AIC: 231.75

Number of Fisher Scoring iterations: 1

              Theta:  3.71
```

Dispersion and many zeroes

A disproportionate number of zeroes can lead to overdispersion.

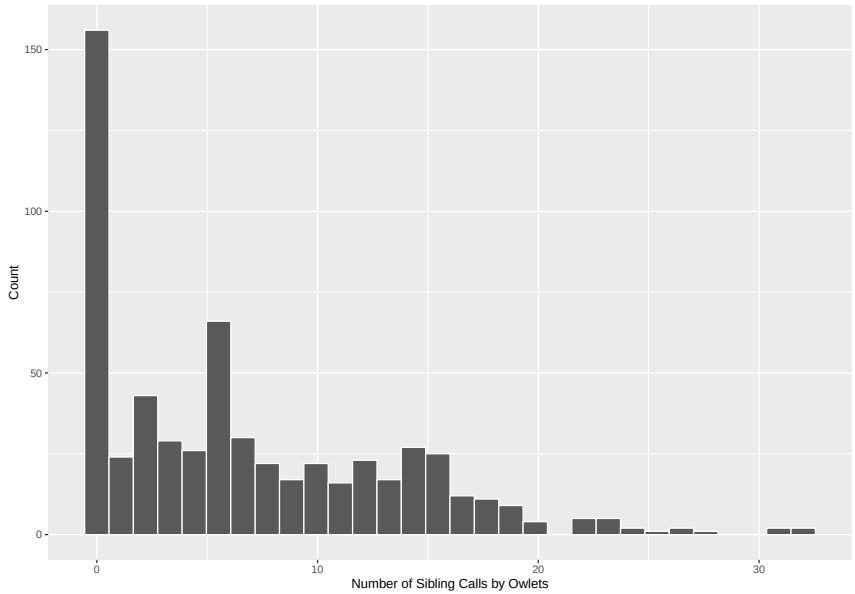
$$\rho = \frac{1}{n-p} \frac{\sum_{i=1}^{n} (Y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}$$

The amount of overdispersion left over when fitting a model (i.e. in the residuals) can be used as a model diagnostic.

Dispersion and many zeroes

Some datasets contain an abundance of zeroes that cannot be captured by the poisson, negative binomial, etc.

e.g. sibling calls of owlets (Bolker et al. 2012)



What are sources of zeros?

What are sources of zeros?

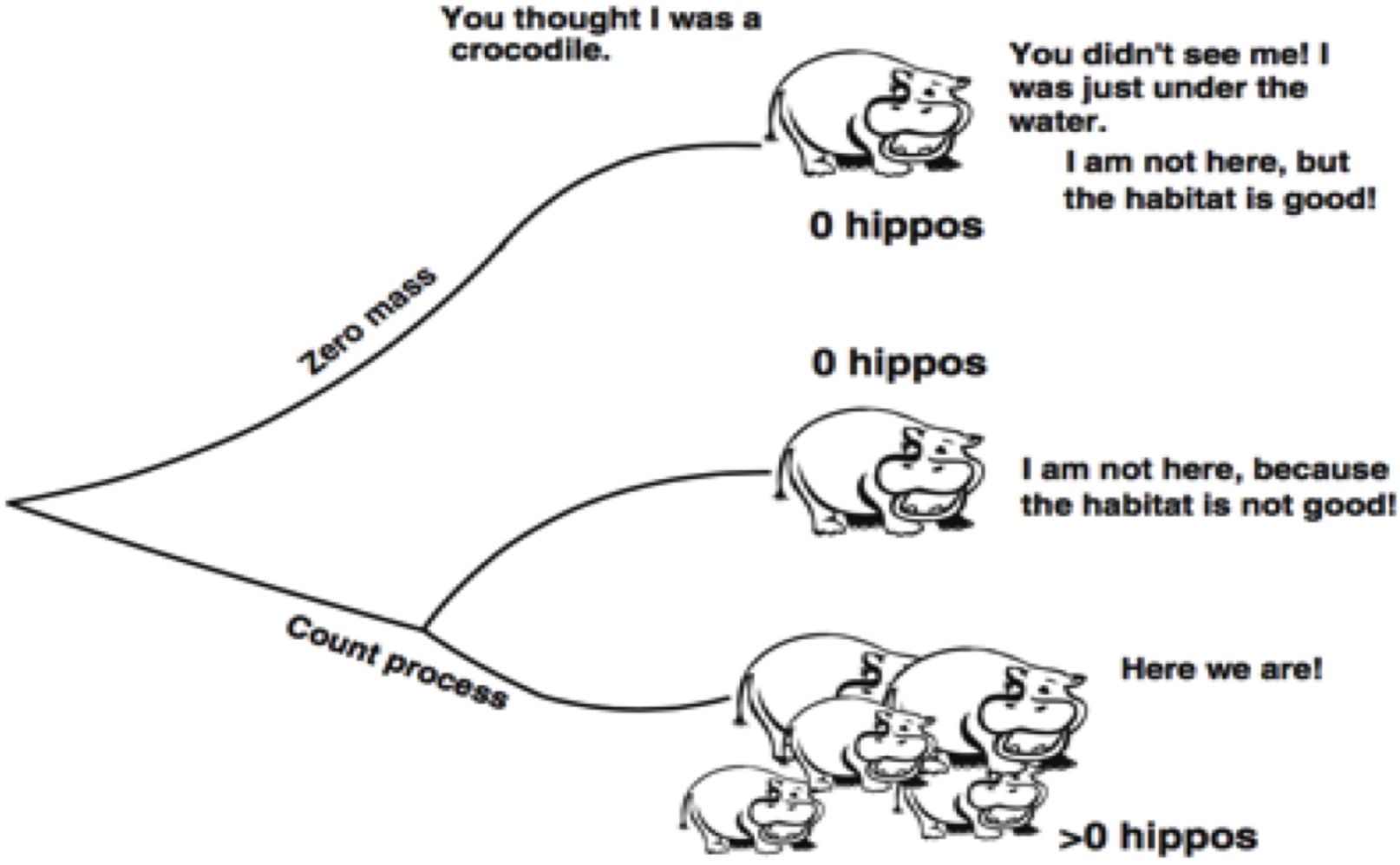
1) Structural error -> Sampling the wrong habitat
2) Design error -> Poor experimental design or sampling
3) Observer error -> Different levels of experiment
4) 'Bird' error -> Habitat is suitable, but not used
5) Naughty naughts -> Bad zeros

How to deal with zeros in count data?

- Ignore them (usually not appropriate unless you know something went wrong that would explain the zero)
- Aggregate data
- Replace zeros (again, usually not the best choice)
- Account for zeros using:
  - Poisson
  - Negative Binomial
  - ZIP: Zero-Inflated Poisson
  - ZINB: Zero-Inflated negative binomial
  - ZAP: Zero-altered Poisson
  - ZANB: Zero-altered Negative Binomial

Zero-altered models are also known as *hurdle* models.

Zero Inflated Mixture Model

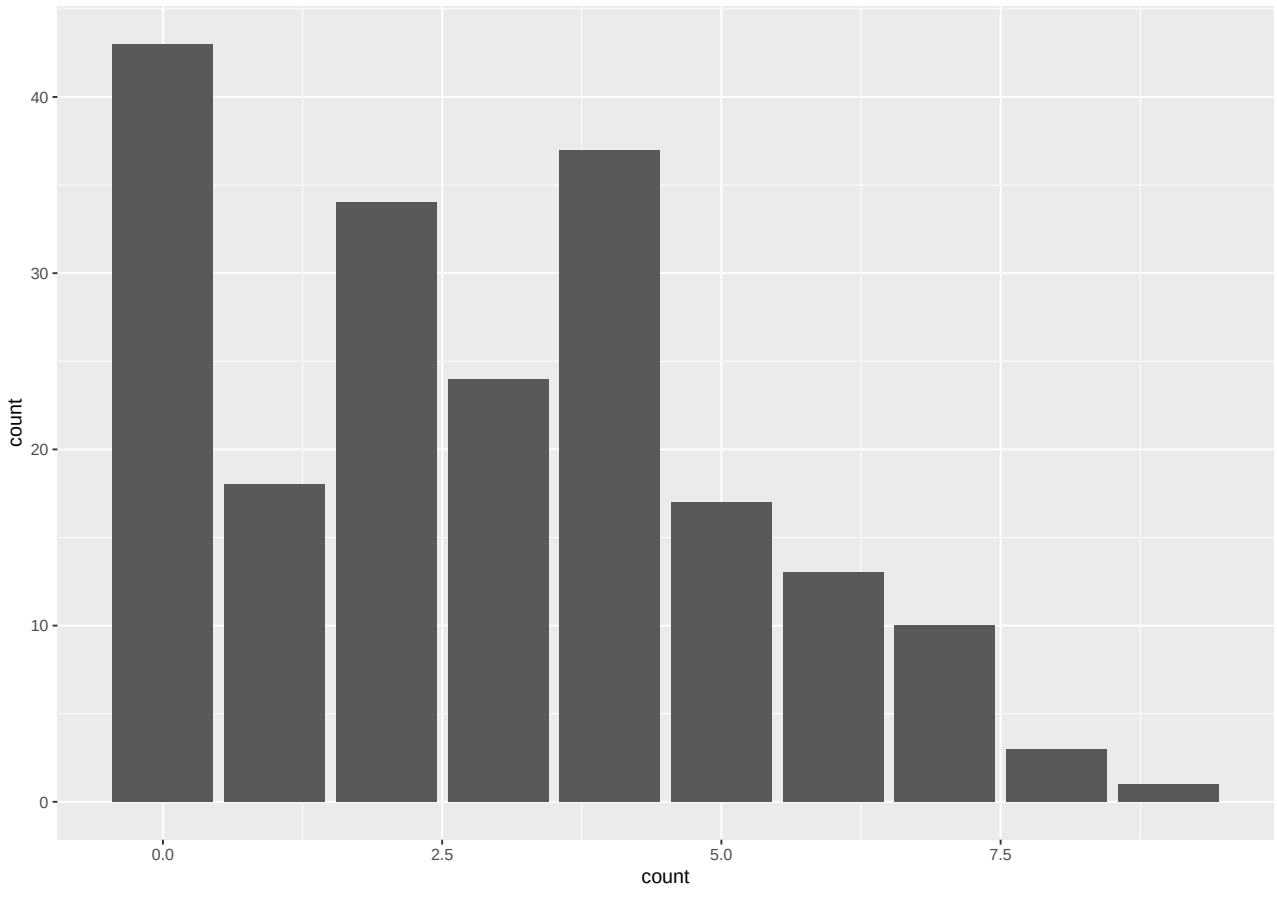


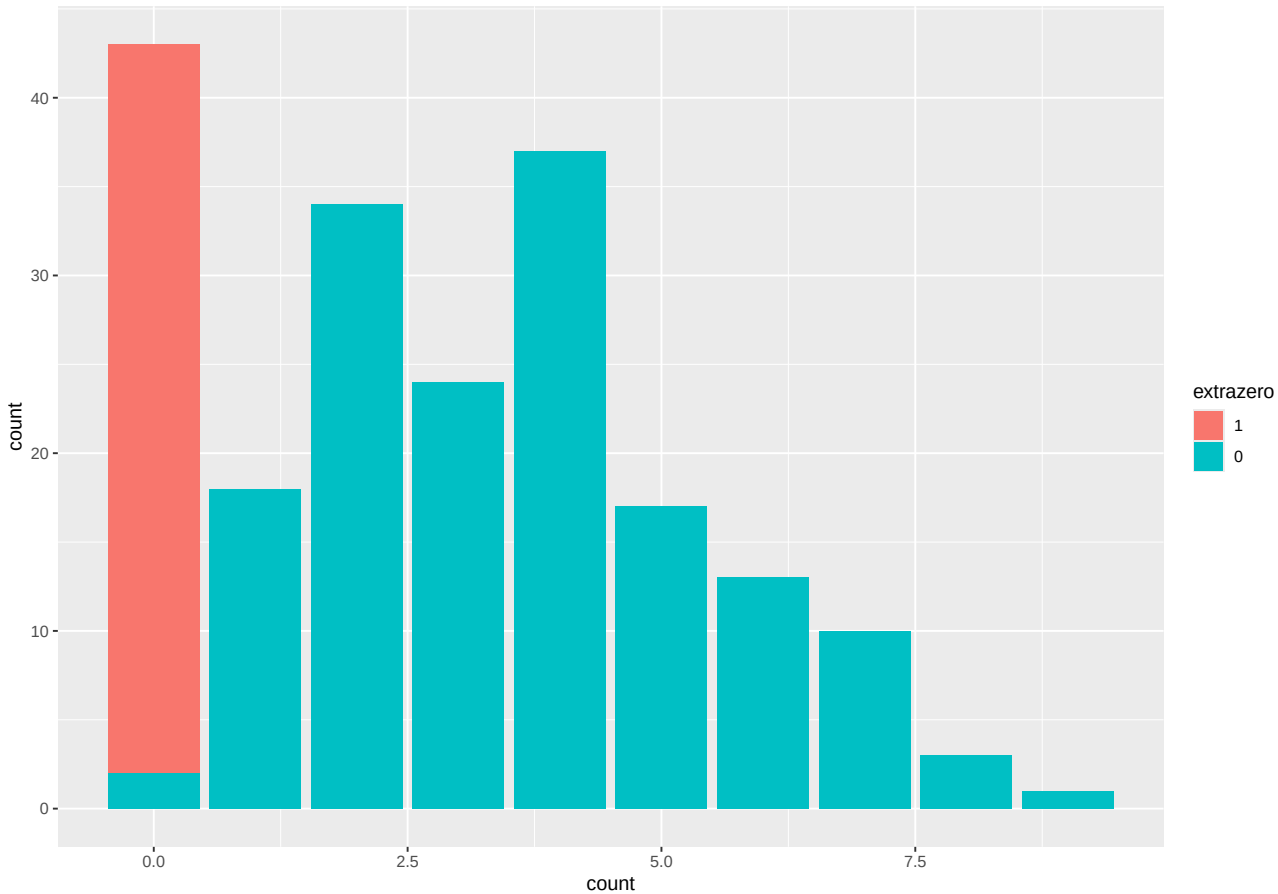
Zero-inflated Poisson

If probability of zero inflation is pz, then assume that response data come from a Poisson with probability $(1 - pz)$.

e.g. river herring counts at a weir

```
set.seed(394)
pz <- 0.2
herring <- data.frame(zi=rbinom(200, size=1, prob=pz))
herring <- herring %>%
  mutate(count = ifelse(zi==1,0,rpois(200,3.5)))
herring$extrazero <- factor(herring$zi)
herring$extrazero <- factor(herring$extrazero,
                            levels(herring$extrazero)[c(2,1)])
g <- ggplot(herring,aes(count))
g + geom_bar()
g + geom_bar(aes(fill=extrazero)) # + scale_x_discrete(bre
```


40
30
20
10
0
count
0.0
2.5
5.0
7.5
count


40
30
20
10
0
count
0.0
2.5
5.0
7.5
count
extrazero
1
0

Poisson GLM on herring

```
herring_poi <- glm(count~1, data = herring,
                   family = poisson)
poi_res <- resid(herring_poi, type="pearson")
disp <- sum(poi_res^2)/herring_poi$df.residual
disp
```

```
## [1] 1.725551
```

```
summary(herring_poi)
```

```
##
## Call:
## glm(formula = count ~ 1, family = poisson, data = herring)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.05082    0.04181   25.13   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 425.58  on 199  degrees of freedom
## Residual deviance: 425.58  on 199  degrees of freedom
## AIC: 905.74
##
```

Zero-inflated Models in R

Zero-inflated count data models can be fit using `zeroinfl()` function from the `pscl` package. Here we use `glmmTMB()` from `glmmTMB` package.

```
#install.packages('pscl')
library(pscl)
library(glmmTMB)
```

```
herring_zip <- glmmTMB(count~1, zi=~1, data = herring, family = poisson
#coefficients
coefs <- tidy(herring_zip)
exp(coefs$estimate[1])
```

```
## [1] 3.537324
```

```
plogis(coefs$estimate[2])
```

```
## [1] 0.1914793
```

```
#dispersion
zip_res <- resid(herring_zip, type="pearson")
disp <- sum(zip_res^2)/df.residual(herring_zip)
disp
```

```
## [1] 1.385981
```

```
summary(herring_zip)
```

```r
summary(herring_zip)
```

```
##  Family: poisson  ( log )
## Formula:          count ~ 1
## Zero inflation:         ~1
## Data: herring
##
##      AIC      BIC   logLik deviance df.resid
##    830.2    836.8   -413.1    826.2      198
##
##
## Conditional model:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.26337    0.04422   28.57   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Zero-inflation model:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.4404     0.1948  -7.394 1.42e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Zero-Altered (hurdle / delta) models

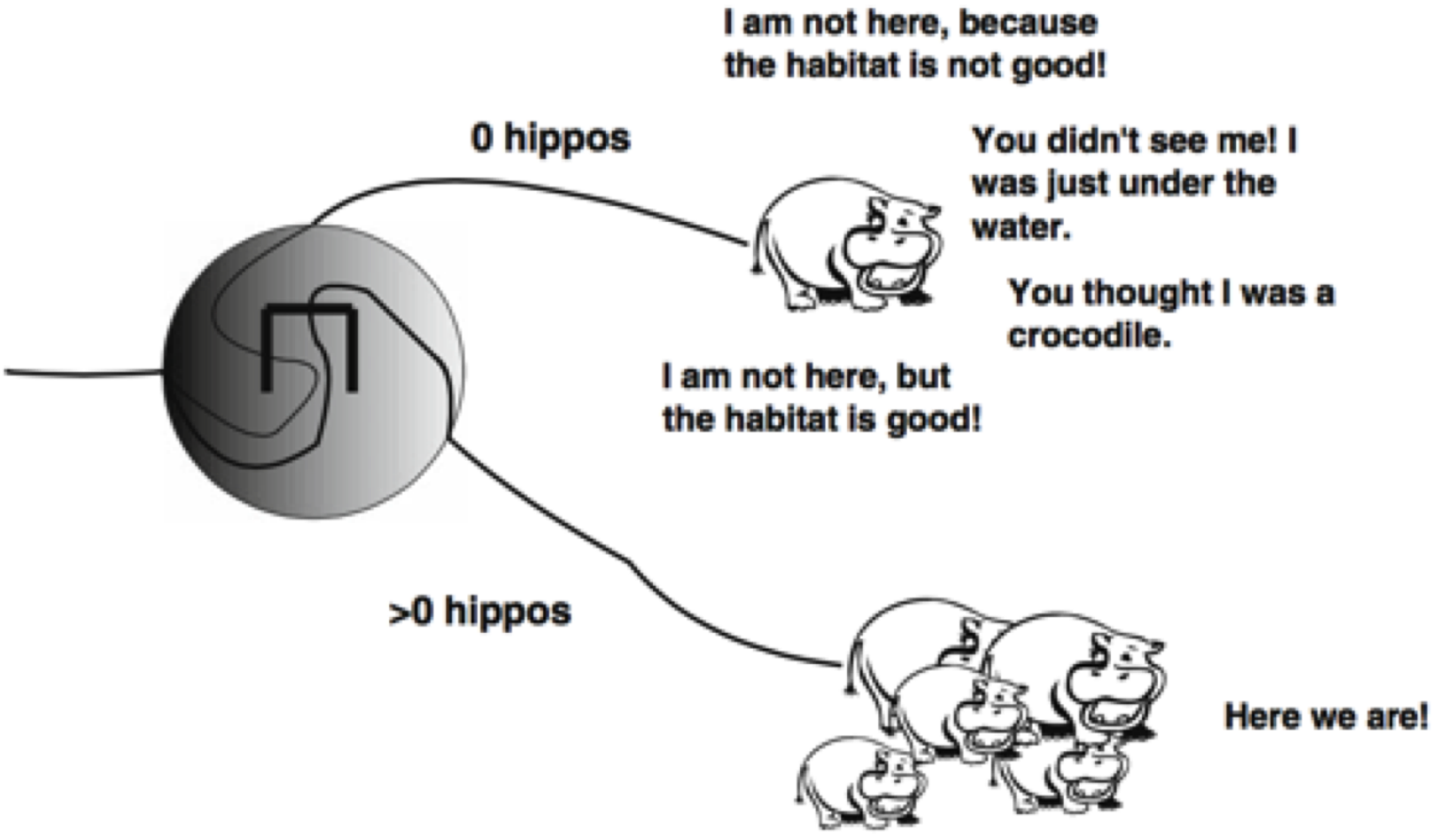
Rather than estimate a mixture distribution, we could fit two separate models:

1. A model for the zeroes (e.g. bernoulli)
2. A model for the non-zeroes (e.g. truncated Poisson)

Commonly referred to as *hurdle* or *delta* models

Unlike in the zero-inflated case, there is now only 1 type of zero.

Zero-Altered (hurdle / delta) models



Zero-Altered Poisson - river herring

```
herring_zap <- glmmTMB(count~1, zi=~1, data = herring,
    family = truncated_poisson)
summary(herring_zap)
```

```
##  Family: truncated_poisson  ( log )
## Formula:          count ~ 1
## Zero inflation:         ~1
## Data: herring
##
##      AIC      BIC   logLik deviance df.resid
##    830.2    836.8   -413.1    826.2      198
##
##
## Conditional model:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.26337    0.04422   28.57   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Zero-inflation model:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.2950     0.1721  -7.524 5.31e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Zero-Altered Poisson - river herring

```
coefs <- tidy(herring_zap)
exp(coefs$estimate[1])
```

```
## [1] 3.537324
```

```
plogis(coefs$estimate[2])
```

```
## [1] 0.215
```

```
#dispersion
zap_res <- resid(herring_zap, type="pearson")
disp <- sum(zap_res^2)/df.residual(herring_zap)
disp
```

```
## [1] 1.481375
```

Positive continuous data with zeroes

A common case is to have positive continuous data that also has zeroes.

e.g. fish trawl survey data where have a positive continuous measure of catch per unit effort but some tows don't have any of a given species.

We can use a hurdle model, the difference is that a truncated distribution is not needed if we use a continuous distribution that only takes positive values (e.g. Gamma or linear model on log scale).

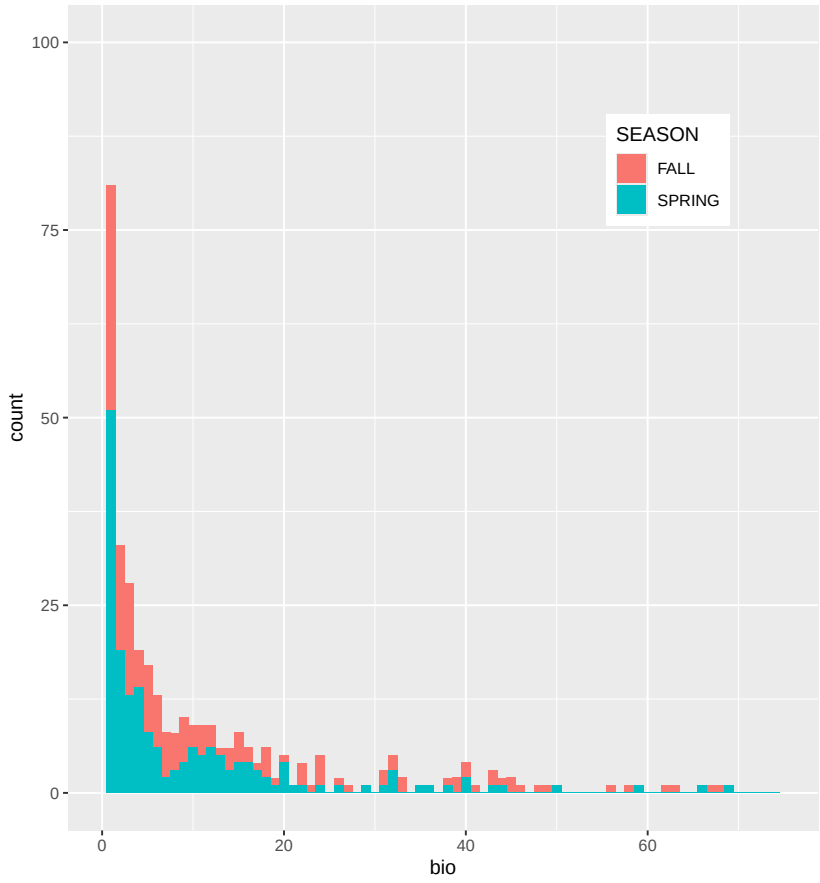
This is commonly referred to as a *delta-GLM* approach.

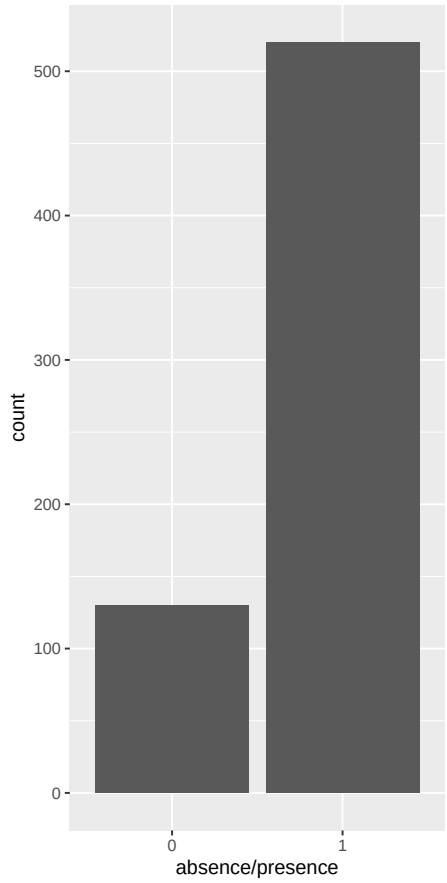
Example delta-GLM, silver hake

2014 NMFS survey data for silver hake.

```
#Read in the survey data
hakedat <- read_csv("../data/hake.csv") |>
 mutate(presence=ifelse(bio>0,1,0))
```

```
## Rows: 650 Columns: 6
## -- Column specification --------------------------------------
## Delimiter: ","
## chr (1): SEASON
## dbl (5): LAT, LON, DEPTH, svp, bio
##
## i Use `spec()` to retrieve the full column specification for this da
## i Specify the column types or set `show_col_types = FALSE` to quiet
```


SEASON
FALL
SPRING
count
bio
0
25
50
75
100
0
20
40
60


count
absence/presence
0
100
200
300
400
500
0
1

fit the delta-GLM

```
m_bin <- glmmTMB(bio~1, zi=~1, data = hakedat,
                 family = ziGamma)
```

```
## Warning in (function (start, objective, gradient = NULL, hessian = N
## NA/NaN function evaluation
## Warning in (function (start, objective, gradient = NULL, hessian = N
## NA/NaN function evaluation
```

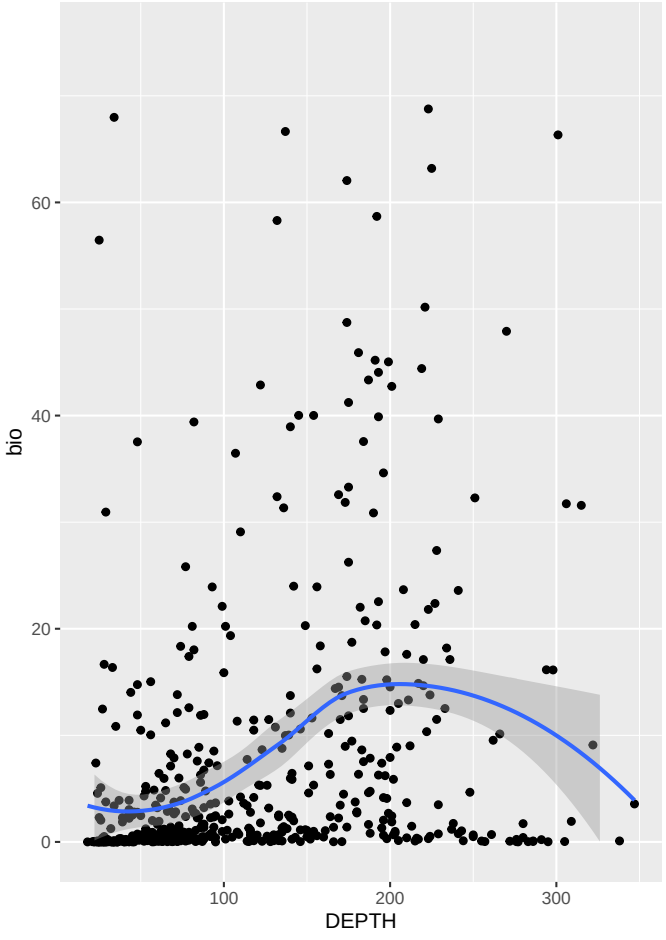
```
tidy(m_bin)
```

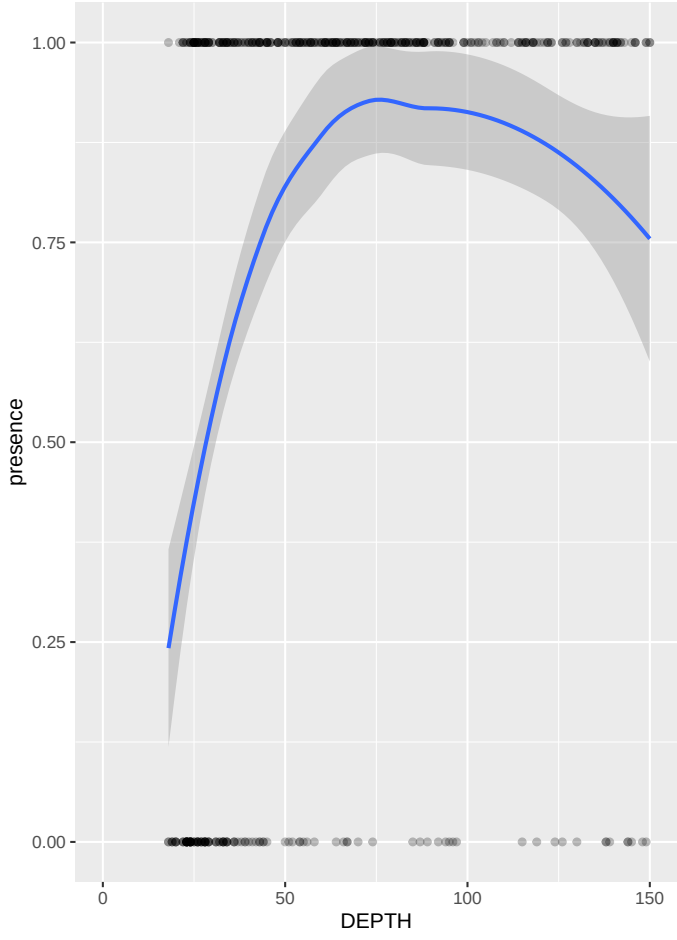
```
## # A tibble: 2 x 7
##   effect component term        estimate std.error statistic  p.value
##   <chr>  <chr>     <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 fixed  cond      (Intercept)    0.124   0.00925      13.4 7.02e-41
## 2 fixed  zi        (Intercept)   -1.39    0.0981      -14.1 2.23e-45
```

```
#model predictions
augment(m_bin)
```

```
## # A tibble: 650 x 4
##       bio .fitted .se.fit .resid
##     <dbl>   <dbl>   <dbl>  <dbl>
##  1 0         0.124 0.00925  -6.46
##  2 0.0929    0.124 0.00925  -6.37
##  3 0         0.124 0.00925  -6.46
##  4 0         0.124 0.00925  -6.46
```

adding depth as a predictor variable





```r
depth_mod <- glmmTMB(bio~DEPTH, zi=~DEPTH, data = hakedat,
          family= ziGamma)
#depth_mod <- glmmTMB(bio~s(DEPTH), zi=~s(DEPTH), data = hakedat,
#           family= lognormal())
summary(depth_mod)
```

```
##  Family: Gamma  ( inverse )
## Formula:          bio ~ DEPTH
## Zero inflation:      ~DEPTH
## Data: hakedat
##
##      AIC      BIC   logLik deviance df.resid
##   3093.9   3116.2  -1541.9   3083.9      645
##
##
## Dispersion estimate for Gamma family (sigma^2): 2.71
##
## Conditional model:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.314e-01  1.804e-02   12.82   <2e-16 ***
## DEPTH       -6.494e-04  6.253e-05  -10.39   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Zero-inflation model:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.064913   0.179207  -0.362    0.717
```

```
tidy(depth_mod)
```

```
## # A tibble: 4 x 7
##   effect component term          estimate std.error statistic  p.valu
##   <chr>  <chr>     <chr>            <dbl>     <dbl>     <dbl>    <dbl
## 1 fixed  cond      (Intercept)  0.231    0.0180       12.8   1.22e-3
## 2 fixed  cond      DEPTH       -0.000649 0.0000625   -10.4   2.86e-2
## 3 fixed  zi        (Intercept) -0.0649   0.179        -0.362 7.17e-
## 4 fixed  zi        DEPTH       -0.0160   0.00224      -7.14  9.53e-1
```

```
augment(depth_mod)
```

```
## # A tibble: 650 x 5
##       bio DEPTH .fitted .se.fit .resid
##     <dbl> <dbl>   <dbl>   <dbl>  <dbl>
##  1 0        229  0.0826     NaN -11.8
##  2 0.0929   338  0.0118     NaN -84.0
##  3 0         58  0.194      NaN  -3.77
##  4 0         36  0.208      NaN  -3.15
##  5 0        338  0.0118     NaN -84.1
##  6 0         24  0.216      NaN  -2.83
##  7 0         94  0.170      NaN  -4.86
##  8 0         64  0.190      NaN  -3.94
##  9 0         28  0.213      NaN  -2.93
## 10 0         34  0.209      NaN  -3.09
## # i 640 more rows
```

Summary

- Methods exist for fitting GLM-type models to data with large numbers of zeroes.
- These methods are preferable to transformations of the response variable.
- Zero-inflated models for count data use a mixture model, and define two types of zeroes.
- Zero-altered (hurdle) models only consider 1 type of zero, but can be applied to continuous positive data.
- Alternative modeling approaches include using a distribution that is very-overdispersed (e.g. Tweedie)

Reading

Bolker et al. 2012. Owls example: a zero-inflated, generalized linear mixed model for count data.
https://groups.nceas.ucsb.edu/non-linear-modeling/projects/owls/WRITEUP/owls.pdf

Zuur, A. F. S., A. A. Ieno. 2012. Zero Inflated Models and Generalized Linear Mixed Models with R.

Zuur A.F., Ieno E. N. 2016. A Beginner's Guide to Zero-Inflated Models with R.

Next Time...

3/25: Zero-Inflated Models

3/26: UMass Marine Science Symposium

3/27: Spatial GLMMs

Another example, Owlet sibling calls