Zero inflated Poisson distribution

For Master 2 Math SV

January 2026

1. The data

We study the abundance of fish species at $n = 89$ sites in the Barents Sea (Fossheim, Nilssen, and Aschan (2006)). The data are available in the file BarentsFish.csv where the first 4 columns correspond to four environmental covariates covariates (latitude, longitude, depth, temperature) and the next 30 columns are the abundances of 30 species.

```r
dataCodBarents <- read.table('BarentsFish.csv', sep=';', header=TRUE)
Covariates <- as.matrix(dataCodBarents[, (1:4)])
Counts <- dataCodBarents[, 5:ncol(dataCodBarents)]
j <- 21 # We focus on Species Tr_es
Abundance <- Counts[, j]; Presence <- (Abundance>0)
Cod_Tr_es<- as.data.frame(Covariates)
Cod_Tr_es$Abundance <-Abundance
Cod_Tr_es$Presence <- Presence


j <- 20 # We focus on Species Se_ma
Abundance <- Counts[, j]; Presence <- (Abundance>0)
Cod_Se_ma<- as.data.frame(Covariates)
Cod_Se_ma$Abundance <-Abundance
Cod_Se_ma$Presence <- Presence
```

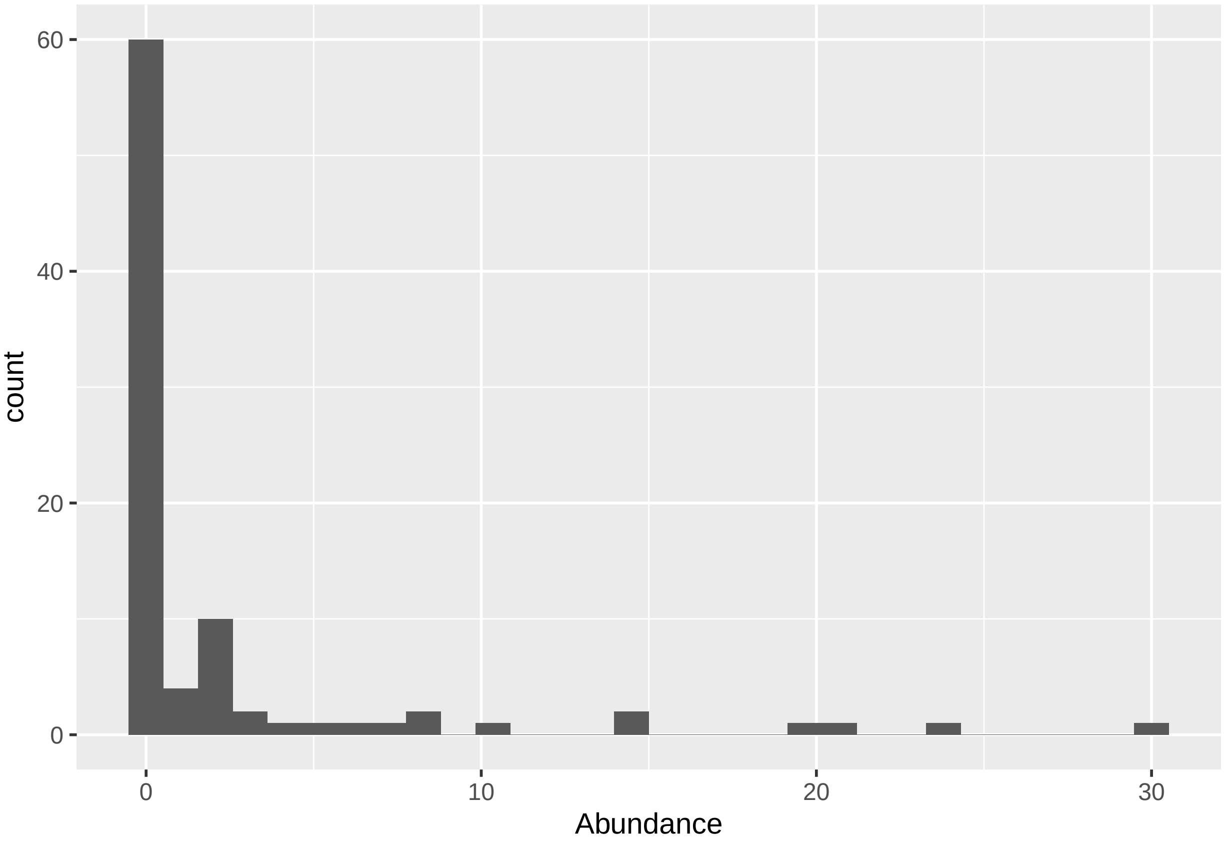
In the following, we will consider only one fish species, for example the 21th (’Tr_es = Trisopterus Esmarkii = Tacaud norvégien) or the 20th (“Se_ma”= Sebastes Marinus) and we will note $1 \leq i \leq n$.

$$Y_i = \text{ abundance of golden redfish in station i.}$$

1. Explore the data with standard tools (means, histograms…) for both species

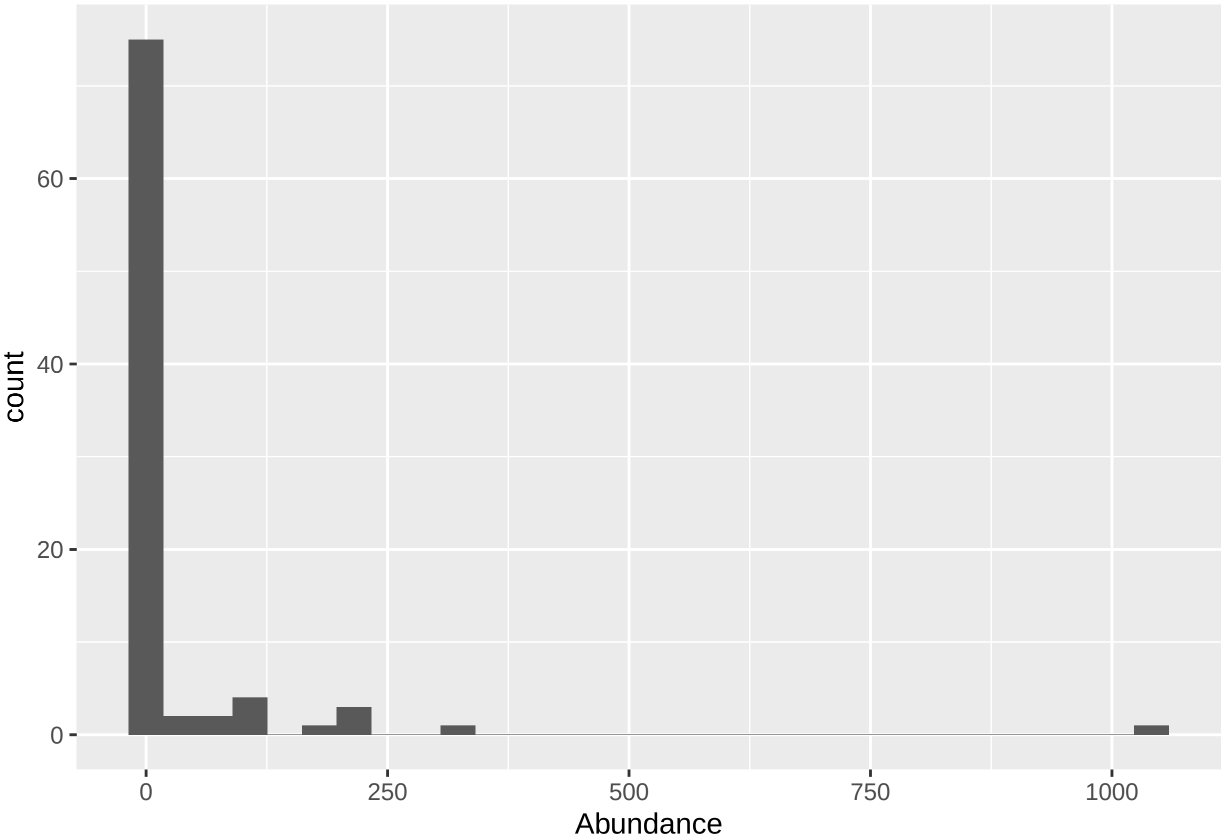
```r
library(ggplot2)
ggplot(Cod_Se_ma,aes(x=Abundance))+geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value `binwidth`.
```



```
ggplot(Cod_Tr_es,aes(x=Abundance))+geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value `binwidth`.
```



Observe an over-representation of null values. We propose to modelize this over inflation of 0.

2. Zero-inflated Poisson model

We propose to consider the following Zero Inflation Poisson distribution (ZIP) Let Z_i be a latent variable such that

$$Z_i \sim_{i.i.d} \mathcal{B}ern(\pi)$$

Then

$$Y_i | Z_i \sim (1 - Z_i)\delta_{\{0\}} + Z_i \mathcal{P}(\mu_i) \quad (1)$$

where $\mathcal{P}$ is the Poisson distribution. Z_i represents the presence of the species.

2. Write the marginal distribution of Y_i
3. Derive $\mathbb{E}[Y_i]$ and $P(Y_i = 0)$
4. Write the complete log likelihood $\log p_\theta(\mathbf{Y}, \mathbf{Z})$ of the model where $\theta = (\pi, \mu)$.

We propose to maximize likelihood with respect to the parameters using the EM algorithm

5. Write the corresponding E-step.
6. Write the corresponding M-step.
7. Suggest an initial value for the parameter θ.
8. Code the EM algorithm.
    a. Test you algorithm on simulated data. Check that the likelihood increases at each iteration of the EM

b. Test you algorithm on the real data. What happens?

3. ZIP with covariates

We now consider a model similar to ZIP but taking into account the environmental covariates. We note x_i the vector comprising these covariates for the site i:

$$x_i = [1, \text{latitude}_i, \text{longitude}_i, \text{depth}_i, \text{temperature}_i].$$

We therefore pose : $(Z_i)_{1,\leq i \leq n}$ independent,$(Y_i|Z_i)_{1,\leq i \leq n}$ independent and

$$\begin{array}{rclll} Z_i & \sim & \mathcal{B}\text{ern}(\pi_i) & \text{with} \quad \log\left(\frac{\pi_i}{1-\pi_i}\right) & = & x_i^T\alpha \\ Y_i|Z_i & \sim & (1-Z_i)\delta_{\{0\}} + Z_i\mathcal{P}(\mu_i) & \text{with} \quad \log\mu_i & = & x_i^T\beta \end{array} \tag{2}$$

The vectors α and β contain the regression coefficients to predict absence and abon- dance conditional on the presence of the species at each site.

9. Write the full log likelihood $p_\theta(\mathbf{Y}, \mathbf{Z})$ of this new model as a function of the parameter $\theta = (\alpha, \beta)$.
10. Write the E-step.
11. Write the M-step. Is it explicit?
12. Propose an initial value for the parameter θ.

References

Fossheim, Maria, Einar M. Nilssen, and Michaela Aschan. 2006. “Fish Assemblages in the Barents Sea.” *Marine Biology Research* 2 (4): 260–69. https://doi.org/10.1080/17451000600815698.