How to deal with count data?
Pollinator deception

Figure 1. *Orchid Mantis and the orchid that mimics*

Many models for biological data do not have constant variance nor are normally distributed. Generalized linear models (GLMs) can evaluate hypothesis for some of these data. A generalized linear model is defined by three properties: the linear predictor, the link function and the error structure. We strongly encourage you to learn more about these models. The estimated values are obtained with a transformation of the data calculated with a linear predictor. The link function relates the values of the response variable to the linear predictor. These models allow you to specify different variance distributions. Count data are integers, bound to an inferior limit, since no count can be less than zero, also they often have many zeroes and their variance frequently increases with the mean (Crawley 2007). The probability distribution Poisson is very useful to describe count data. It estimates the probability of obtaining a count x when the mean count per unit is λ (Crawley 2007), and it works fine when the mean is fairly equal to its variance. When the variance in counts is much greater than the mean, the data are better described by the Negative Binomial Distribution (Crawley 2007). The link for these types of models is the logarithmic link.

That some mantis mimic flowers to attract pollinators as prey has been a favorite hypothesis since first proposed by Charles Darwin. However, it has rarely being tested. A recent experiment was designed to formally evaluate its support. Hanlon et al. (2014) designed and implemented an experiment to compare if, as predicted, the Malaysian orchid mantis *Hymenopus* coronatus are indistinguishable from the sympatric flowers that are visited by their hymenopteran prey (Figure 1). In each trial, a live mantis was placed on top of one stick, a live *Asystasia intrusa* flower was tethered to another, and a third stick was left bare as a control stimulus. They were observed simultaneously for an hour in different sites and visiting insects were tallied for a total of 30 observations. The authors kindly provided these data that we evaluate below. We read the data, calculate the average number of counts per type of stimulus, and plot their histograms (Figure 2).

```
rm(list=ls())
library(lattice)

cd <- read.table("mantis.txt", header=T)
names(cd)

## calculate means
mean(cd$total[type=="Total_Mantid"])
mean(cd$total[type=="Total_Flower"])
mean(cd$total[type=="zTotal_Control"])
```

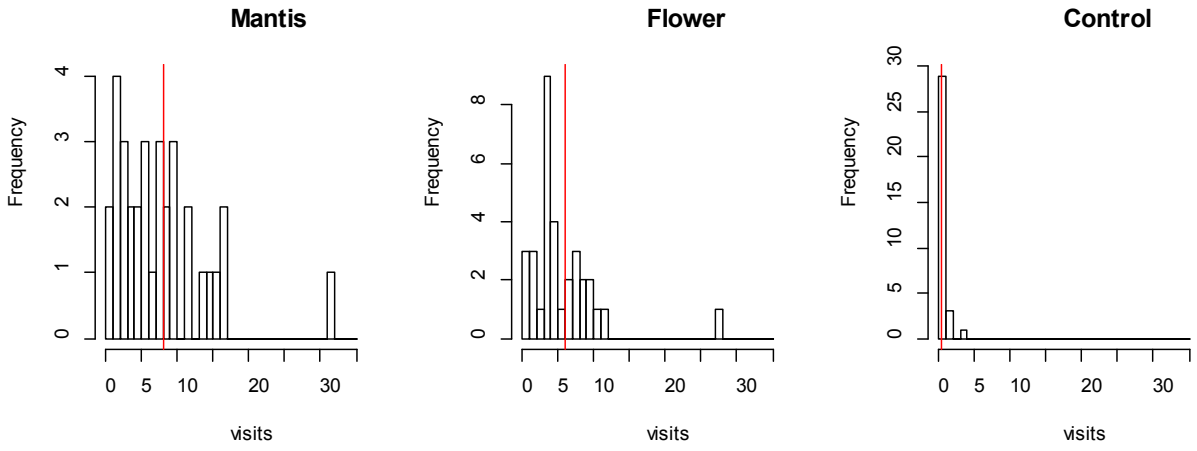

Figure 2. *Histograms of the data, mean count in red*

Notice that the data around the mean are not normally distributed and their spread increases with the mean. Consequently, we evaluate two GLMs for these data. For the first we use Poisson distribution, we then compensate for over-dispersion and evaluate a GLM with negative binomial distribution (Table 1; Zuur et al. 2015).

(1.1) For the ***Poisson*** the likelihood distribution is given by:

$$\boxed{\begin{aligned} &Number\_of\_Insects_i \sim P(\mu_i) \\ &E(N\_insects_i) = \mathrm{var}(N\_insects_i) = \mu_i \end{aligned}}$$

(1.2) The link function is the log of μ:

$$\boxed{\log(\mu_i) = \eta_i}$$

(1.3) We define the predictor function η is a function of the covariates:

$$\boxed{\eta = \beta_i[treatment]_i}$$

(2.1) For the ***Negative binomial*** the likelihood distribution is given by:

$$Number\_of\_Insects_i \sim NB(\mu_i, k)$$
$$E(N\_insects_i) = \mu_i$$
$$var(N\_insects_i) = \mu_i + \mu_i^2 / k$$
$$var(N\_insects_i) = \mu_i + \alpha \times \mu_i^2$$

(2.2) The link function is the log of μ:

$$\log(\mu_i) = \eta_i$$

(2.3) The predictor function η is a function of the covariates:

$$\eta = \beta_0 + \beta_1 [treatment]_i$$

```
## Poisson model
model1 <- map2stan(
  alist(
    total ~ dpois(lambda),
    log(lambda) <- t[tid],
    t[tid] ~ dnorm(0,1)
    ),
  data = cd, chains =3
)
> precis(model1,digits=1,depth=2)
     Mean StdDev lower 0.89 upper 0.89 n_eff Rhat
t[1]  2.1    0.1        2.0        2.2  2860    1
t[2]  1.8    0.1        1.7        1.9  2709    1
t[3] -0.8    0.2       -1.2       -0.4  2797    1
```

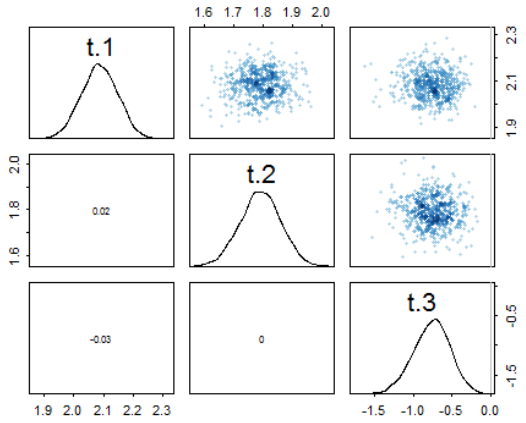

Figure 3. The plot on top shows distribution and the correlation of the parameters of the Poisson model.

```
## Negative binomial model

model2 <- map2stan(
  alist(
    total ~ dgampois(pbar,scale),
    log(pbar) <- a + b*man + c*con,
    a ~ dnorm(0,10),
    b ~ dnorm(0,2),
    c ~ dnorm(0,2),
    scale ~ dcauchy(0,2)
  ),
  data = cd,
  constraints =list(theta ="lower=0"),
  start =list(a=1,b=1,c=1,scale=3),
  iter=4000, warmup=1000, chains =3
  )

precis(model2,digits=1,depth=2)
      Mean StdDev lower 0.89 upper 0.89 n_eff Rhat
a      1.8    0.1        1.6        2.0  5288    1
b      0.3    0.2        0.0        0.6  5099    1
c     -2.6    0.3       -3.1       -2.1  6353    1
scale  2.4    0.6        1.5        3.2  7632    1
```

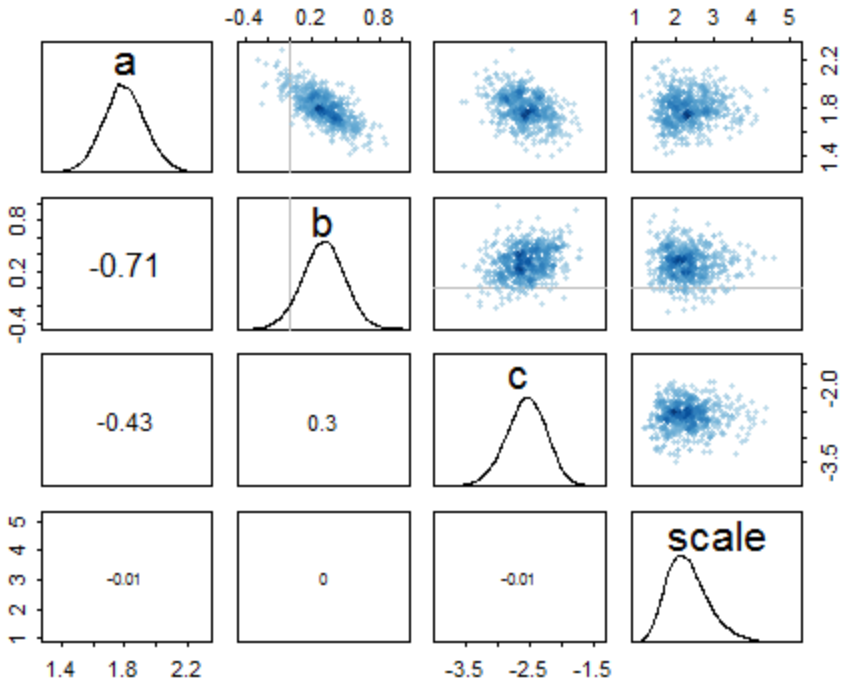

Figure 4. The plot on top shows distribution and the correlation of the parameters of the Negative binomial model.

```
compare(model1,model2)

        WAIC pWAIC dWAIC weight    SE   dSE
model2 456.0   5.9   0.0      1 29.76    NA
model1 561.6  10.3 105.6      0 70.12 47.47
```

Table 1. Parameters and their standard errors after the two GLM models

	Poisson		Negative binomial	
Coefficient	Estimate	Std. Error	Estimate	Std. Error (coeffs.)
Intercept (Flower)	1.8	0.1	1.8	0.1
Mantis	2.1	0.1	1.8 + 0.3 = 2.1	0.2
Control	-0.8	0.2	1.8 - 2.6 = -0.8	0.3
Dispersion parameter (scale)			2.4	0.6
Dispersion statistic		3.66		1.29

Notice that we use different implementations for these models. Based on the model with Poisson errors, we could conclude that visitation of mantis (mean = exp (2.1) = 8.12) was significantly higher than that for flowers (mean = 6.06). However, the negative binomial is more likely to explain the data and the fact that the parameter variances were larger in the negative binomial model indicates over-dispersion (extra, unexplained variation in the response; Crawley 2007). A more precise way to evaluate for over -dispersion is to calculate the dispersion statistic. Do not confound the dispersion *statistic* (see below) with the dispersion *parameter* α (see definition of variance of Negative binomial above; Zuur et al. 2015).

$$dispersal\ statsitc\ = \frac{x^2}{residual\ degrees\ of\ freedom}$$

$$x^2 = \sum_{i=1}^{N} \frac{(Y_i - E(Y_i))^2}{var(Y_i)}$$

We found that the dispersal statistic for the Poisson model is 3.66 and the one for the Negative binomial is 1.29. Simulations indicate that a Poisson model well fitted should have a dispersal statistic close to 1.0. The negative binomial model, which is more informative than the one with Poisson errors (based on WAIC: 456 vs 562), confirms this interpretation. Both models consistently provide evidence that visitation rates for the procedure control were significantly lower than the other two stimuli (mean = 0.45). Once again Darwin was correct!

NOTE: all the materials for this demo can be found at:

https://sciences.ucf.edu/biology/d4lab/methods-2/

References

Crawley, M.J. 2007. The R Book. Wiley.

Hanlon, J.O, G. L. Holwell and M. Herberstein. 2014. Pollinator deception in the Orchid Mantis. American Naturalist 183: data: http://dx.doi.org/10.5061/dryad.g665r.

McElreath, R.M. 2016. Statistical Rethinking: a Bayesian course with examples in R and Stan. Chapman and Hall.

Photo credits: https://photoplusbyritasim.wordpress.com/tag/purple/page/7/

Zuur, A, J.M. Hilbe and E N. Leno. 2015. A beginner's guide to GLM and GLMM with R. Highland Statistics, Ltd.