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 normally distributed errors. 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 values calculated with the linear predictor. The link function relates the values of the response variable to the linear predictor. These models allow you to specify different error distributions. Count data are integers, bounden in their inferior limit, since no count can be less than zero, they also 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 its mean is fairly equal to its variance. When the variance in counts is much greater than the mean, the data are better described by 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. However, it is only recently that an experiment was designed to test 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"])
```

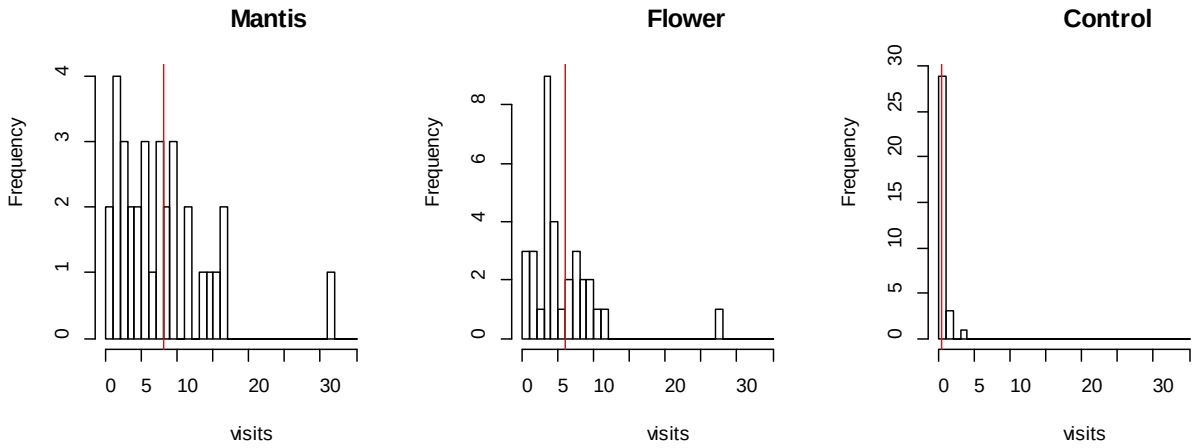

Figure 2. *Histograms of the data, mean count in red*

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

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

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

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

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

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

$$\boxed{\eta = \beta_0 + \beta_1[treatment]_i}$$

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

$$\begin{aligned} &Number\_of\_Insects_i \sim NB(\mu_i, k) \\ &E(N\_insects_i) = \mu_i \\ &var(N\_insects_i) = \mu_i + \mu_i^2 / k \\ &var(N\_insects_i) = \mu_i + \alpha \times \mu_i^2 \end{aligned}$$

(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:

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

```
## models with different families for errors

model1 <- glm(total ~ type, family = poisson)
summary(model1)
model2 <- glm(total ~ type, family = quasipoisson)
summary(model2)
model3 <- glm.nb(formula = total ~ type, init.theta = 0.1, link = log)
summary(model3)

AIC(model1,model3)
```

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

	Poisson		Quasi-Poisson		Negative binomial	
Coefficient	Estimate	Std. Error	Estimate	Std. Error	Estimate	Std. Error
Intercept (Flower)	1.802	0.071	1.802	0.135	1.802	0.133
Mantis (difference)	0.293	0.093	0.293	0.179	0.293	0.184
Control (difference)	-2.590	0.268	-2.590	0.512	-2.590	0.311
Residual Variance		296.44		296.44		103.44
Degrees of freedom		96		96		96
Dispersion parameter		1		3.664		2.397
Dispersion statistic		3.664				1.292

Based on the model with Poisson errors, we could conclude that visitation of mantis (mean=8.12) was significantly higher than that for flowers (mean=6.06). However, the fact that the residual variance was larger than the residual degrees of freedom 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 of 1.0. We can conclude that the Negative binomial model is superior. We compensate for the over-dispersion by refitting the model using quasi-Poisson rather than Poisson errors. This compensation increased the *p* value to 0.105, providing no evidence to support this difference. Zuur et al. (2015; page 21) cautions that "quasi-Poisson distribution only modifies the standard deviation of the parameters in the Poisson GLM and not the parameter estimates and argues that therefore quasi-Poisson is less useful as a solution for over dispersion". The negative binomial model, which is more informative than the one with Poisson errors (based on AIC: 451.7 vs 554.08), confirms this interpretation. All models consistently provide evidence that visitation rates for the procedure control were significantly lower than the other two stimulus (mean = 0.454). The residuals of the model with negative binomial errors had a tighter distribution (Figure 3).

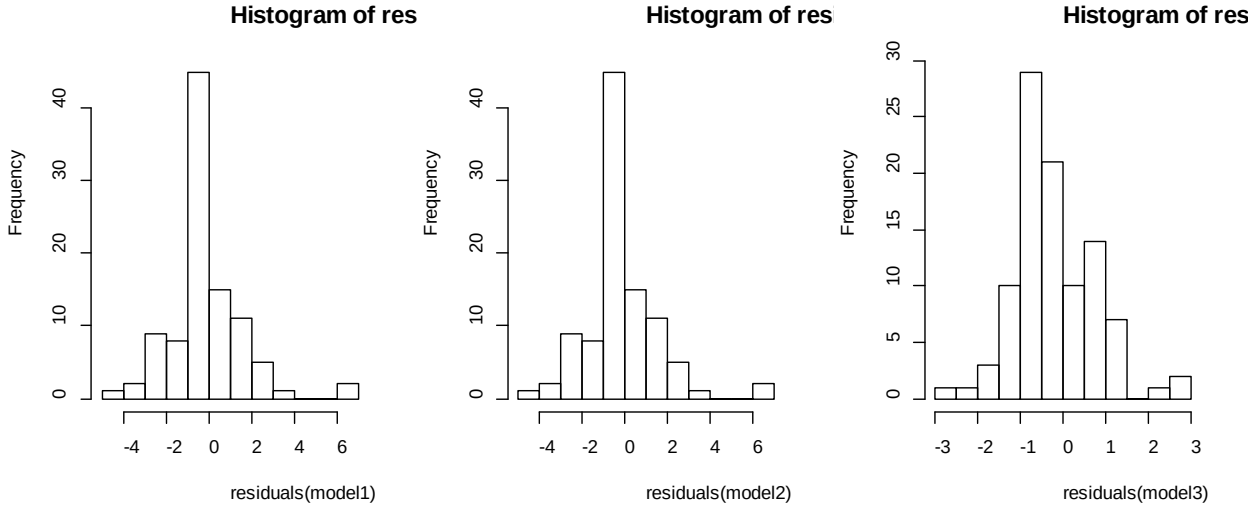

Figure 3. *Histograms of the residuals of the three models*

The code for a Bayesian model with no informative priors and Poisson distribution and that with Negative Binomial distribution of these data gives commensurate results and it is included below for your information. The posterior distributions of the parameters are shown in Table 2.

```
### Bayesian version ###

##########################################
### Bayesian version ###
library(rjags)

n_obs <- length(cd$type)
x <-c(rep(2,length(cd$type[cd$type=="Total_Mantid"])),
rep(1,length(cd$type[cd$type=="Total_Flower"])),
rep(3,length(cd$type[cd$type=="zTotal_Control"])))
y <- cd$total

#Write model
## a) Bayesian analysis with uninformed priors ##

model =  "mantis_jags Poisson.R"

X <- model.matrix(~~ type,data=cd)
K= ncol(X)

#Bundle data
win.data <- list(Y=y,X=X, K=K, n=n_obs)


# Inits function
inits <- function(){
  list(beta = rnorm(K,0,0.1))}


# Parameters to estimate
params <- c("beta","Fit","Fitnew")

# MCMC settings, start Gibbs sampler and plot results and diagnostics
jm=jags.model(model,data=win.data,inits=inits,n.chains=3,n.adapt=5000)
update(jm, n.iter=10000)
zc=coda.samples(jm,variable.names=params,n.iter=10000)
gelman.diag(zc)
plot(zc)
summary(zc)
mean(zc[[1]][,1] > zc[[1]][,2])

##############################################

#Write model
## a) Bayesian analysis with uninformed priors ##

model =  "mantis_jags Neg Binom.R"

#Bundle data
win.data <- list(Y=y,X=X,K=K, n=n_obs)

# Inits function
inits <- function(){
  list(beta = rnorm(K,0,0.1))}


# Parameters to estimate
params <- c("beta","Fit","Fitnew","size")

# MCMC settings, start Gibbs sampler and plot results and diagnostics
jm=jags.model(model,data=win.data,inits=inits,n.chains=3,n.adapt=5000)
update(jm, n.iter=10000)
zc=coda.samples(jm,variable.names=params,n.iter=10000)
```

```
gelman.diag(zc)
plot(zc)
summary(zc)
mean(zc[[1]][,1] > zc[[1]][,2])
```

Table 2. Parameters and their standard errors after the three GLM models

Bayesian	Poisson		Negative binomial	
Coefficient	Estimate	Std. Error	Estimate	Std. Error
Intercept (Flower)	1.80	0.071	1.81	0.133
Mantis (difference)	0.293	0.093	0.293	0.184
Control (difference)	-2.59	0.267	-2.62	0.317
Size			2.507	
Bayesian P	1		0.8348	
Frequentist				
Intercept (Flower)	1.80	0.071	1.80	0.133
Mantis (difference)	0.293	0.093	0.293	0.184
Control (difference)	-2.59	0.268	-2.59	0.311
Size			2.397	

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.

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.