How to deal with count data?
Pollinator deception

Figure 1. Orchid Mantis and the orchid it mimics

Many models for biological data do not have constant variance or normally distributed errors, and generalized linear models (GLMs) can help us evaluate hypothesis for some of these data, so we strongly encourage you to learn more about them. A GLM is defined by three properties: the *linear predictor*, the *link function* and the *error structure*. 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. Counts are a good example of data that should not be analyzed with simple linear regressions. They are recorded as integers and are bounded in their inferior limit by zero; they also often have many zeroes so their variance frequently increases with the mean (Crawley 2007). The Poisson probability distribution is very useful to describe count data as 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 the variance. When the variance in counts is much greater than the mean (i.e. the data is *over-dispersed*), the data are better described by a negative binomial distribution (Crawley 2007). The link for both of these models is the logarithmic link.

Orchid mantises are hypothesized to mimic orchids in order to attract pollinators to consume as prey. O'Hanlon *et al.* (2014) designed and implemented an experiment to establish whether, as predicted, the Malaysian orchid mantis *Hymenopus coronatus* were undistinguishable from the sympatric flowers 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. The sticks were observed simultaneously for an hour in different sites and visiting insects were tallied for a total of 30 observations. Using the data kindly provided by the authors, we start off by calculating the average number of counts per type of stimulus (treatment), and plotting the corresponding histograms (Figure 2).

For this demo you will need the following files: `mantis.R`, `mantis_negbinom.R`, `mantis_Poisson.R`, and `mantis.txt` as well as the `JAGS` software and the `MASS` and `rjags` packages.

```
cd <- read.table("mantis.txt", header=T)
mean(cd$total[type=="Total_Mantid"])
mean(cd$total[type=="Total_Flower"])
mean(cd$total[type=="zTotal_Control"])
b = seq(0,35,1)
par( mfrow=c(1,3))
hist(cd$total[cd$type=="Total_Mantid"],breaks =b,xlab="visits", main="Mantis")
abline(v=8.121212, col="red")
hist(cd$total[cd$type=="Total_Flower"],breaks =b,xlab="visits", main="Flower")
abline(v=6.060606, col="red")
hist(cd$total[cd$type=="zTotal_Control"],breaks =b,xlab="visits", main="Control")
abline(v=0.4545455, col="red")
```

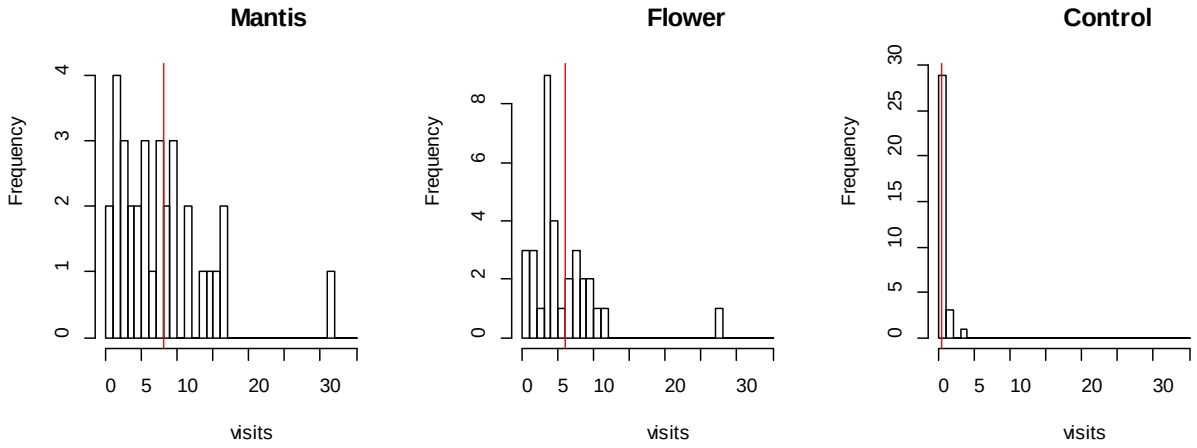

Figure 2. Histograms of the count data per treatment (mean number of visits in red)

Notice that the data are not normally distributed around the mean, and that their spread increases as the mean increases. Consequently, we will evaluate three GLMs for this data. First we use Poisson errors, then we 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) The ***Poisson*** error distribution is given by:

$$\begin{aligned} &Number\_of\_\text{Insects}_i \sim P(\mu_i) \\ &E(N\_\text{insects}_\text{i}) = \text{var}(N\_\text{insects}_\text{i}) = \mu_i \end{aligned}$$

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

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

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

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

(2.1) The ***Negative binomial*** error 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:

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

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

Table 1. Parameters and their standard errors for 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
Mantid (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
Residual 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 much larger than the residual degrees of freedom indicates over-dispersion (a lot of 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).

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

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

We found that the dispersal statistic for the Poisson model was 3.66 and the one for the negative binomial was 1.29. Simulations indicate that a well fitted Poisson model should have a dispersal statistic of 1.0. We can conclude that in this case the negative binomial model has a better fit. We also tried to compensate for the over-dispersion by refitting the model using quasi-Poisson rather than Poisson errors. This compensation increased the *p-value* of the differences between Flowers and Mantids from 0.0017 to 0.105, meaning model 2 no longer provides evidence to support a difference. Zuur *et al*. (2015; page 21) caution that the quasi-Poisson distribution only modifies the standard deviation of the parameters in the Poisson GLM and not the parameter estimates themselves, and therefore argue that quasi-Poisson is a less useful solution for over dispersion. Based on AIC scores (451.7 vs 554.1) the negative binomial model is more informative than the one with Poisson errors, and provides better evidence to support the hypothesis that insects cannot differentiate between flowers and mantises. Additionally, 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 around the mean (Figure 3).

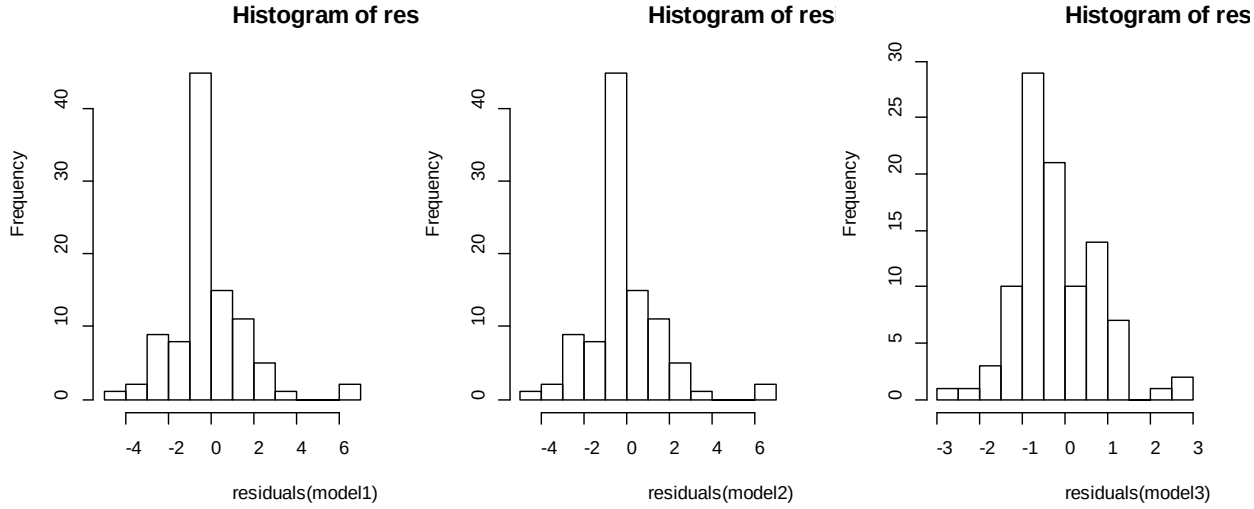

Figure 3. Histograms of the residuals for the three models

The code for a Bayesian model with uninformative priors, and both Poisson and negative binomial error distributions, is included in the R script for your information. Results are commensurate, as shown by the posterior distributions of the parameters in Table 2 (next page).

Table 2. Parameters and their standard errors for the two GLM models under the uninformed Bayesian and frequentist approaches.

	Poisson		**Negative binomial**	
Coefficient	Estimate	Std. Error	Estimate	Std. Error
Bayesian				
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	-

References

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

O'Hanlon, J.C., G.I. Holwell and M. Herberstein. 2014. Pollinator deception in the Orchid Mantis. *The American Naturalist* **183**:126-132.

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