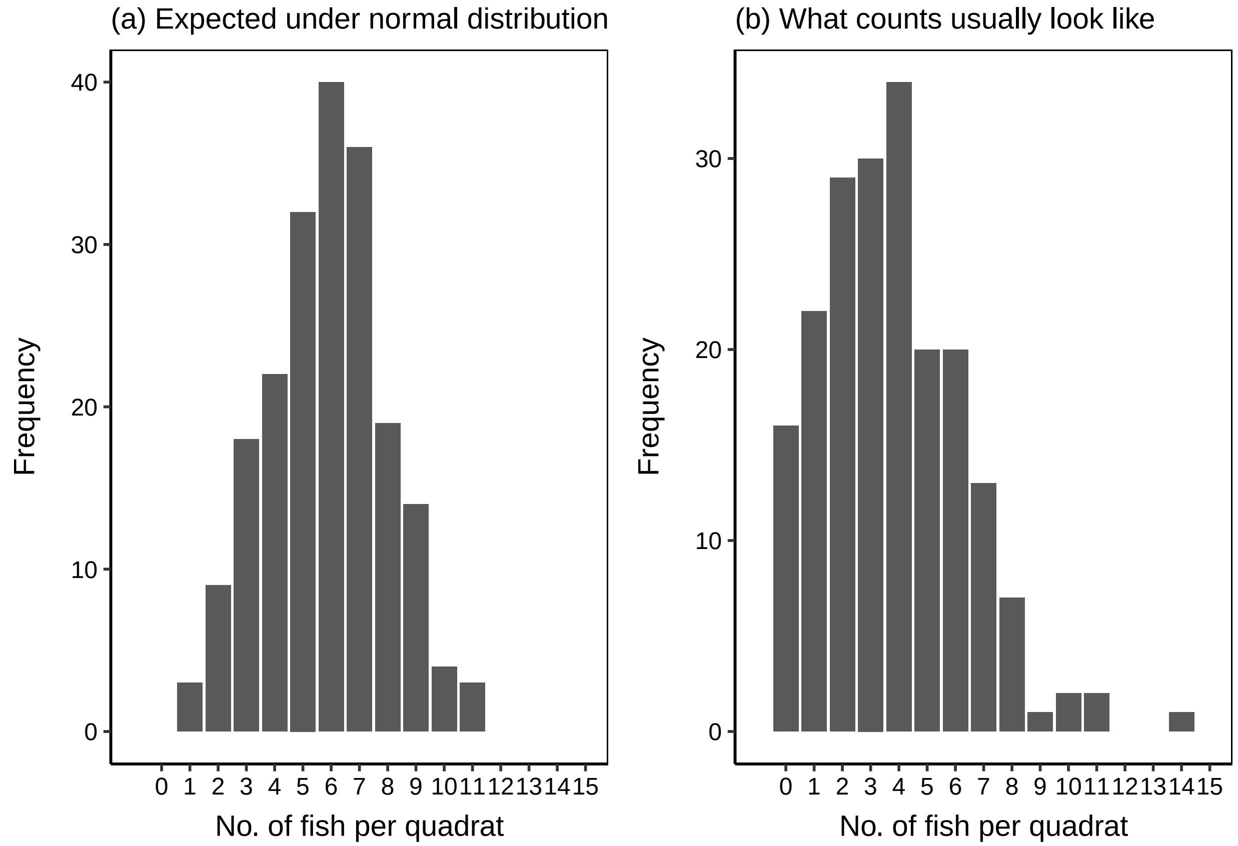
Tutorial #7: Count models - Poisson GLM

Guy F. Sutton

Centre for Biological Control, Rhodes University

7.1. Analysis of count (abundance) data

In the previous tutorials, we have been analysing continous response variables (i.e. numbers that can have a decimal place), assuming the data comes from a normal or Gaussian distribution. However, many ecological studies collect data where the variable of interest is a count or measurement of abundance (e.g. numbers that cannot have a decimal place). Count data typically DO NOT follow a normal distribution, and therefore typically should not be analysed assuming a normal distribution (e.g. Gaussian GLM, linear regression, ANOVA). Moreover, the assumption of equal variance is usually an issue when analysing count data, which again, makes standard linear regression/ANOVA-type models inappropriate. Count data often follow a strong mean-variance relationship, which simply means that the variance in your data increases with the mean…

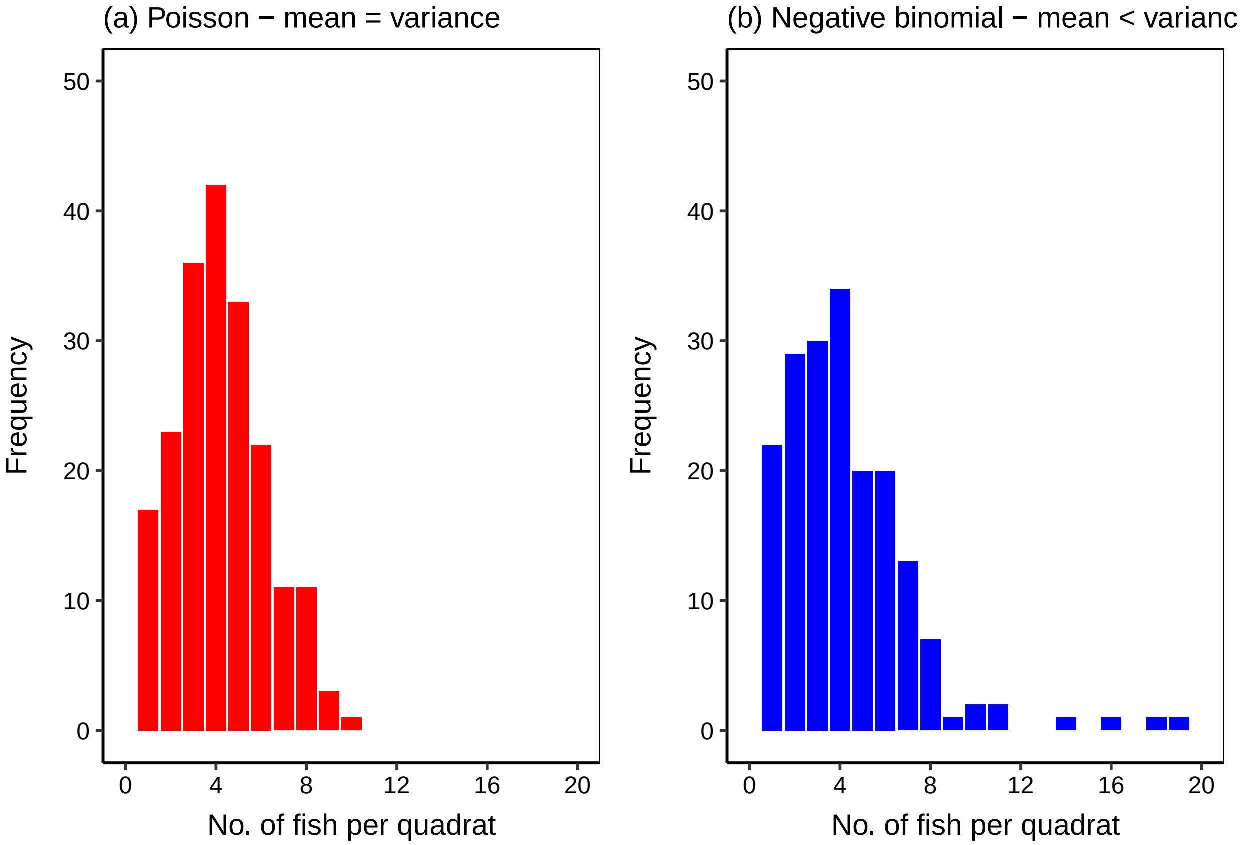

(a) Expected under normal distribution

(b) What counts usually look like

7.2. Entry-level count data models

There are two entry-level options for modelling count data:

1. *Poisson GLM* - The poisson distribution assumes the mean = variance.
2. *Negative binomial GLM* - Expects more variance than the Poisson. Variance > mean.



7.3. Poisson GLM

Let's assume we counted the number of fish per quadrat (`no_fish_quadrat`) (i.e. a measure of abundance) at each of three different sites (`site`). We want to know whether abundances differ between the 3 sites?

The Poisson GLM should be our starting point when analysing count data. If the Possion GLM is not an appropriate fit (evaluated using model diagnostics), then we can proceed to using slightly more complex count models, such as the negative binomial GLM.

```
## Rows: 60
## Columns: 2
## $ no_fish_quadrat <int> 6, 4, 6, 5, 3, 2, 3, 4, 0, 6, 7, 6, 7, 4, 4, 5, 9, ...
## $ site            <chr> "A", "A", "A", "A", "A", "A", "A", "A", "A", "A", "...
```

Fitting the model

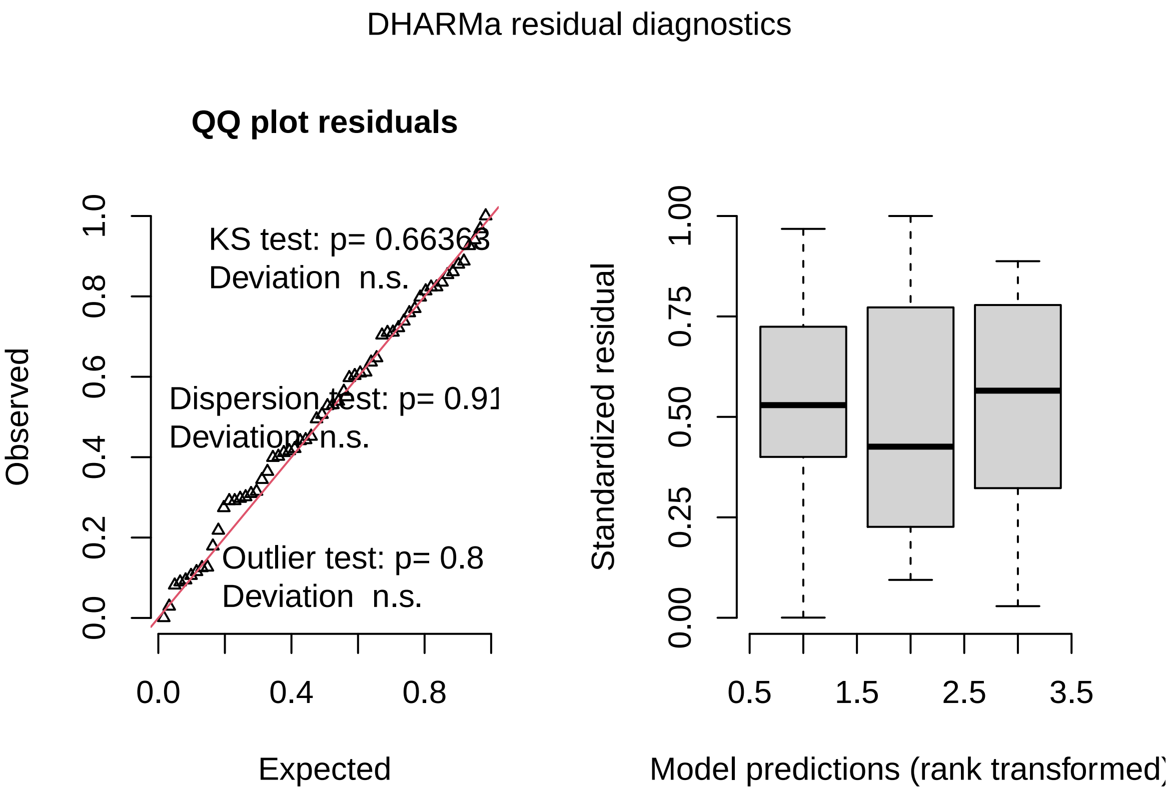
Our first option for analysing count data is the Poisson GLM. Specifying a Poisson GLM is very similar to the gaussian GLM's we fit during previous tutorials by changing the `family` argument.

```
mod1 <- glm(no_fish_quadrat ~ site,
            data = data,
            family = poisson(link = "log"))
```

Evaluate model fit

Evaluating the fit of a Poisson (actually any GLM's) is exactly the same as for the Gaussian GLM's that we have already covered, with a few additional tests.

```
mod1_diagnostis <- DHARMa::simulateResiduals(mod1)
plot(mod1_diagnostis)
```



LHS plot (QQplot): This plot tells us whether our data conforms to a Poisson data distribution. If our GLM is a good fit to the data, the white triangles will fall approximately on the red 1:1 line and the KS test P-value will be greater than 0.05. The Kolmorogorov-Smirnoff test (KS test) evaluates whether our data follow the distribution we specified in the `family` argument in the GLM call above (remember: we said "family = poisson").

- Interpreting this plot tells us that the Poisson data distribution was a good fit. Note that the points fall approximately along the 1:1 red line and the KS test result was not significant - which tells us that our model approximates a Poisson distribution well.

RHS plot (Fitted vs Residuals): This plot tells us whether there are issues with homogeneity of variances. GLM's, like ANOVA and linear regression, assume that the variance in the different groups, or across the range of x values, are approximately equal. We want to see the bold black lines for the different groups fall approximately at y = [0.25, 0.50, 0.75]. If there are issues, the bold lines will fall away from the y = [0.25, 0.50, 0.75], and all the lines will be red (not black).

- Interpreting this plot we can see that the bold lines fall approximately on the y = [0.25, 0.50, 0.75] lines, as would be expected. There is a no evidence for a deviation from equal variances.

Testing for overdispersion For a Poisson model, we can directly measure whether there was more variation in the data than under the assumption of a Poisson distribution. This phenomenon is called overdispersion. Poisson models assume that mean and variance increase linearly. We can evaluate this by calculating an overdispersion statistic from Gelman and Hill (2007).

- If the dispersion ratio is close to one = a Poisson model fits well to the data.
- Dispersion ratios larger than one indicate overdispersion, thus a negative binomial model or another distribution that assumes larger variances might fit better to the data.
- A p-value $< .05$ and/or ratio > 2 indicates overdispersion.

```
# Test for overdispersion
performance::check_overdispersion(mod1)
```

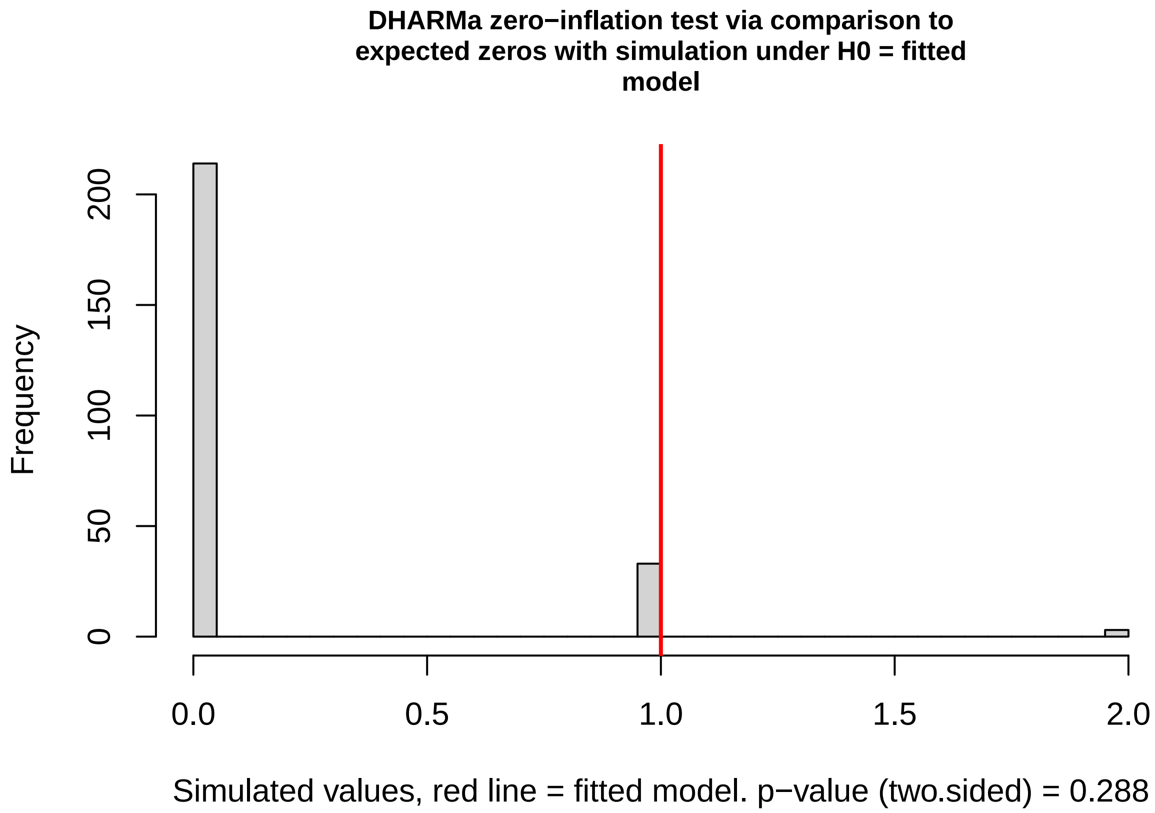
```
## # Overdispersion test
##
##        dispersion ratio =  0.972
##   Pearson's Chi-Squared = 55.385
##                 p-value =  0.536
```

```
## No overdispersion detected.
```

The dispersion ratio is approximately 1, indicating that there is as much variation in our data as expected under the Poisson distribution.

Testing for zero-inflation Count data often contains a large amount of zeroes. Anyone who has done fieldwork counting animals or plants will be acutely aware of this - I spent most of my PhD counting no insects during each of my surveys. We can specifically test for whether our data has too many zeroes than expected under the Poisson distribution (this is called 'zero-inflation' in the literature).

```
# Test for zero-inflation
DHARMa::testZeroInflation(mod1)
```



```
##
##  DHARMa zero-inflation test via comparison to expected zeros with
##  simulation under H0 = fitted model
##
## data:  simulationOutput
## ratioObsSim = 6.4103, p-value = 0.288
## alternative hypothesis: two.sided
```

If there are too many zeros in our data, the p-value will be significant (< 0.05).

- This example tells us that there aren't too many zeros in our dataset (P > 0.05).

Take home: The GLM that we fit was a good fit to the data. We can proceed with looking and the results.

Results

Now we get to the bit we are interested in: assessing statistical significance and calculating p-values. To do this, we will use a Likelihood Ratio Test (LRT). When we only have 1 predictor variable (here: `site`), we need to calculate p-values using type I sum-of-squares (SOS). SOS's are just different ways that we ask `R` to calculate p-values.

PLEASE DO NOT USE SUMMARY() - THIS WILL PRODUCE THE WRONG P-VALUES WHEN YOU HAVE MORE THAN 1 PREDICTOR VARIABLE

```
# Perform LRT  with type I sum-of-squares
car::Anova(mod1,
           # Specify we want a Likelihood Ratio Test
           test = "LR")
```

Parameter significance

```
## Analysis of Deviance Table (Type II tests)
##
## Response: no_fish_quadrat
##      LR Chisq Df Pr(>Chisq)
## site   105.42  2  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

So, the LRT tells us that the number of fish we counted per quadrat (`no_fish_quadrat`) was significantly different between sites (`site`) ($\chi^2 = 105.4$ d.f. $= 1$, P < 0.001).

***Post hoc* analysis** We often want to determine where the significant differences occur that we have identified above. To do this, we can perform *post hoc* tests. Below, we will perform Tukey's *post hoc* comparisons, which is arguably the most common and widely used test. We have to feed in the model name (`mod1`), and then tell `R` that we can't pairwise comparisons amongst our predictor variable (here: `site`).

```
emm1 <- emmeans(mod1,
                specs = pairwise ~ site,
                adjust = "tukey")

# Get 95% confidence intervals
emm1$contrasts %>%
  summary(infer = TRUE)
```

```
##  contrast estimate     SE  df asymp.LCL asymp.UCL z.ratio p.value
##  A - B     -0.9808 0.1197 Inf    -1.261    -0.700 -8.196  <.0001
##  A - C     -1.0524 0.1185 Inf    -1.330    -0.775 -8.878  <.0001
##  B - C     -0.0716 0.0868 Inf    -0.275     0.132 -0.824  0.6879
##
## Results are given on the log (not the response) scale.
## Confidence level used: 0.95
## Conf-level adjustment: tukey method for comparing a family of 3 estimates
## P value adjustment: tukey method for comparing a family of 3 estimates
```

So, we can see that the number of fish recorded per quadrat was higher at `siteB` than `siteA` ($P < 0.001$) and higher at `siteC` than `siteA` ($P < 0.001$), but was not significantly different at `siteB` versus `siteC` ($P = 0.688$).