# Tutorial #1: Introduction to GLM's

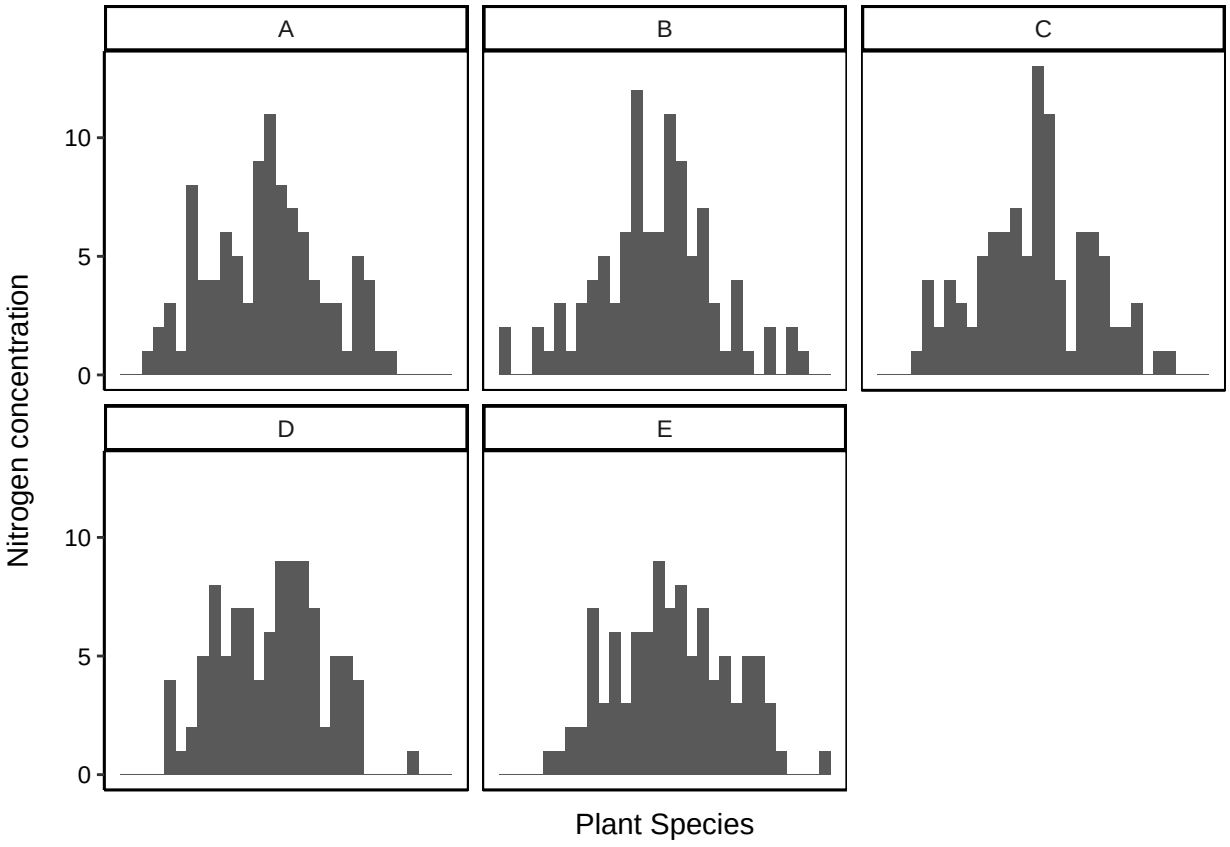
Guy F. Sutton

Centre for Biological Control, Rhodes University

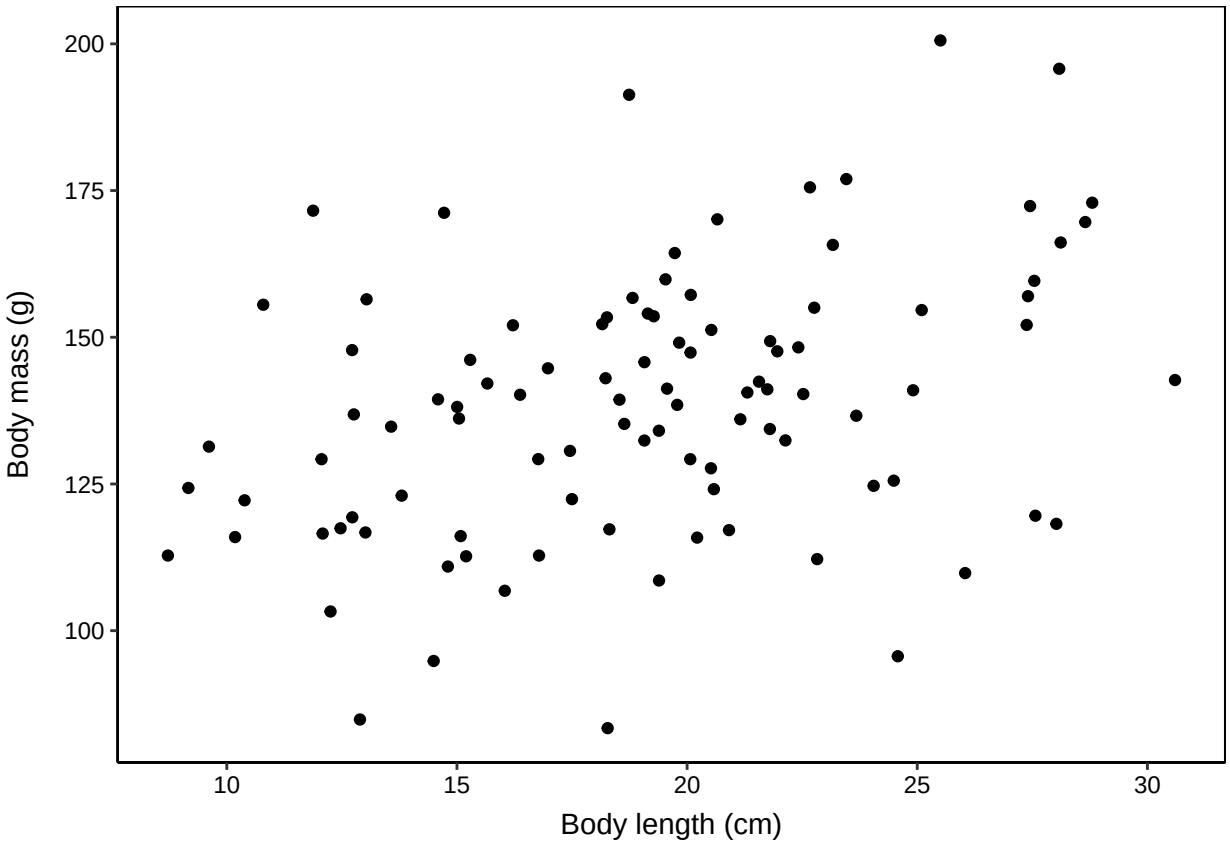
## 1. General Linear Model (GLM)

GLM has traditionally referred to a **General Linear Model**. A GLM allows us to fit a linear regression model to a continuous response variable given continuous and/or categorical predictor variables.

- For example, let's say you measure the nitrogen content from leaves from 5 different plant species and want to know whether leaf nitrogen differs amongst species, you will be fitting a GLM. However, most undergraduate statistics courses will refer to this model as an *Analysis of Variance (ANOVA)* because you have a single categorical predictor variable (`plant_species`).



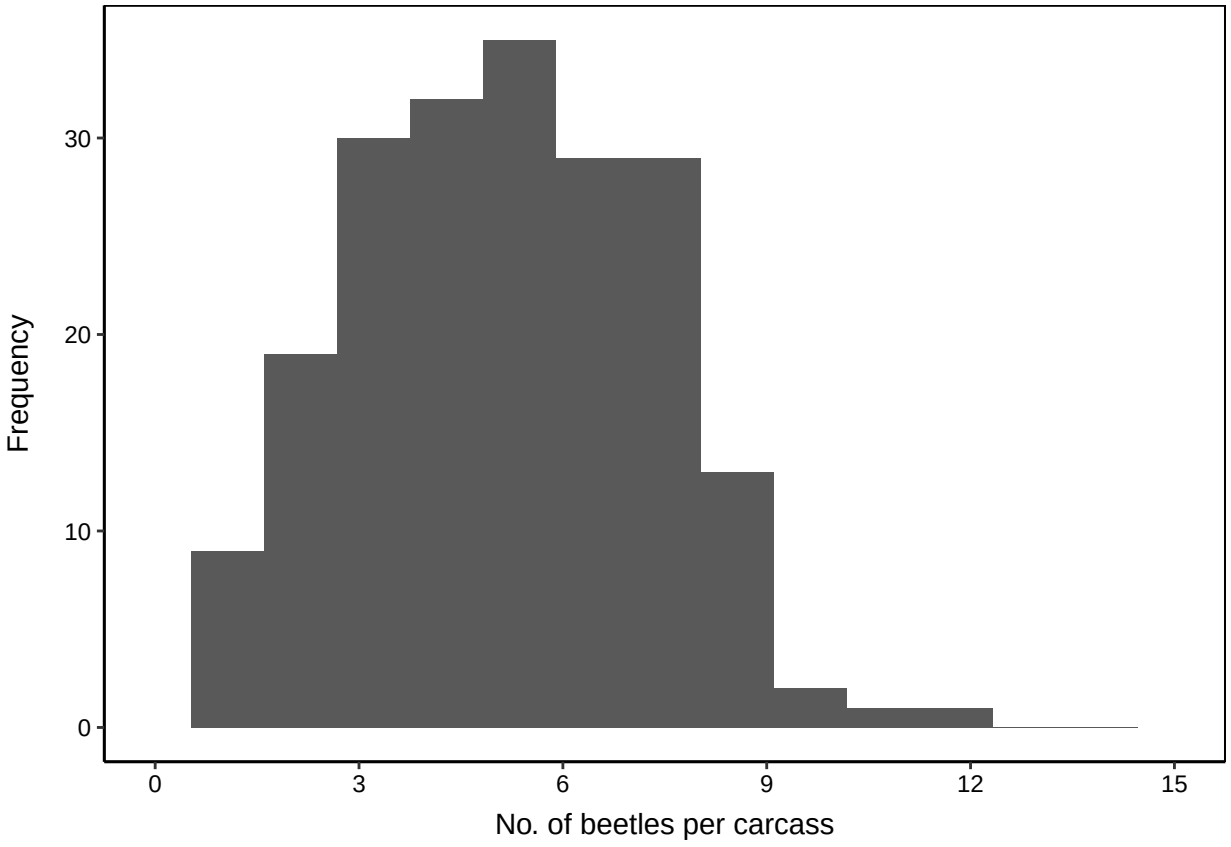
Alternatively, let's say you measure the body length of 100 individuals of a single fish species, and want to know whether body length correlates with body mass, you will be fitting a GLM. However, most undergraduate statistics courses will refer to this model as an *Linear Regression* because you have a single continuous predictor variable (`body_length`).



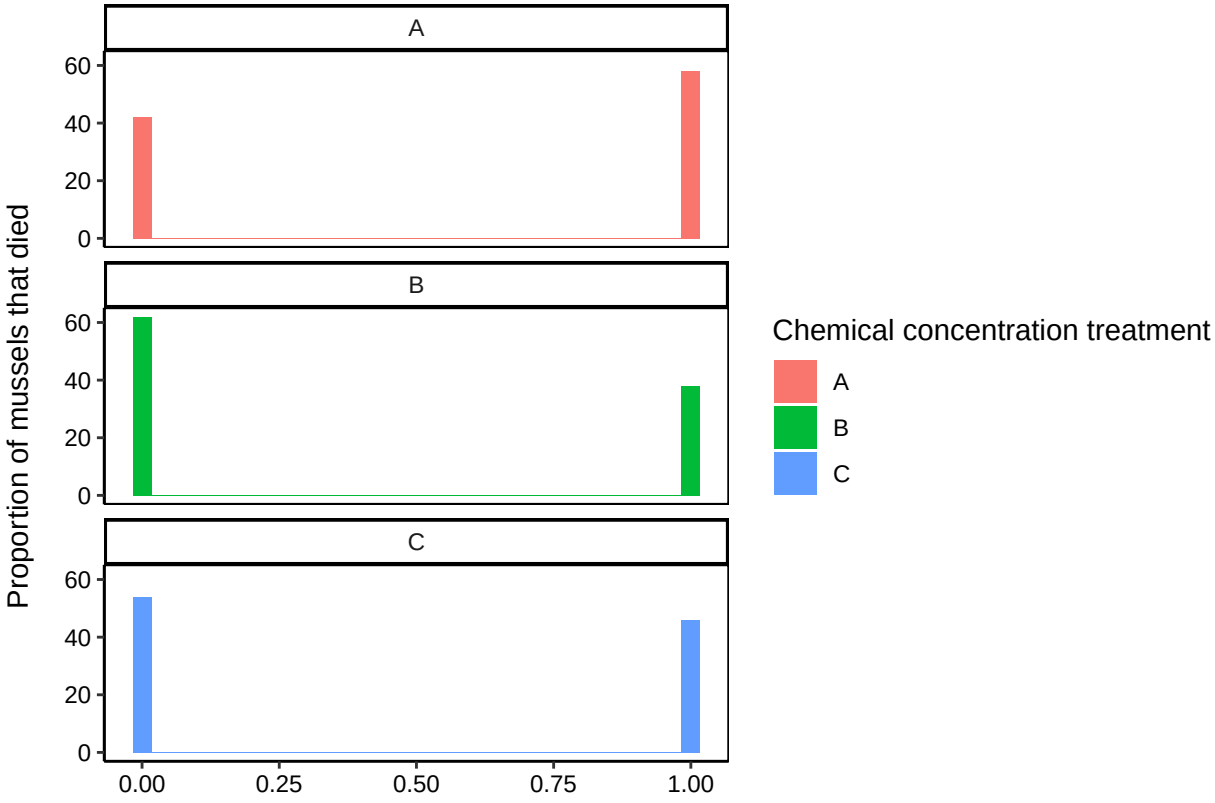

So, without probably knowing it, you have been using GLM's in your undergraduate research projects and/or honours projects.

As you progress through your research career, you will encounter datasets that are not suitable to analyse using linear regressions or ANOVA's.

- Many projects will collect count/abundance data, such as: counting the number of beetles per dead animal carcass, or number of fish caught in 15 minutes, or the number of birds recorded in a 20 meter transect.



- Other research projects will collect data that is either binary or proportion data. For example, you may be interested in whether larger aloes are more likely to die than smaller aloes. You go into the field, and record whether each of 30 aloes died or not and their respective heights. Alternatively, you could be interested in what proportion of mussels in your experimental buckets died when applying 3 different toxic chemical concentrations.




Notice the different structure of the datasets when we visualise the data!!! None of these datasets would be suitable for analysis using linear regressions or ANOVA's.

## 2. Generalized Linear Model (GLM)

In recent times, GLM has become synonymous with the **Generalized Linear Model**. This is basically the same as the GLM we discussed above, however, our response variable no longer has to be a continuous number. Our response variable can be a continuous number (e.g. nitrogen concentration), but it can also be a count (e.g. number of birds per transect), binary (scoring aloes as dead or alive), or a proportion (proportion of mussels that died at different chemical concentrations).

As such, **GLM** is an umbrella term for many different statistical models, with the general structure of trying to explain some response variable given a single or combination of predictor variables. The distinction between which GLM to choose largely depends on the structure and variance of your response variable (i.e. variable of interest) - more on this later.

## Course structure

In this course, we are going to cover when to use the different GLM's, how to specify these models in R, how to interpret the results, and how to report the results in your project write-ups/scientific articles. The following models are the most commonly used GLM's in ecology:

(1) *Gaussian GLM* (same as ANOVA/linear regression): Continuous response variable, and continuous and/or categorical predictor variables.

(2) *Poisson GLM*: Count/abundance response variable, and continuous and/or categorical predictor variables.

(3) *Negative Binomial GLM*: An extension to the Poisson GLM for analysing count/abundance response variables, when there is more variance in your dataset than expected by the Poisson GLM.

(4) *Logistic GLM*: Binary or proportion response variable, and continuous and/or categorical predictor variables.

(5) *Multivariate GLM*: Analysis of species community composition data (e.g. counting the abundances of multiple species in a community).