**Introduction to non-Gaussian traits (count, binary and proportion data) and generalized linear mixed-effects models**

***Goal:***to understand what kinds of traits are ‘non-Gaussian’, how they are different from Gaussian (normally distributed) traits, and how they can be modeled using the ‘generalized’ linear mixed-effects model (GLMM) framework.

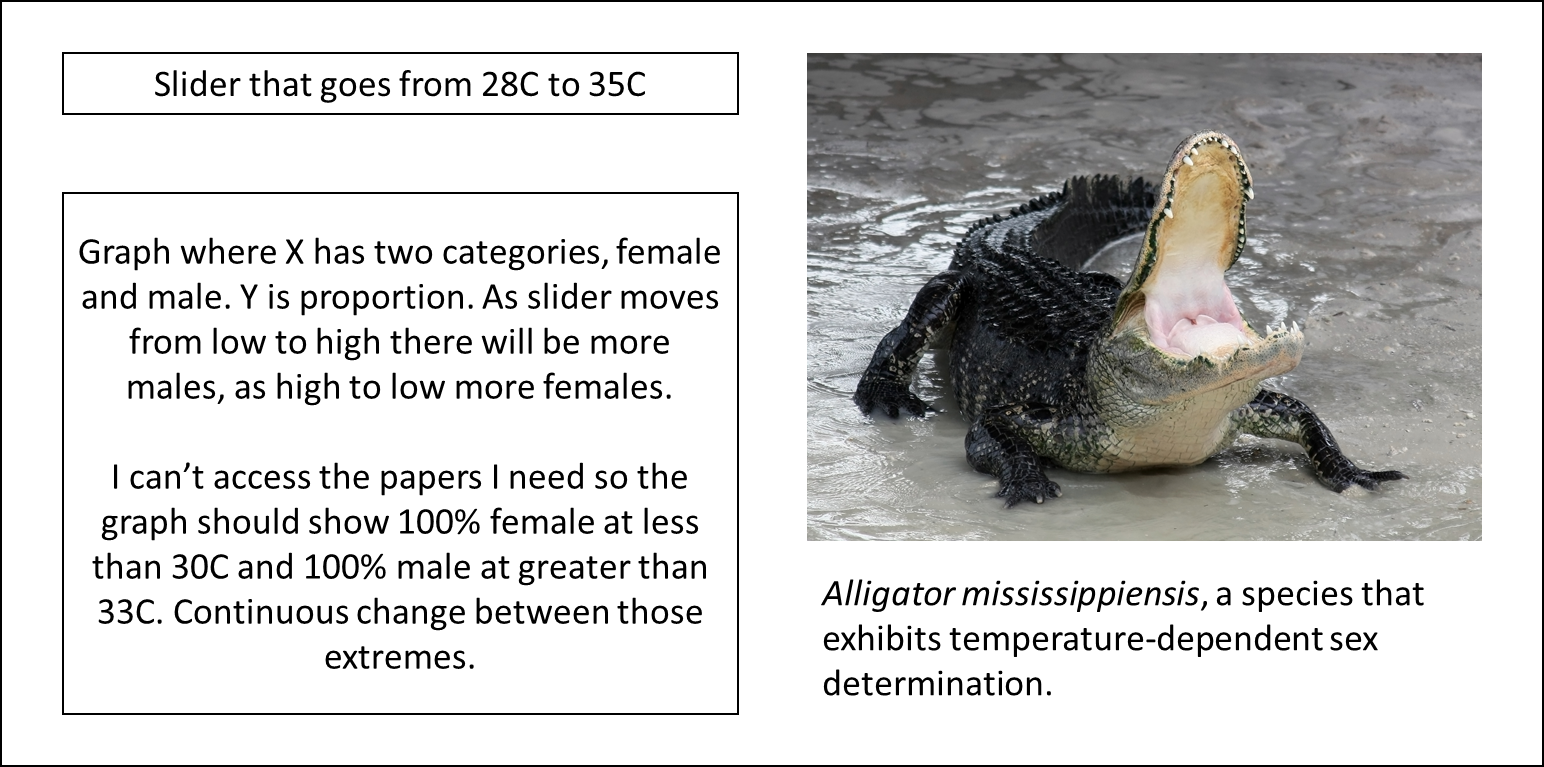
***Step 1. The biology that produces binary, proportional, and count data.***

**Sub-goal: Introducing how biology may result in non-normally distributed data**

**Introduction:** Before you start analyzing your data, one thing you want to consider is the biology of what you’re looking at and the how the phenomenon you’re interested in is expressed by organisms. Most of the statistical methods we’re familiar with make assumptions about how data are distributed (or, really, how errors are distributed), specifically they assume that our data are normally distributed (i.e., having a Gaussian error structure). However, is this always an appropriate assumption?

While the assumption of normality (in residuals) may work in many instances, it is also easy to think of biological traits that clearly don’t fit this framework. For example, the sex-ratio of your offspring will not be something that can be measured in such a way that the data would be normally distributed and thus can’t be analyzed with typical models. This is most apparent in species with temperature-dependent sex determination where the temperature an egg is incubated at determines whether it is female or male:

Picture: @wikimedia



As you can see from playing with the temperature at which a clutch of American alligators is incubated, sex ratio does not conform to a normal distribution and so using something like a regression to understand how temperature affects sex ratio would be inappropriate.

A simpler case than this one is one where sex is determined via particular combinations of allosomes (i.e. sex chromosomes). In such cases the sex of offspring will be determined via what’s known as a Bernoulli process. As an example, let’s consider how the coin-flip process of genetic sex determination interacts with the number of coin-flips to give a distribution of sexes:

--here there would be a field where the user could enter a number of offspring, a run button and then a graph with two bars, one for females and one for males and the y-axis would be proportion. The graph then changes with the sample number but the probability is 50/50.

The outcome of a single coin flip, like that for a single egg, represents what is known as **binary data**, i.e., this data takes the form of 1 or 0 (A or B, female or male, etc.). Many other biological processes take a similar form: at a particular instance one of two outcomes is possible. Importantly, this process also works when the odds of either outcome are not 50/50:

--Same as above type set up but with an additional entry field where the user will input the probability of condition ‘A’.

This situation is now starting to approach what we saw with temperature-dependent sex determination: the trait we’re measuring is always one of two conditions (e.g. female or male) but the probability of either condition is not 0.5 and, in fact, is dependent on some other environmental condition.

Many biological traits are expressed in binary terms. This includes traits expressed repeatedly over one’s lifetime, e.g. whether a bird mates during a particular breeding season or whether a plant flowers during a particular year. As you’ll see in later steps of this module, this Bernoulli process can be statistically modelled using what is known as a binomial distribution restricted to binary outcomes. For now, let’s consider some other types of distributions and generating patterns we might see.

**Proportional Data**

What if instead of a single egg, we consider the sexes of a clutch of animals? In this case the sex of each individual egg is the product of a Bernoulli process but the data we might collect is actually in proportional form, e.g., proportion of males in the clutch. In fact, we actually were plotting data in this sort of manner in all three of our figures above. In this case what matters is the probability of either outcome and, as you saw above, the number of coin flips (or eggs) and the probability of either of the two outcomes.

Exercise: explore the effects of the number of coin flips/eggs/trials and the probability.

Here there would be an interactive component where the user enters the number of coin flips and the probability of success. What is plotted is a histogram of proportions on the x from 1000 random simulations of their parameters.

Many types of biological data might similarly be expressed as proportions: for example, the proportion of seeds that successfully germinate and the proportion of females in a group. This type of proportional data is **binomially distributed**.

*When proportional data are not from a binomial distribution*

The important characteristic of the data types above is that they are generated via a Bernoulli process: an egg is either female or male, a seed either germinates or it doesn’t. These outcomes are, generally, independent of each other and this is actually the key assumption for analyzing proportional data. Importantly, data that is expressed or summarized as a proportion but is not generated from a Bernoulli cannot be analyzed in the same manner. For example, we might often express time spent performing a particular action as a proportion but how we divide the time increments over which we record are typically not independent of each other. For example, if an individual is currently foraging, it is likely still foraging a second later but way less likely a week later. This distinction can be very important in analyses and when data not generated by a Bernoulli processes are analyzed as a proportion, any resulting p-values and confidence estimates will be inappropriate.

**Count Data**

Besides binary proportional data, another major class of data is count data. Examples include traits like how many clutches a bird has in a single year and how many flowers a plant produces during a growing season. In both of these examples the key aspect is that some act is performed or something is produced additively. The average production over some sampling period (breeding season, life-time, etc.) is then the average of what is known as a Poisson distribution, i.e. the rate (λ) of the Poisson process as discussed in the next step of this module. This mean then determines the shape of resulting distribution:

Here there would be an interactive component where the user enters the mean for a Poisson distribution. What is plotted is a histogram of counts from 1000 random simulations.

As you may have found above, if you set a high enough mean, the Poisson distribution begins to visually resemble a normal distribution (if you didn’t do so earlier enter a large number for the above figure). Importantly, however, this distribution is not actually normal and should not be considered as one because Poisson distributions have the additional property that the variance of the distribution is equal to its mean.

***Step 2: Introduction to Bernoulli, binomial and Poisson distributions***

**Sub-goal:**  to learn statistical properties of the three key non-Gaussian distributions

**Introduction:** In Step 1, we considered biological scenarios where non-Gaussian data could be produced. The three important types of data are binary, proportional and count data, each of which corresponds to a uniquely named statistical distribution, namely Bernoulli, binomial and Poisson. Remember that a Gaussian (normal) distribution is characterized by a mean (*μ*) and a variance (*V*). These two quantities are called statistical parameters for the Gaussian distribution. Statistical parameters for the three non-Gaussian distributions are not the mean and variance (at least they are not called so). Let’s look at the statistical parameters for each non-Gaussian distribution now.

A **Bernoulli** distribution is characterized by only one parameter, *p*, which is often interpreted as a probability of success. Or more biologically, for example, you can see it as the probability of female offspring (see Step 1). More formally, we can write a trait *y*,which follows a Bernoulli distribution along with its mean (µ) and variance (V) as:

A **binomial** distribution has one more statistical parameter, which is *m*, the number of trials. A biological example is the number of offspring in a brood (as *m*) with the probability of female being *p*. More formally, we can write a proportional trait, *y* as:

As you can see a Bernoulli distribution is a special case of a binomial distribution with *m* being 1.

Like the Bernoulli distribution, the **Poisson** distribution has only one statistical parameter. This parameter is often called, *λ* (termed the ‘rate’ parameter). We can formally write a count trait following a Poisson distribution as:

As you can see, the mean equals the variance in a Poisson distribution. Also, for both Bernoulli and binomial distributions, the mean and variance are also tightly related. In fact, a unique feature of a Gaussian distribution is non-existence of the relationship between the mean and variance. Non-Gaussian distributions, in general, have, what is called, a mean-variance relationship. It may be interesting to note that a binomial distribution becomes a form of a Poisson distribution when *p* is very small because:

**Exercise:** We explore the mean-variance relationship for Bernoulli distributions with *p* ranging from 0 to 1.

Here there have an interactive component simulating 10000 (a lot) outcomes of Bernoulli trials with different p and plotting. 7 different ps? (0, 0.1, 0.3, 0.5, 0.7, 0.9, 1). It would be good to plot these separately and also plot means and variances

We now do the same for Poisson distributions where *λ* varies between 0.5 and 100.

Here there have an interactive component simulating 10000 (a lot) outcomes of Poisson trials with different *λ* and plotting. 8 different *λ*s? (0.5, 1, 3, 5, 7, 10, 20, 100); the same as above

***Step 3: Introduction to generalized linear mixed-effect models (GLMMs)***

**Sub-goal:** to understand the idea of the link and inverse link function and overdispersion and how (dis)similar it is to model Gaussian and non-Gaussian traits.

**Introduction 1:** In other modules, we assumed the trait of interest, *y* follows a Gaussian distribution and we used a mixed-effects modeling framework to partition variance in *y*. For a non-Gaussian trait, *y*, somehow we would like to make *y* normally distributed, removing the mean-variance relationship. Then, we could use the same mixed-effects model framework we have been using in other modules so far. This ‘transformation’ can be done using the ‘link’ function in generalized linear models (GLMs) if *y* values are independent of each other or in generalized linear mixed models (GLMMs) when *y* values are repeated measures or have a correlational structure. Also, the ‘inverse link’ function can bring results from the ‘link’ scale (*y’*) back to the original non-Gaussian data scale. This link scale is often called as the latent scale. For example, a commonly used link function for binary (and proportional) data is the logit link function. The logit function and its inverse can be written as:

Unlike *y*, *y’* is assumed to follow a Gaussian distribution.

**Exercise 1:** To understand the concept of the link and inverse link function better, we will simulate *y’* (a non-Gaussian trait on the latent scale) and convert simulated values first to *p* and the non-Gaussian trait *y*.

We now simulate 10000 data points using rnorm(10000, 0, 1). Then, use plogis() to change this vector into binary data using binom(10000, size = 1, prob = vec).

Note that we can get binary data from a normal distribution by using the inverse link function and Bernoulli distributions. But the reverse operation requires modeling (which we will practice below) because the logit transformation of 0 and 1 is negative and positive infinity, respectively.

**Introduction 2**: Another concept you need to learn for modeling non-Gaussian data is overdispersion, which we will explain more below (‘dispersion’ relates to variability of a distribution like variance but it is a more general term). Imagine that we record the number of female and male offspring that female animals have per breeding season across the number of years (*y* being a concatenation of female and male offspring per female *i* and per season *h*. Then, using the logit link function with a binomial error structure, *y* can be expressed as a GLMM:

where logit-1 is the inverse link function and *ohi* is overdispersion, which is normally distributed as for the individual effect *Ii* (note that we skip the notation *y’* which we used above). All the key parameters (population mean *β0* and individual-specific deviation *Ii*) are estimated on the latent scale where random effects are (assumed to be) normally distributed. It is also interesting to notice that *ohi* is very much like *ehi* (residuals), which we have seen in other modules; in the statistical literature, *o* is known as additive (over)dispersion because there is an alternative way of implementing dispersion known as multiplicative overdispersion (for more details of additive and multiplicative dispersion, see Nakagawa and Schielzeth 2010). The overdispersion *ohi* is also known as the observation-level random factor (effect) because the number of categories of this random effect matches the number of data points, this aspect is also somewhat analogous to the residuals in a normal mixed-effects model.

Overall the GLMM above looks very much like normal mixed-effects models apart from the overdispersion and the (inverse) link function, which connects *phi* and *yhi* via binomial distributions. For a binary trait (dead or alive, male or female, present or absent), *mhi* is always 1. Also *ohi* is always 0 because overdispersion is not ‘identifiable’ in binary data (see module XXX.

Now, imagine that the number of matings (*y*) per male (*i*) animal and per season (*h*). Then, using the log link function with a Poisson error structure, *y* can be written as a GLMM:

where *exp* is the inverse link function (of log), and the other notations are the same as above. Again, on the latent scale, random effects are all normally distributed, and the equation is very similar to a Gaussian trait.

**Exercise 2:** We now set the 3 parameters (*β0*, *Vi* and *Vo*) on the latent (link) scale in the Poisson GLMM with the log link function and generate count data. The simulation is done with 100 individuals that are sampled 5 times each.

Here we can set all the 3 parameters on sliders like other modules (e.g. Basic lessons Step 2).

Once they generated data and then use glmer (lme4) package to fit a model

CODE

overdispersion<- as.factor(1:length(Phenotype)) # the observation-level random effect

LMM<- lme4::glmer(Phenotype ~ 1 + (1|Individual)+(1|overdispersion), family = poisson(link=log), data = sampled\_data)

You should be able to recover these parameters

**Conclusion:** The underlying biology that generates phenotypes is important to consider because not all biological processes will produce Gaussian data. Deciding how you are going to analyze your data requires that you consider how it was generated and this biology will determine the types of analyses you conduct. Once we identify non-Gaussian data such as binary, proportional and counts. Then, wee can use GLMMs to model these data (with appropriate link function and error structure, e.g., log link function with a Poisson error structure). On the latent scale, we can estimate regression coefficients and variance components just like how we model Gaussian data.