**Understanding and using generalized linear mixed-effects models: analysing non-normal data.**

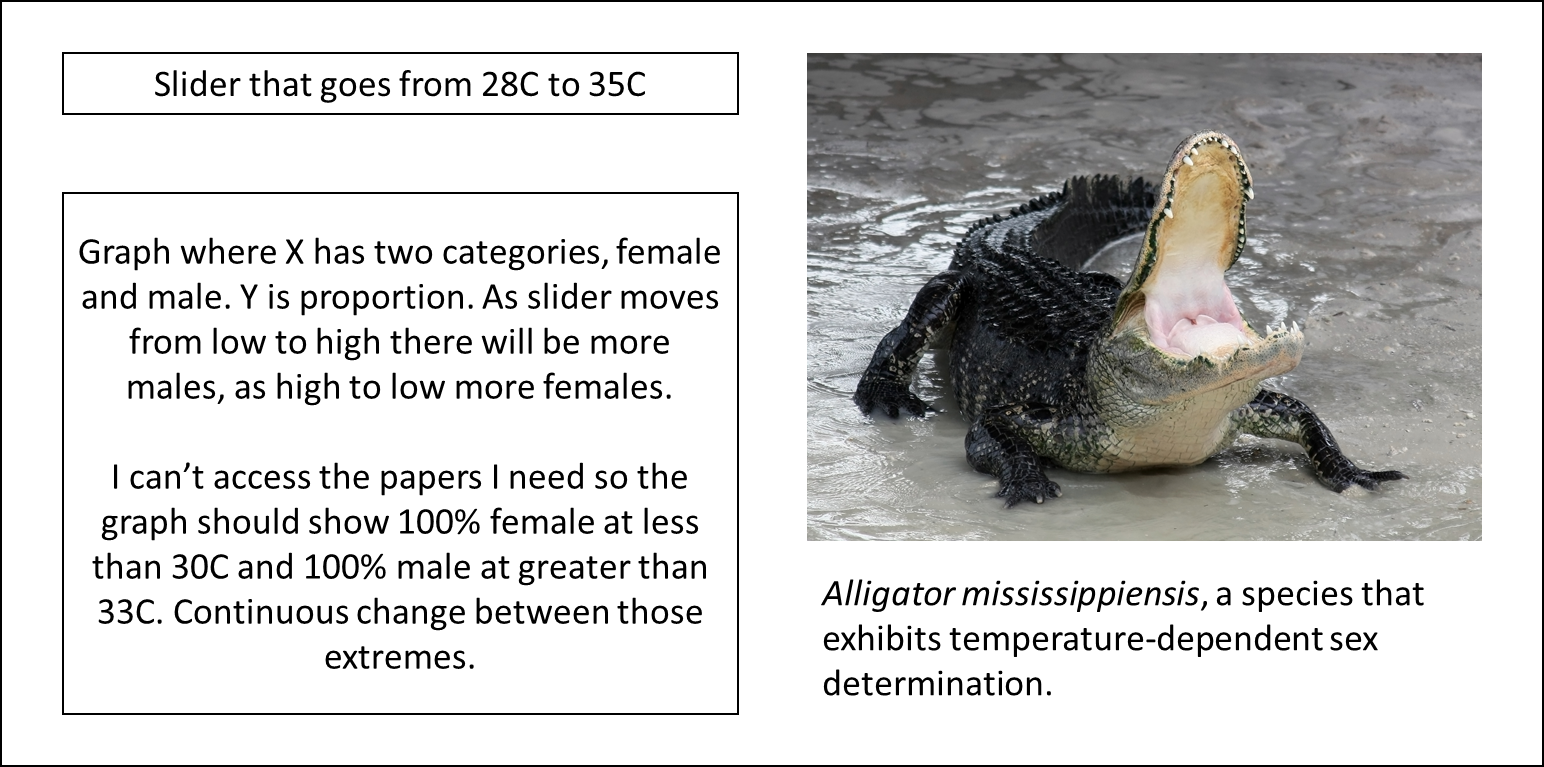
***Goal:***to develop understanding of how the distribution of data influences analyses and inferences.

***Step 1. The biology that produces binary, proportional, and Poisson distributed 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 deeply consider is the biology of what you’re looking at and the how the phenomenon you’re interested in are expressed by organisms. Most of the statistical methods we’re familiar using make assumptions about how data are distributed (or, really, how errors are distributed), specifically they assume that our data are normally distributed. However, is this an appropriate assumption?

While the assumption of normality 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:



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 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 process 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 odds 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 number of number of coin flips/eggs/trials and 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: proportion of seeds that successfully germinate, ADD IN OTHER EXAMPLES. This type of proportional data is known **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. 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 normally distributed data, binary data, and 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 fit as one because Poisson distributions have the additional property that the variance of the distribution is equal to its mean.

In the next steps of this module you’ll first learn more about Binomial and Poisson processes and then how the use of general**ized** linear mixed-effects models can allow you to address similar questions about labile phenotypes as covered in other modules.

**Conclusion:** The underlying biology that generates phenotypes is important to consider, not all biological processes will produce normally distributed 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.