# Nonlinear, Non-normal models

## Lecture 05.2: Generalized Linear Models (GLMs)

Lauren Sullivan

Module: Linear, Nonlinear, and Mixed Effects Models

# Readings

**Required for class:**

- NA

**Optional:**

- Crawley, M. *Statistics: An Introduction Using R*
- Bolker, B. *Ecological Models and Data in R - Ebook version*
- R-tutorials, Nonlinear Regressions

# General*ized* Linear Models (GLMs)

Generalized linear models calculate nonlinear regressions when you have non-constant variance in the data, or when your dependent variables are not normally distributed.

- Sometimes you know you have nonlinear or non-normal data and you a-priori know you need to use a non-standard linear model. For example:
  - Exponential growth.
  - Binary response data (survived till the next time point yes/no)
  - Count data

# Model Assumptions

Linear Model assumptions:

1. Relationship between dependent and independent variables are linear
2. (Multivariate) Normal distributions of error variance $\epsilon$
3. Equal variance (aka - Homoscedasticity)
4. Independence of observed samples

Generalized Linear Models take #1 and #2 a step further, and assume that the error variance of the dependent variable comes from a family of distributions knows as the *Exponential Family*.

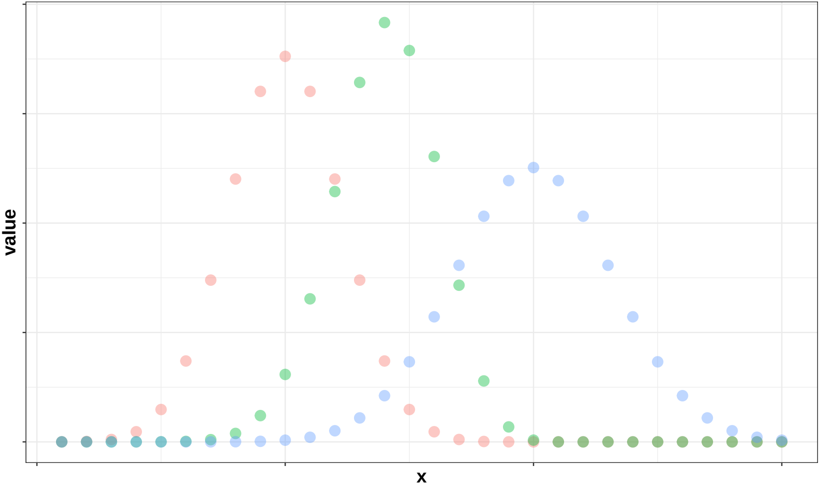
- The Exponential Family includes many of the most of the common distributions, including: Normal, Exponential, Beta, Binomial, Chi-Square, Gamma, Poisson, Wishart, etc.

# glm()

To analyze GLM's in R, you use the function glm(), with an additional argument for `family`.

- This `family` argument depends on the type of data you are trying to model.
  - Binomial
  - Poisson
  - Inverse Gaussian

# Binomial Data



The Binomial distribution is discrete, this is called a proability mass function. Often, an observation of 1 is a success (survival) and 0 is a failure (death).

# Binomial Distribution

The Binomial Distribution is discrete and describes the number of successes in a sequence of $n$ independent experiments/trials.

- Parameters: $n \in \{1, 2, 3...\}$ (number of trials), $p \in [0, 1]$ (success probability for each trial)
- Support: $k \in \{1, 2, 3..., n\}$ (number of successes)
- PMF:

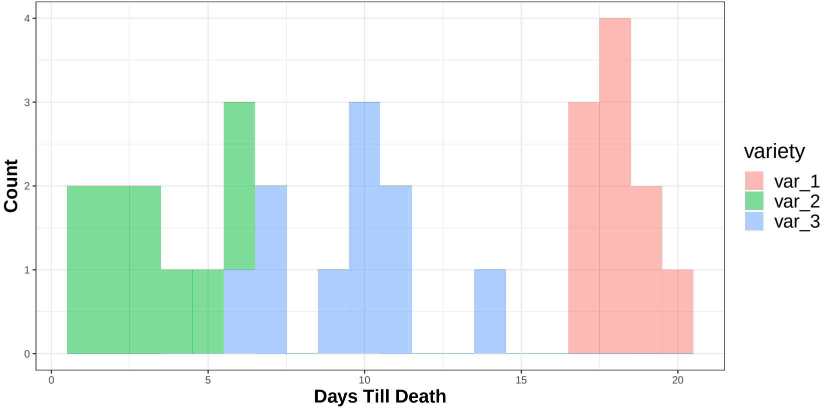
$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Some Examples:

- Survival Analysis
- mRNAseq reads, where you have $n$ reads (or trials), with a probaiblity of success $p$ for a particular transcript/gene in each read
- Anything with 0/1 data through time/space

# Binomial Experiment

Imagine seeding survival (or days until death due to dessication) of three different varieties of soybean.



17 days till death means 17 days surving (1's), and the rest of the days dead (0's).

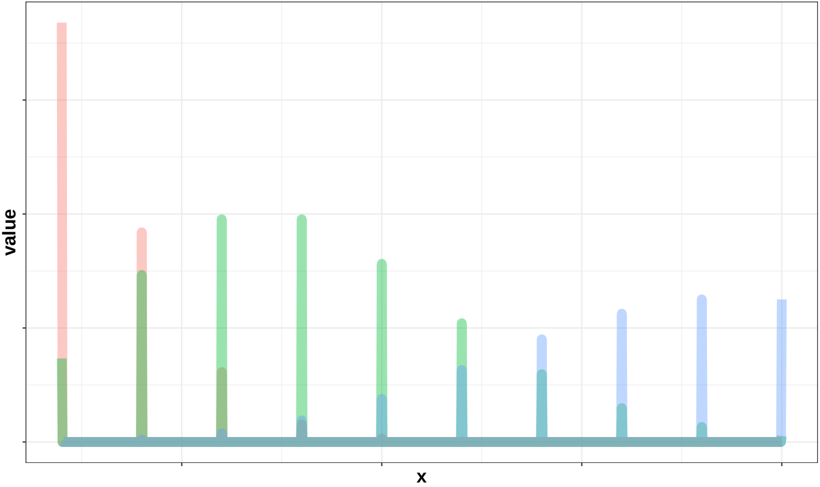
# Binomial GLM

Does the survival of soybean seedlings depend on the variety?

```r
summary(glm(alive ~ variety, family = "binomial",
            data = seedling))
```

```
##
## Call:
## glm(formula = alive ~ variety, family = "binomial", data = seedling)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1945  -0.6005   0.4343   0.4343   1.8983
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    2.3136     0.2470   9.366   <2e-16 ***
## varietyvar_2  -3.9351     0.3120 -12.615   <2e-16 ***
## varietyvar_3  -2.4137     0.2847  -8.477   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 831.11  on 599  degrees of freedom
## Residual deviance: 576.92  on 597  degrees of freedom
## AIC: 582.92
##
## Number of Fisher Scoring iterations: 4
```

# Poisson Data



The Poisson distribution is also discrete, thus it appears as points of probability at each $k$.

# Poisson Distribution

The Poisson Distribution describes the probability of a given number of events occurring in a fixed interval of time or space.

- Parameters: $\lambda \in \mathbb{R}^+$ (rate - must be a real number)
- Support: $x \in \mathbb{N}_0$ (natural numbers starting from 0)
- PMF:

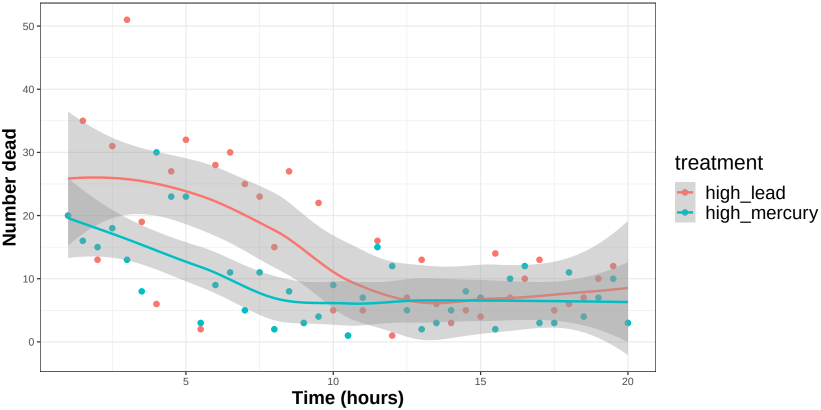
$$f(x; \lambda, k) = \frac{\lambda^k exp^{-\lambda}}{k!}$$

Some Examples:

- The number of meteorites greater than 1 meter diameter that strike Earth in a year
- Death or births of individuals at a given time
- Number of dispersal events at a given distance

# Poisson Experiment

Let's look at the survival of 2 populations of 500 mice each on a different diet: one high in lead, one high in mercury. We have time in hours since the diet was fed to the mice, and the number of mice from each population that died at each time point.



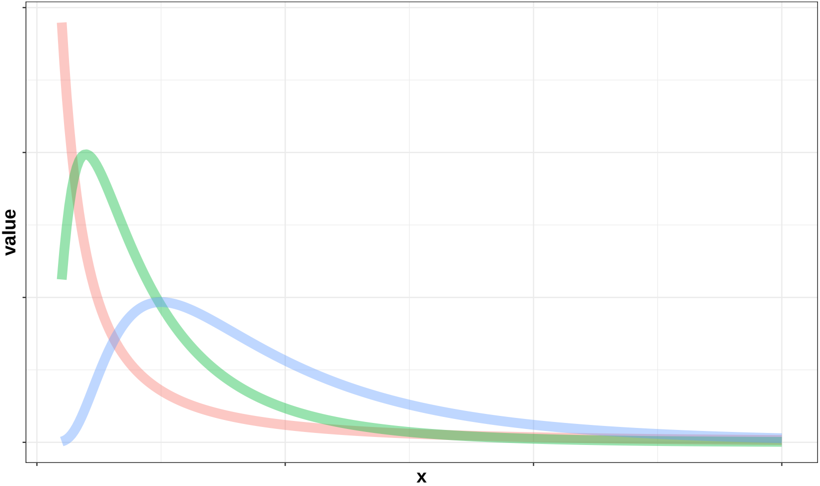
# Poisson GLM

Is the number of mice that die a function of time, and is there a difference between the two treatments?

```
summary(glm(number_dead ~ hours + treatment, family = "poisson",
            data = mice_diet))
```

```
##
## Call:
## glm(formula = number_dead ~ hours + treatment, family = "poisson",
##     data = mice_diet)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -5.1115  -1.6565  -0.3549   1.4570   4.7409
##
## Coefficients:
##                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)            3.435851   0.065169  52.722  < 2e-16 ***
## hours                 -0.083401   0.006234 -13.378  < 2e-16 ***
## treatmenthigh_mercury -0.442624   0.067450  -6.562  5.3e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 565.68  on 77  degrees of freedom
## Residual deviance: 330.82  on 75  degrees of freedom
## AIC: 647.42
##
## Number of Fisher Scoring iterations: 5
```

# Inverse Gaussian Data



The inverse gaussian distribution has two parameters, a central tendancy ($\mu$) and a dispersion ($\lambda$)

# Inverse Gaussian Distribution

The Inverse Gaussian Distribution does not have a nice description like the other distributions, but is a helpful 2-parameter, skewed distribution. As $\lambda$ tends towards infinity, this looks more like a gaussian distribution.

- Parameters: $\mu > 0, \lambda > 0$
- Support: $x \in (0, \infty)$
- PDF:

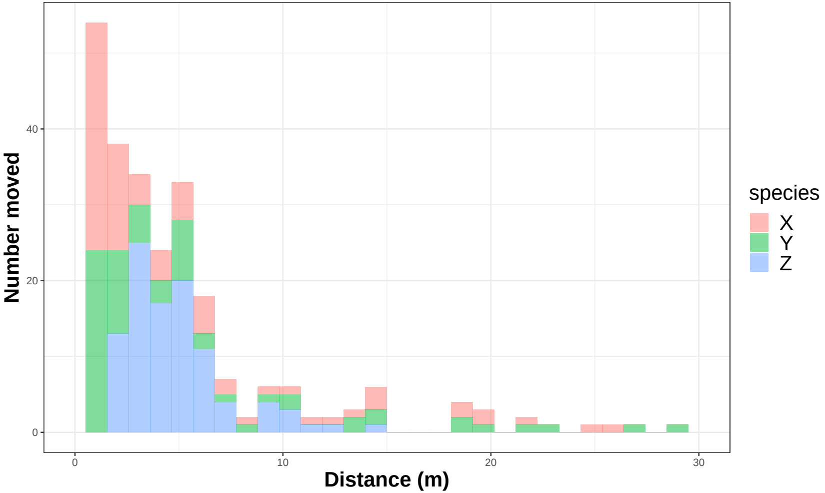
$$f(x; \mu, \lambda) = \sqrt{\left(\frac{\lambda}{2\pi x^3}\right)} exp \left[ -\frac{\lambda(x-\mu)^2}{2\mu^2 x} \right]$$

Some Examples:

- Dispersal events over an infinite distance
- non-negative, positively skewed data.

# Inverse Gaussian Experiment

Dispersal of 3 different frog species away from a home pond. Is there a difference in movement by species?



# Inverse Gaussian GLM

## Do species of frogs disperse with different patterns?

```
summary(glm(distance ~ species, family = "inverse.gaussian",
            data = frog))
```

```
## 
## Call:
## glm(formula = distance ~ species, family = "inverse.gaussian", 
##     data = frog)
## 
## Deviance Residuals: 
##      Min        1Q    Median        3Q       Max  
## -2.80282  -0.79175  -0.26663   0.01261   1.48327  
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)   
## (Intercept)  0.02708    0.01038   2.609  0.00955 **
## speciesY    -0.02648    0.01040  -2.546  0.01140 * 
## speciesZ     0.01541    0.01788   0.862  0.38933   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for inverse.gaussian family taken to be 0.6045674)
## 
##     Null deviance: 222.91  on 299  degrees of freedom
## Residual deviance: 200.84  on 297  degrees of freedom
## AIC: 1745.8
## 
## Number of Fisher Scoring iterations: 13
```