Generalized Linear Model (GLM)

Emily Haeuser

Eva Malecore

Ecology Konstanz

TReND

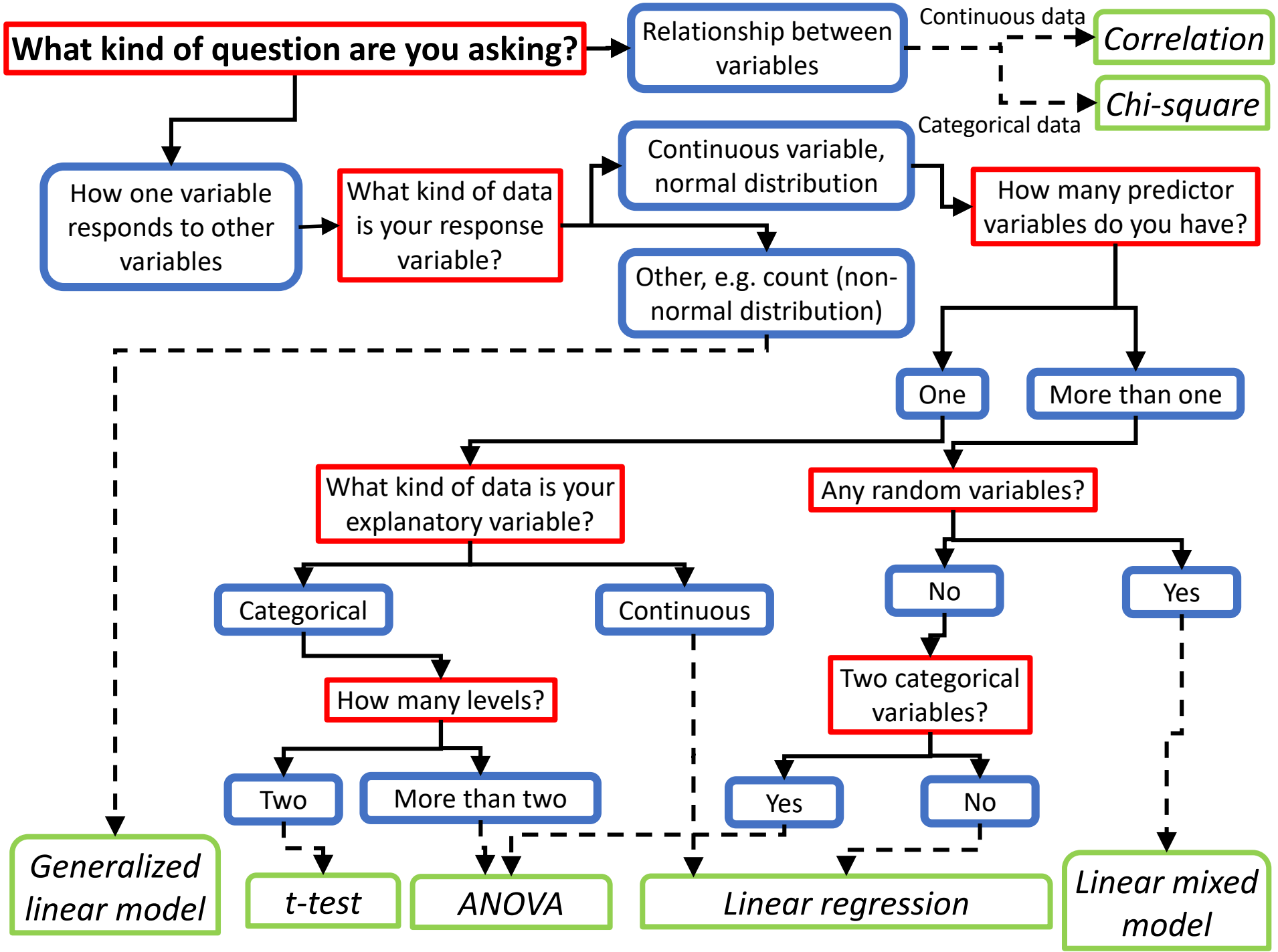
Institute for Health Metrics and Evaluation

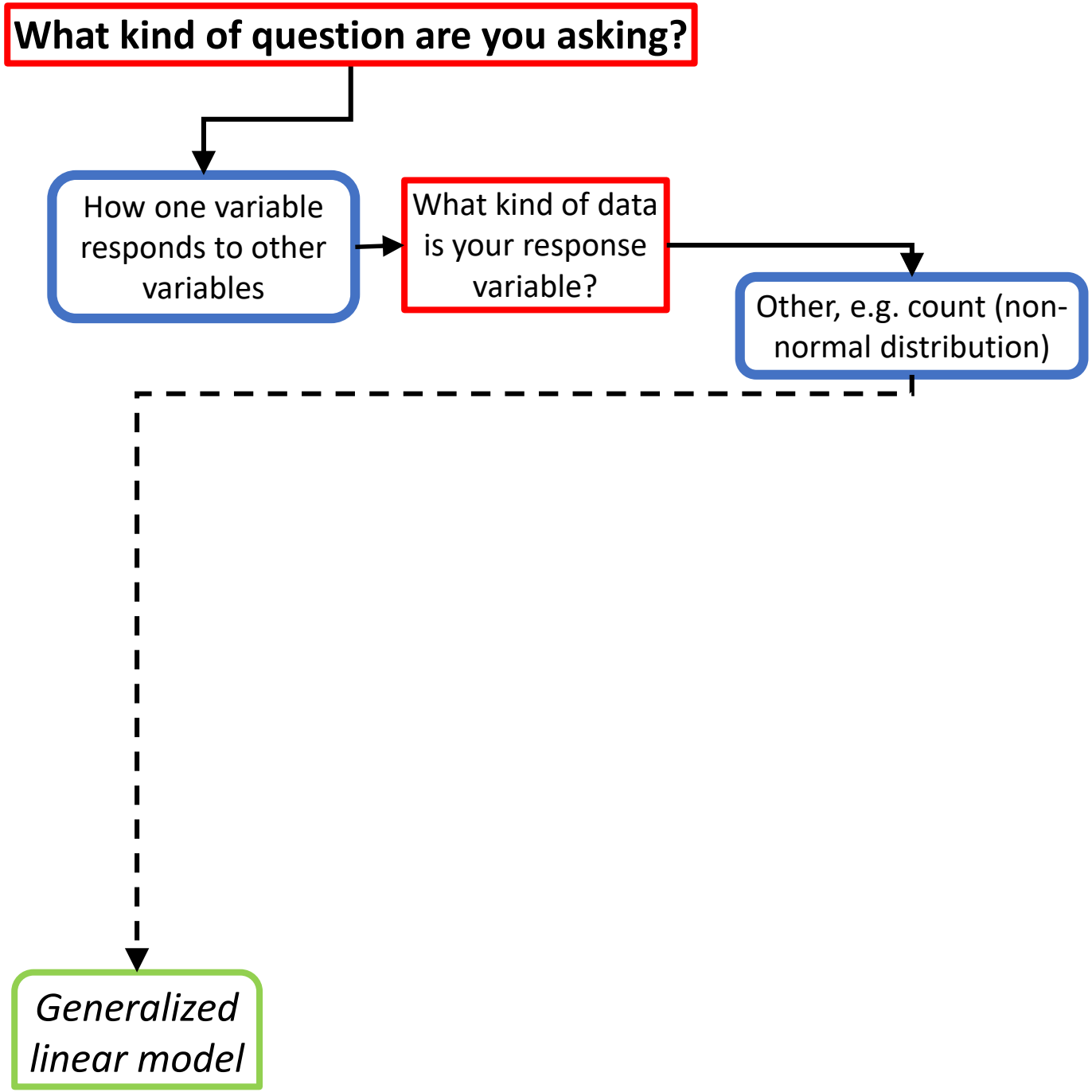
The Company of Biologists

International Max Planck Research School for Organismal Biology

Overview

Topics
1 – Generalized Linear Models intro
2 – Poisson GLM
3 – Binomial GLM


What kind of question are you asking?
Relationship between variables
Continuous data
Correlation
Categorical data
Chi-square
How one variable responds to other variables
What kind of data is your response variable?
Continuous variable, normal distribution
Other, e.g. count (non-normal distribution)
How many predictor variables do you have?
One
More than one
What kind of data is your explanatory variable?
Any random variables?
Categorical
Continuous
No
Yes
How many levels?
Two categorical variables?
Two
More than two
Yes
No
Generalized linear model
t-test
ANOVA
Linear regression
Linear mixed model


What kind of question are you asking?
How one variable responds to other variables
What kind of data is your response variable?
Other, e.g. count (non-normal distribution)
Generalized linear model

Reminder: LM assumptions

- Linear models (ANOVA, linear regression) imply several **assumptions**:
  - **Normality of residuals**
  - **Homogeneity of variance**
  - **Independence**

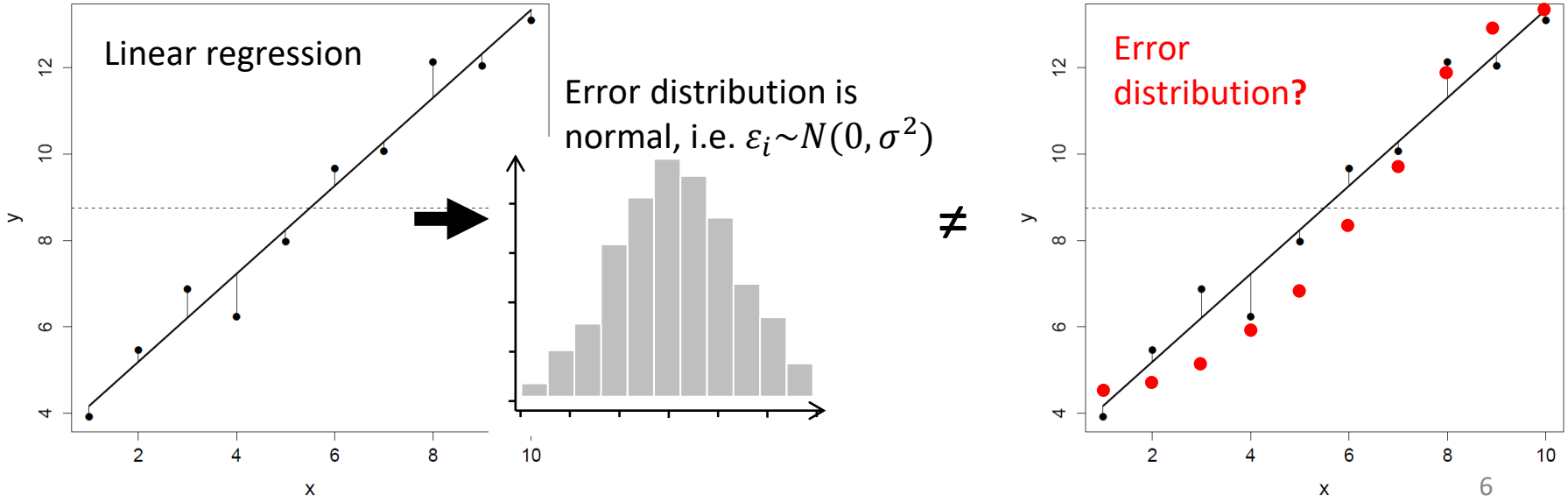
If not met:
Generalized Linear Model

If not met:
Linear Mixed-effects Model

If both are not met:
Generalized Linear Mixed-effects Model

Non-normal data

- Examples of non-normal data
  - Count data
  - Success and failure
- When you model non-normal data with a LM, the residuals will not be normally distributed, hence the model not valid.



Linear predictor and link function

- A GLM consists of:
  - A **probability distribution**, of **mean μ**
  - A **linear predictor**: $\eta_i = \alpha + \beta X_i$

    In other words: $some\ function\ of\ Y = \alpha + \beta X_i$
  - A **link function**: $g(\mu_i) = \eta_i$

    The link function provides the relationship between the linear predictor and the mean (μ) of the probability distribution function. It allows to obtain predicted values of the response variable: $\mu_i = g^{-1}(\eta_i)$

    In other words: $Y_i = inverse\ of\ some\ function\ of\ (\alpha + \beta X_i)$

Linear predictor and link function

- Note that the **Linear Model** (LM) is a particular case of the Generalized Linear Model (GLM) in which the link function is the **identity link function**:
  - The probability distribution is a **normal distribution** (of mean μ)
  - The linear predictor is still: $\eta_i = \alpha + \beta X_i$
  - And here the link function is: $g(\mu_i) = \mu_i = \eta_i$
  - In other words, we can directly predict the response variable as a linear combination of explanatory variables: $Y_i = \alpha + \beta X_i$

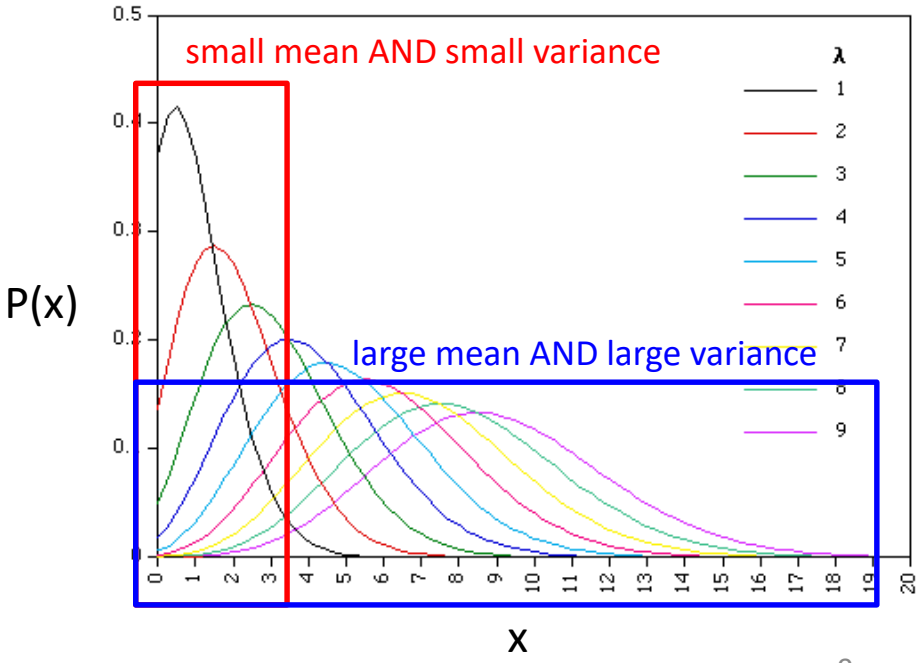
- Error distributions for GLMs belong to the **exponential family**:
  Normal, Poisson, Binomial, Inverse-Normal, Gamma, Negative Binomial…

The Poisson distribution

- A **discrete probability distribution** expressing the **probability of a number of events (x)** occurring in a fixed time/space interval.

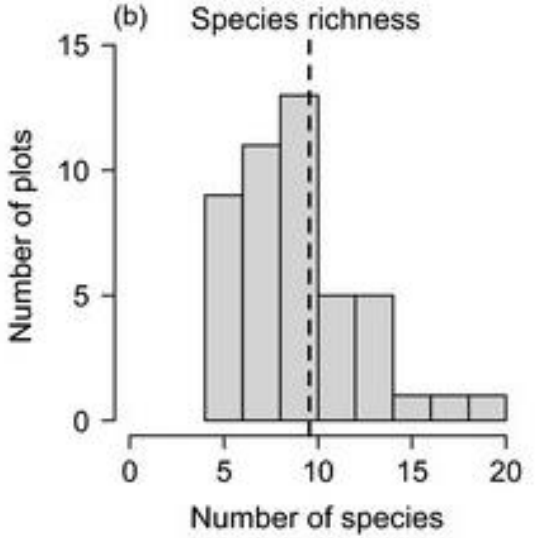
- Typical distribution of **count data**

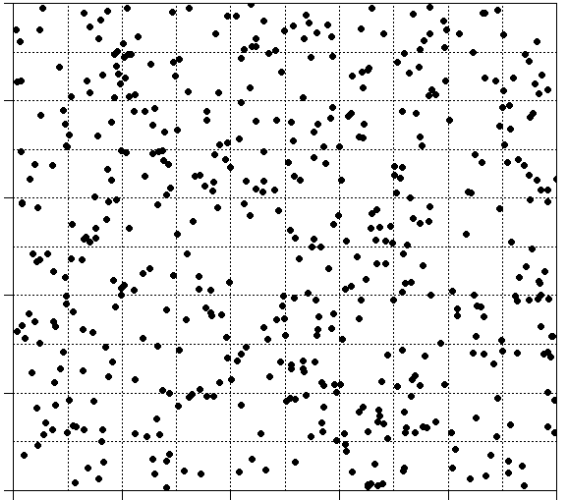
- Characteristics:
  - **The variance equals the mean (=λ)**
  - As the mean increases, the distribution gets closer to a normal distribution.

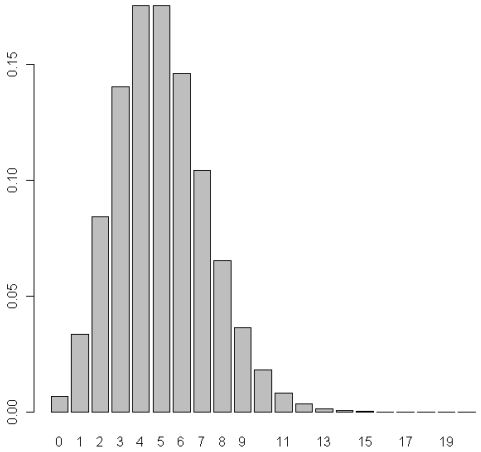


The Poisson distribution

- Examples
  - Number of plant species in vegetation plots
  - Number of raindrops per square







Poisson GLM

- For **Poisson GLMs**, we transform the linear predictor using the **log link function**, to predict the number of Y events (e.g. species, pollinator visits...).
  - The probability distribution is a **Poisson distribution** (of mean μ)
  - The linear predictor is still: $\eta_i = \alpha + \beta X_i$
  - But the link function is now: $g(\mu_i) = \ln(\mu_i) = \eta_i$
- To predict the response variable, we thus use: $Y_i = e^{\alpha+\beta X_i}$
- Because Y is modelled as an exponential, it is **always positive** (which is useful, as we can't have negative counts!).

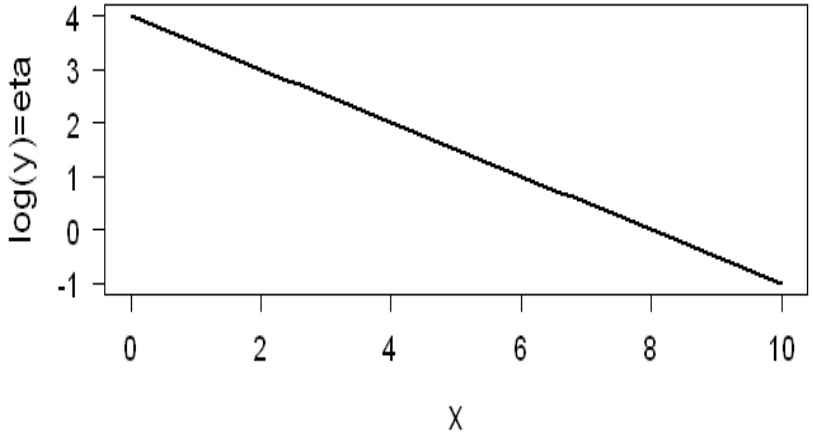
Poisson GLM

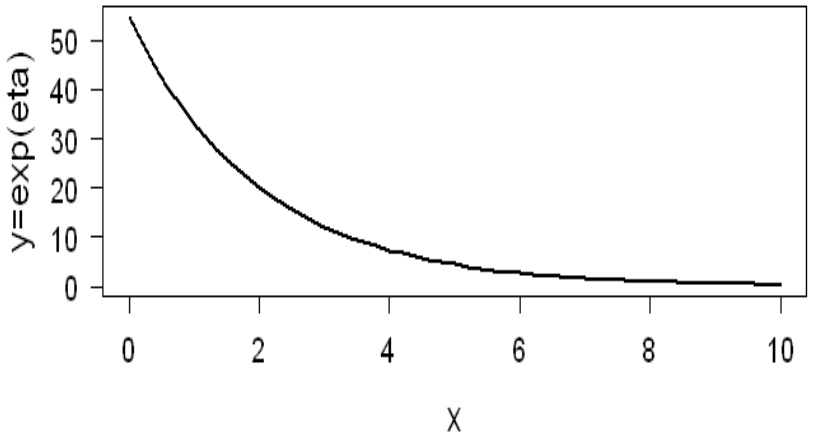
Linear predictor

$$\log(\lambda) = \alpha + \beta X_i$$

Response

$$\lambda = e^{\alpha+\beta X_i}$$

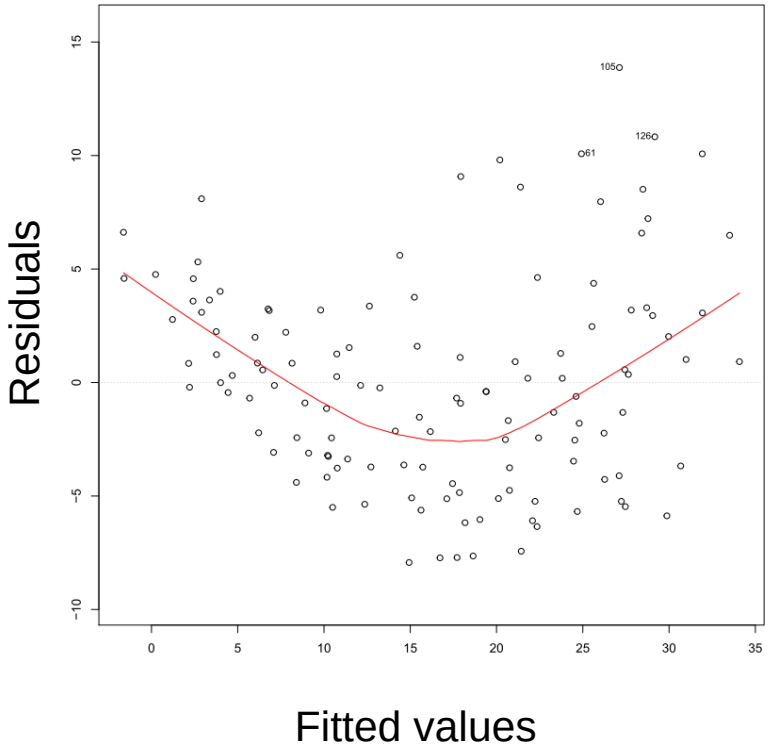


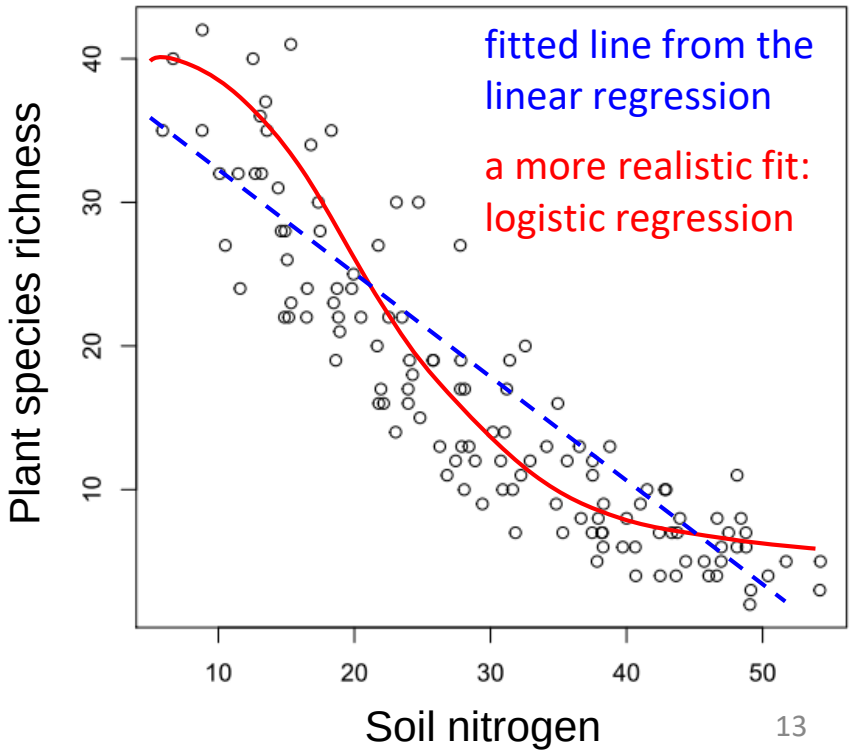


The relationship is linear on the scale of the linear predictor.

Example of a Poisson GLM

- Is there a relationship between soil nitrogen and plant species richness?
  - You successfully accounted for non-independence of the observations…
  - … but you found an odd pattern for the plot of residuals vs. fitted values:





Poisson GLM in R

Poisson GLM

To run a Poisson GLM without random effects, first create the model:
glmPoisson<-glm(Response~Explanatory, data=Data, family="poisson")

Then get the summary table
summary(glmPoisson)

Poisson GLM in R

Let's go through the Poisson GLM summary table

```
Call:
glm(formula = Species_richness ~ N_mg_g_soil, family = "poisson",
    data = soiln)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7203  -0.7692  -0.1295   0.5280   2.7124

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.073083   0.050281   81.01   <2e-16 ***
N_mg_g_soil -0.048557   0.001978  -24.55   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 788.78  on 125  degrees of freedom
Residual deviance: 121.27  on 124  degrees of freedom
AIC: 684.65

Number of Fisher Scoring iterations: 4
```

Poisson GLM in R

Let's go through the Poisson GLM summary table

```
Call:
glm(formula = Species_richness ~ N_mg_g_soil, family = "poisson",
    data = soiln)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7203  -0.7692  -0.1295   0.5280   2.7124

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.073083   0.050281   81.01   <2e-16 ***
N_mg_g_soil -0.048557   0.001978  -24.55   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 788.78  on 125  degrees of freedom
Residual deviance: 121.27  on 124  degrees of freedom
AIC: 684.65
```

The Poisson GLM also indicates that plant species richness is significantly negatively related to soil nitrogen.

Poisson GLM in R

Let's go through the Poisson GLM summary table

```
Call:
glm(formula = Species_richness ~ N_mg_g_soil, family = "poisson",
    data = soiln)
```

In a GLM, since we do not use Least Squares, we do not have a coefficient of determination R^2. Instead, we have a null deviance and a residual deviance.

The residual deviance is like the residual sum of squares in a linear regression. It is the sum of the squared deviance residuals: the smaller the better. A deviance residual is the difference between an observed and fitted value.

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 788.78  on 125  degrees of freedom
Residual deviance: 121.27  on 124  degrees of freedom
AIC: 684.65

Number of Fisher Scoring iterations: 4
```

Poisson GLM in R

Let's go through the Poisson GLM summary table

```
Call:
glm(formula = Species_richness ~ N_mg_g_soil, family = "poisson",
    data = soiln)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7203  -0.7692  -0.1295   0.5280   2.7124

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
```

We also have Akaike's Information Criterion (AIC) value. It is based on something called the 'likelihood' of the model. We want models that maximize the likelihood, and thus minimize the AIC.

AIC can be used for model comparison.

```
Residual deviance: 121.27  on 124  degrees of freedom
AIC: 684.65

Number of Fisher Scoring iterations: 4
```

Poisson GLM: model validation

- Because Poisson GLMs allow for larger spread of residuals for larger fitted values, **it doesn't make sense to look at residuals as observed minus fitted values**.

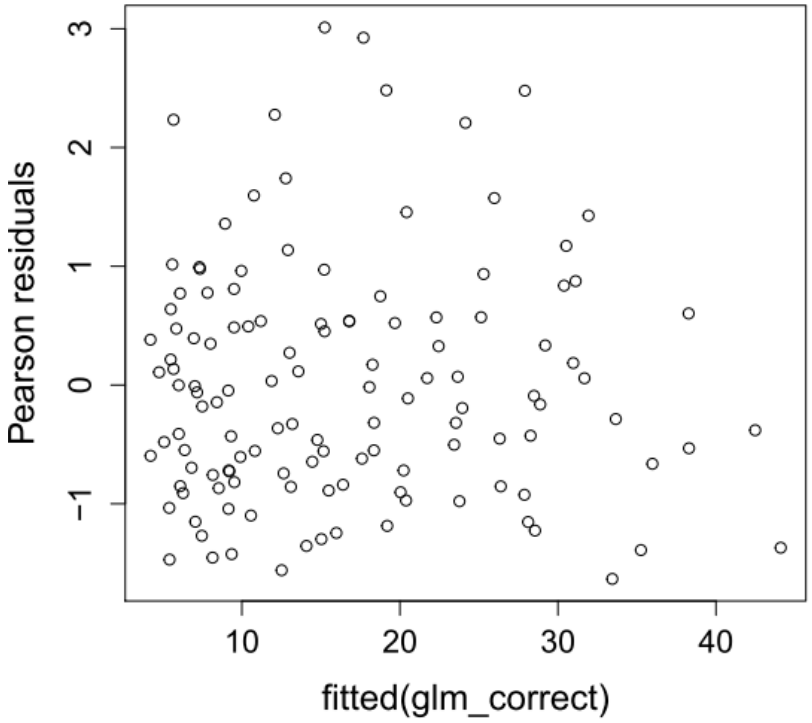
- For non-Gaussian (=non-normal) GLMs, we use **Pearson residuals**:

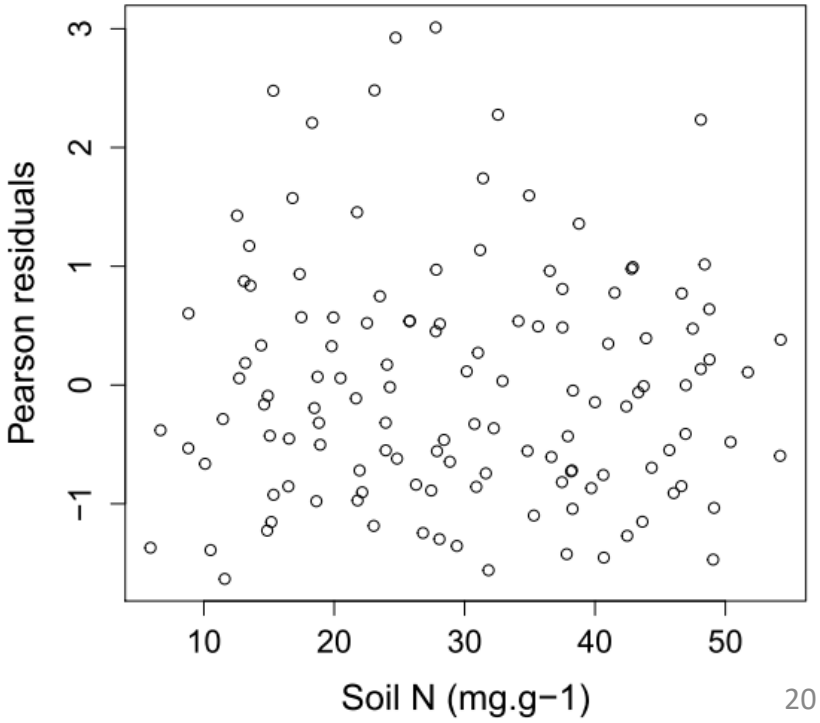
$$Pearson\ residuals = \frac{Y_i - \widehat{\mu_i}}{\sqrt{\widehat{\mu_i}}}$$

The Pearson residuals scale the (observed – fitted) differences by dividing by the square-root of the fitted value.

Poisson GLM: model validation

- No patterns should be visible when we plot the Pearson residuals
  - Against the fitted values
  - Against the explanatory variable(s)



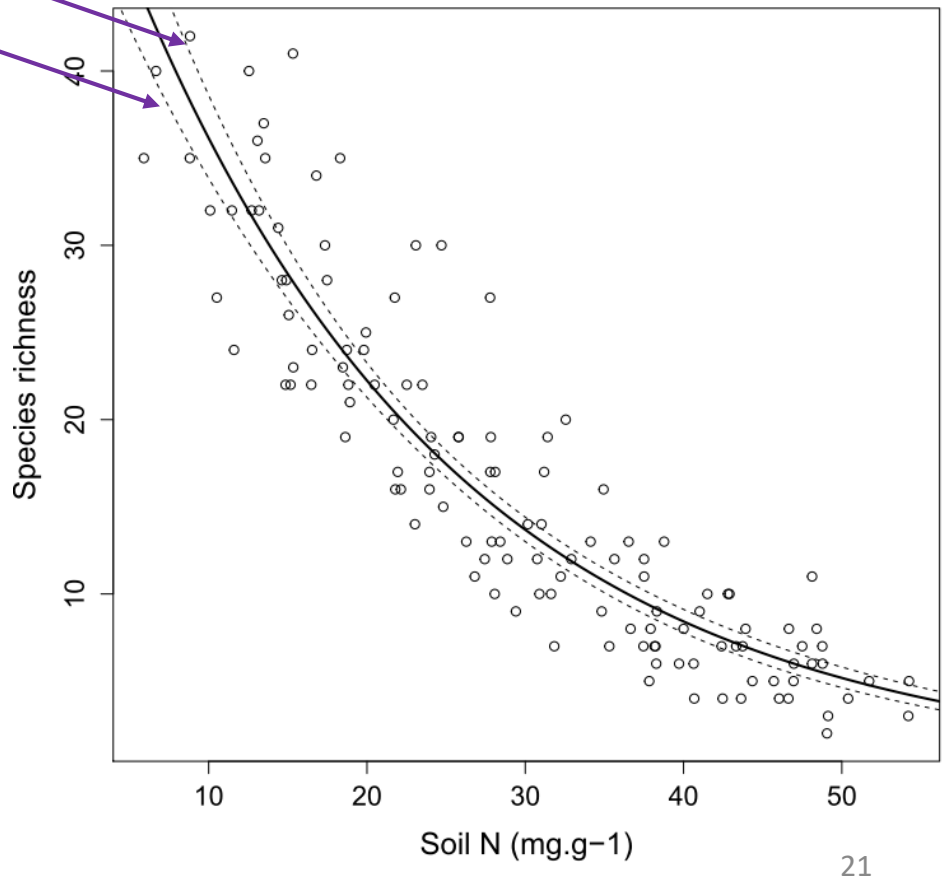


Plotting a Poisson GLM

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.073083   0.050281    81.0
soilN       -0.048557   0.001978   -24.5
---
```

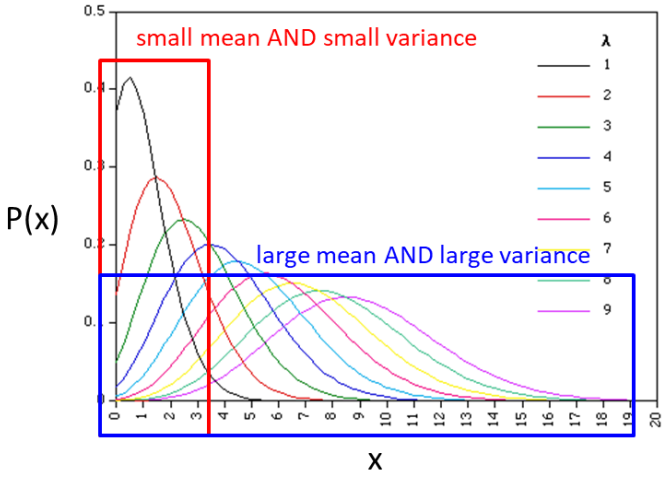
Fitted line:

$$y_i = e^{4.073 - 0.049 * x_i}$$



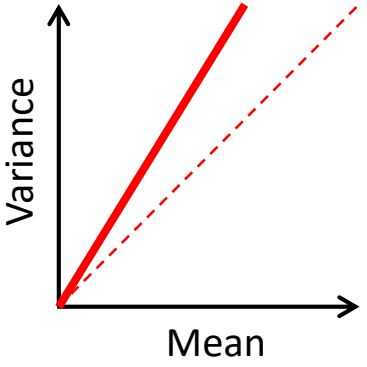
Overdispersion

- Remember that in a Poisson GLM **the mean should equal the variance** ($\mu=\sigma^2$).



- Dispersion can be characterized by the **dispersion parameter, ρ**.
  - ρ should be 1 if $\mu=\sigma^2$
  - **If ρ>1**, i.e. the variance exceeds the mean, then we have **overdispersion**.

    Overdispersion can be thought of as **extra variation in the response** that cannot be captured by a Poisson GLM.
  - Note that underdispersion may also occur, although more rarely with ecological data at least.



Overdispersion

- If residual deviance > residual degrees of freedom, then ρ > 1,
  i.e. there is overdispersion. **As a rule of thumb, up to ~ 1.5 is ok**.

- Accounting for overdispersion is important, as it increases standard errors.
  Ignoring overdispersion can result in Type 1 errors (false positives).

- What can you do in case of overdispersion?
  - Use a '**quasi-poisson**' GLM
    It calculates ρ based on our mean and variance ($variance = \rho * \mu$), still applying the Poisson distribution.
  - Use a **negative binomial** distribution,
    where we estimate the variance as: $variance = \mu + \frac{\mu^2}{k}$
    (k is an estimated dispersion parameter)

Time for an exercise

Check if your Poisson GLM suffers from overdispersion.

More advanced exercise!

- So far, you have learned
  - How to do a linear mixed-effects model (using `lmer()`)
  - How to do a Poisson generalized linear model (using `glm()`)

- Now, find out how to do a **Poisson generalized linear mixed-effects model** in order to correctly model the effect of soil nitrogen on plant species richness!

Tip: you need to use the function `glmer()` from the package 'lme4'.

The binomial distribution

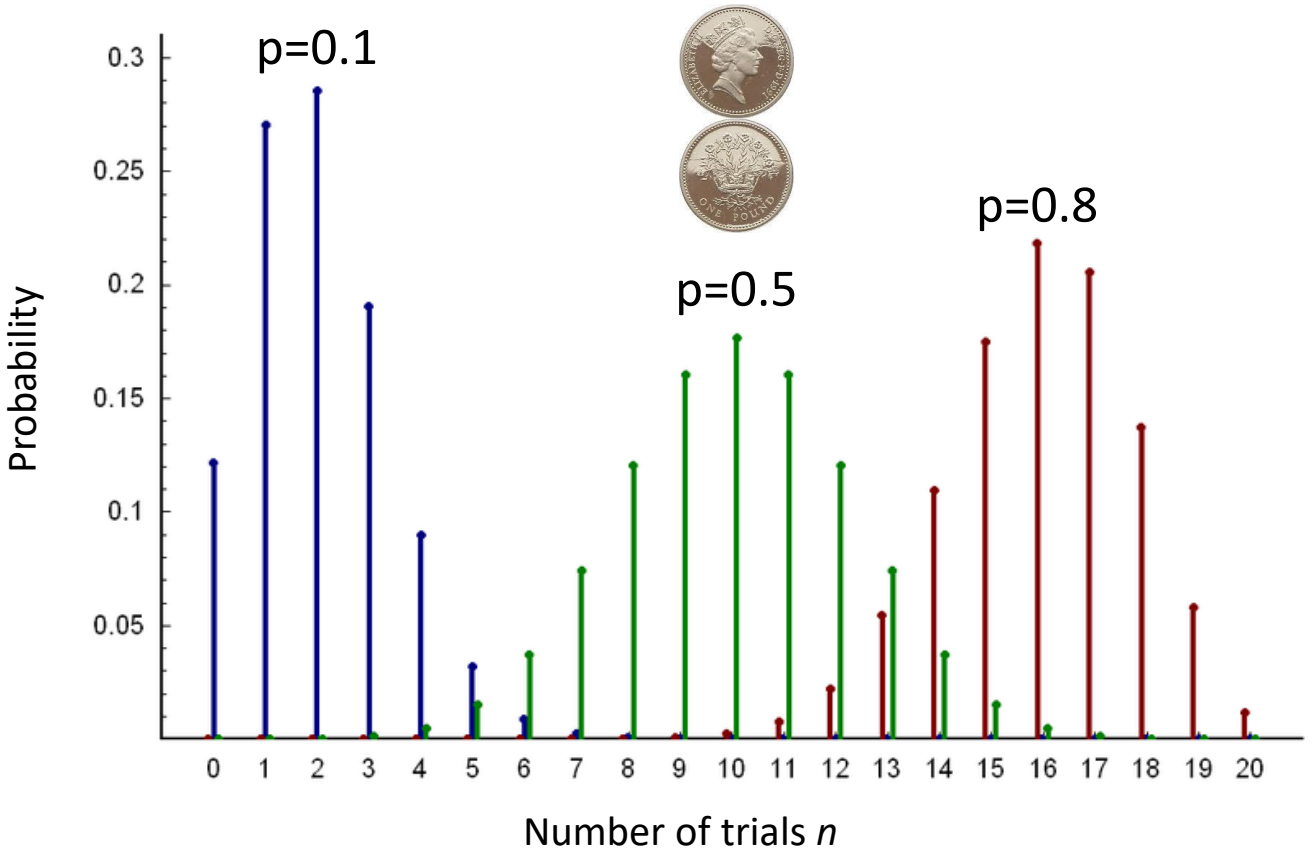
- A **discrete probability distribution** of the **number of successes** in a sequence of independent yes/no trials.

- The classic example of the coin!
  Let's imagine we flip a coin 20 times.
  - What is the probability of getting a head once?
    0.5
  - How many heads do we expect in total?
    10

- **Mean=np, variance=np(1-p)**
  where n=number of trials, and p=probability of 'success'

The binomial distribution

- Examples with n=20 and different probabilities p:



The binomial distribution

- **We do not know the probability of success *p***, so **we have to model it** as a function of explanatory variables using a GLM.

- Examples in ecology
  - Probability of germination (seeds)
  - Probability of herbivory (leaves)
  - Probability of survival to fire
  - Probability of reproduction
  - …

The Bernoulli distribution

- In a **Bernoulli distribution**, we only have **one event: n=1**.
  In other words, x ~ Binomial(p, 1) = x ~ Bernoulli(p)

- We **model the probability of success or failure**.
  (e.g. the probability of alien species establishing)

- However, in practice, GLMs for binomial and Bernoulli distributions work in the same way.

Probability and odds

- Probability is bounded by 0 and 1.

- **Odds** express probabilities without an upper bound: $\boldsymbol{odds} = \frac{\boldsymbol{p}}{\boldsymbol{1-p}}$

- If we take the log of odds, our probability is no longer bounded by 0 and 1.

Binomial GLM

- For **binomial GLMs**, we transform the linear predictor using (most often) the **logit link function**, to predict the probability of successes.
  - The probability distribution is a **binomial (or Bernouilli) distribution**
  - The linear predictor is still: $\eta_i = \alpha + \beta X_i$
  - But the link function is now:

$$g(\mu_i) = \text{logit}(\mu_i) = \log(odds) = \log\left(\frac{p}{1-p}\right) = \eta_i$$

- To predict the response variable, we thus use: $Y_i = \frac{e^{\alpha+\beta X_i}}{1+e^{\alpha+\beta X_i}}$

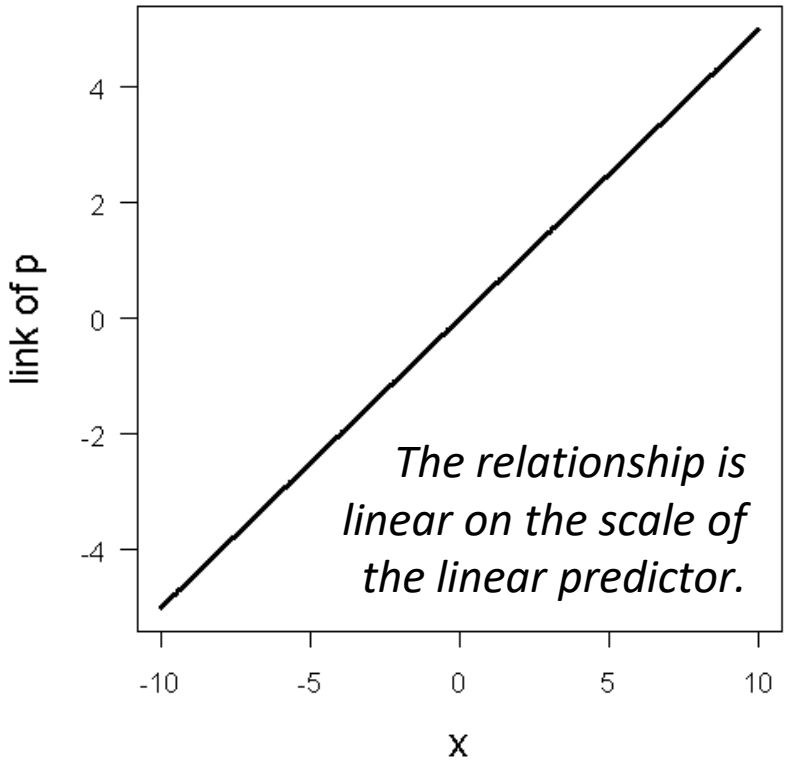
Binomial GLM

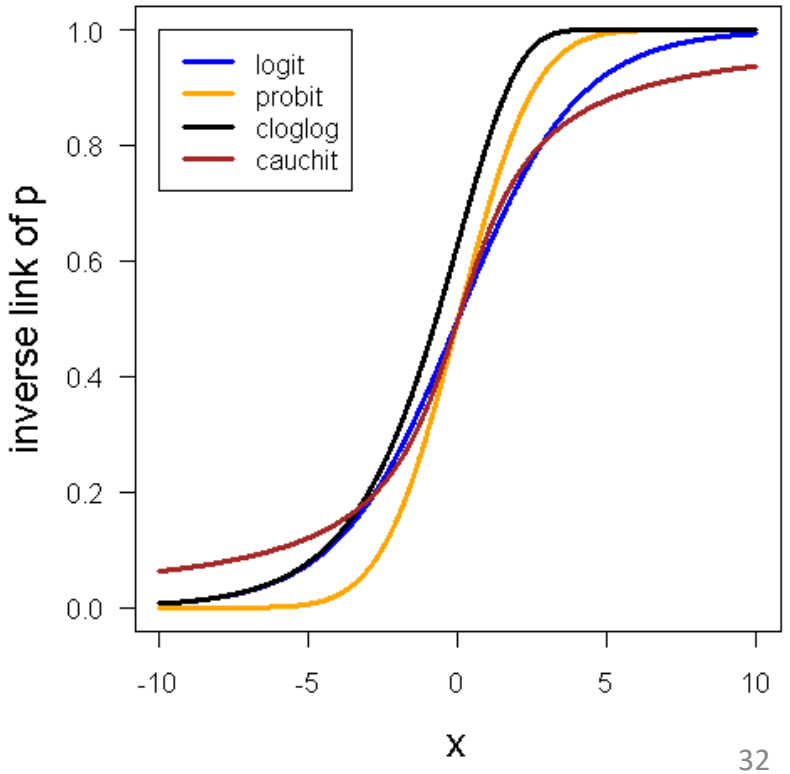
Linear predictor

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \alpha + \beta X_i$$

Response: Probability

$$p = \frac{e^{\alpha+\beta X_i}}{1+e^{\alpha+\beta X_i}}$$





Binomial GLM in R

Binomial GLM with Bernoulli data

Just as usual, first create your model:
glmBern=glm(Response~Explanatory, data=Data, family="binomial")

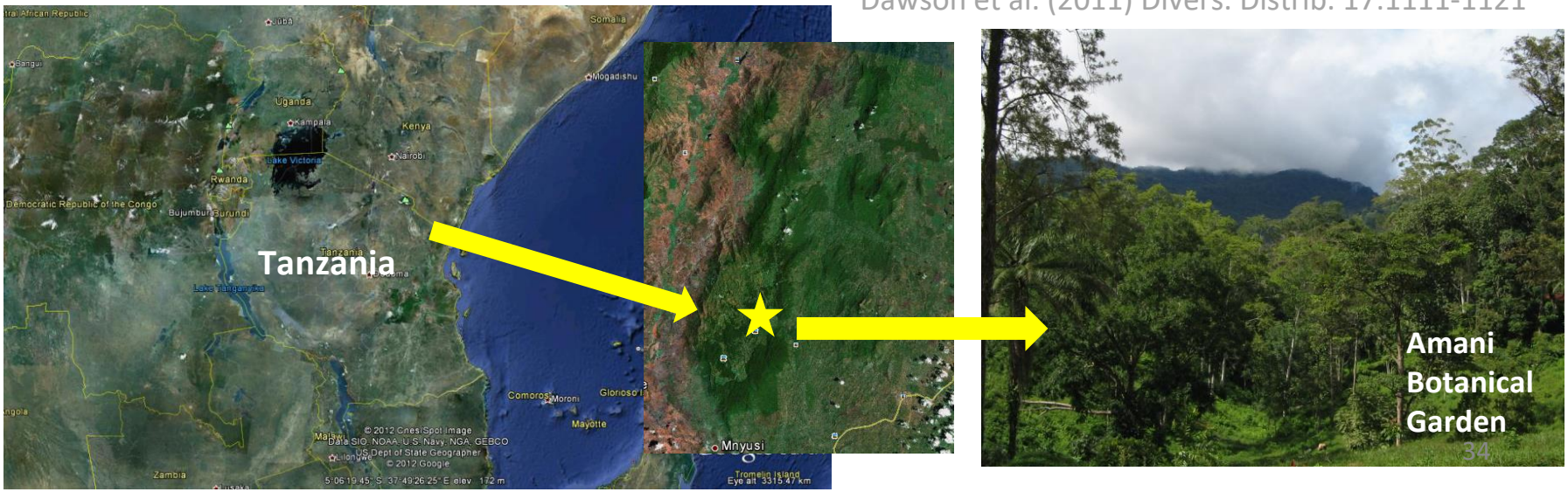
Then get the summary table:
summary(glmBern)

Exercise with a new dataset

Load the “naturalizationABG.txt” file. The data for each woody plant:

- **Species:** Species name
- **nat:** Naturalization success (1) or failure (0)
- **no_plants:** Number of plants originally planted in Amani Botanical Garden
- **surviving:** Survived in ABG (1) or not (0)
- **growth.form:** P=palm vs. S=Shrub vs. T=Tree
- **Tree:** S=Shrub vs. T=Tree or palm tree
- **origin:** continent of origin

Dawson et al. (2011) Divers. Distrib. 17:1111-1121



Exercise with a new dataset

(1) Try to make your own binomial GLM to answer the following question:

For those species that have managed to survive in ABG, is the probability of naturalization outside ABG positively related to the number of plants originally planted (i.e. propagule pressure)?

Tips:

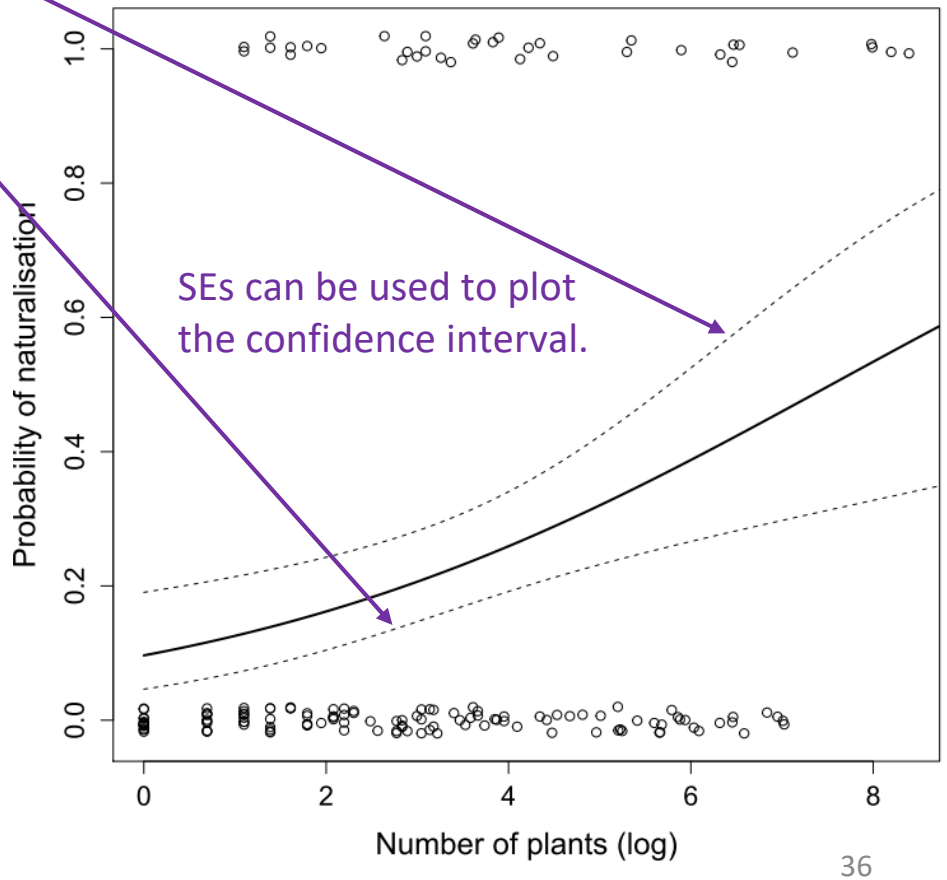
1) Examine variables you will need
2) Subset the data, if needed
3) Do a binomial GLM

Plotting a binomial GLM

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.23656    0.40294  -5.551 2.85e-08 ***
plants       0.29655    0.09257   3.204  0.00136 **
---
```

Fitted line:

$$y_i = \frac{e^{-2.237+0.297*x_i}}{1+e^{-2.237+0.297*x_i}}$$



Proportional data

- With **proportional data**, we need to **represent the data as the number of 'successes' and the number of 'failures'**, using `cbind()`.
  - Example: you sow 10 seeds of 30 species, you record how many seeds germinate for each species, and you then model as a response variable what *proportion* has germinated, out of 10 seeds.

Overdispersion

- Just like in a Poisson GLM, there may be overdispersion:
  - In a binomial GLM with proportional data, **we expect the variance to equal np(1-p)**, where n = number of trials, and p = proportion of 'successes'.
  - If the variance is **bigger than np(1-p)**, then we have **overdispersion**. In that case, we can use a '**quasi-binomial**' GLM, modelling a dispersion estimate ρ such that: $variance = \rho * np(1-p)$
  - Again, there may be underdispersion, but this is less common.

Proportional data in R

Binomial GLM with proportional data

In R, you need to indicate in the formula of your model both the number of successes:
glmProp=glm(cbind(successes,failures)~Explanatory,data=Data,family="binomial")

Then, as usual, get the summary table:
summary(glmProp)

Overview of models

	Independent observations	Dependent obervations

Linear relation
Error distribution normal
	Linear model (LM)
	
	`aov()`
	`lm()`
		Linear mixed-effects model (LMM)
		
		`nlme::lme()`
		`lme4::lmer()`

Linear relation Error distribution different	Generalized linear model (GLM) `glm()`	Generalized linear mixed-effects model (GLMM) `lme4::glmer()`

Acknowledgements

People:

Noelie Maurel
Wayne Dawson
Fränzi Körner

International Max Planck Research School for Organismal Biology

Ecology Konstanz

Institute for Health Metrics and Evaluation

TReND

The Company of Biologists

Supported by

The Company of Biologists

Development

Journal of Cell Science

Journal of Experimental Biology

Disease Models & Mechanisms

Biology Open