Generalized Linear Models

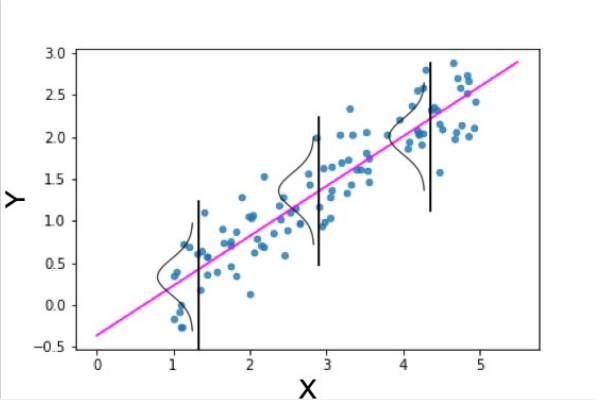
Underlying hypotheses of linear models

ANOVA

$$Y_{ik} = \mu + \alpha_i + \epsilon_{ik}$$

Linear regression

$$Y_{ik} = \beta_0 + \beta_i X_i + \epsilon_{ik}$$



Predicted values (Y variable) results from a linear combination (addition) of explainatory variables, X variable(s)

Distribution of the residuals is normal, centered on 0 and their variances are homogenous.

Examples of variables measured on plants

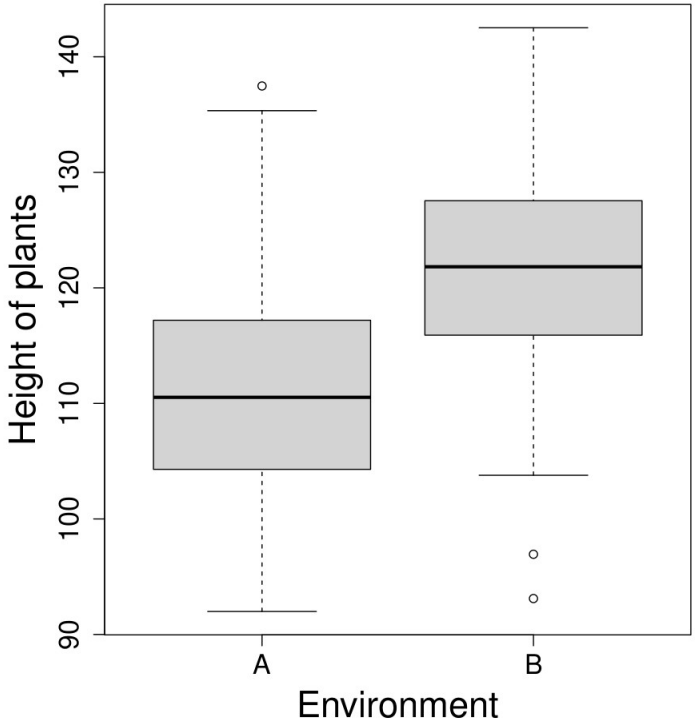
Let’s consider 100 plants grown in two different environments : A and B

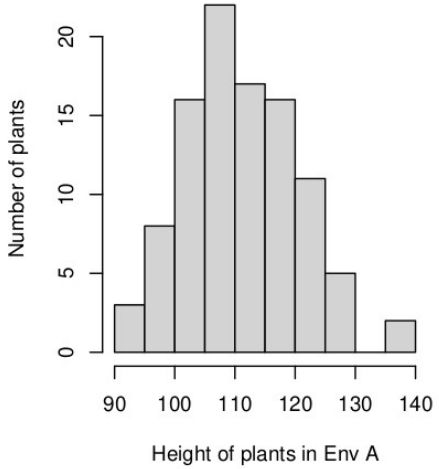
Several variables can be evaluated :

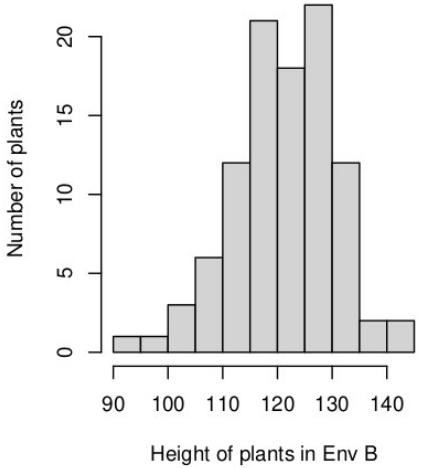
- Height of a plant
- Number of seeds produced by a plant
- Number of seeds that germinated among 100 sowed seeds

Type of variables to model : gaussian variables

Y is the random variable representing the height of plants (1)







$$Y_{ik} = \mu + \alpha_i + \epsilon_{ik}$$

Limits of linear models

Linear models are not appropriate in two mains situations :

- when the range of the variable is limited, for example

  - X>0 for count data :
  Number of seeds produced by a plant

  - 0<X<1 for probabilities :
  Number of seeds that germinated out of 100 sowed seeds

- the residual variance depends on the mean

Data transformation vs GLM

A solution could be the transformation of the response variable (log transformation for example) before applying a linear model.

$$\log\left(y_{ik}\right)=\mu+\alpha_i+\epsilon_{ik}$$

$$E\left(\log\left(Y_i\right)\right)=\mu+\alpha_i$$

This solution is not always satisfying…

- The transformation has to improve linearity
- The homogeneity of variance needs to be improved
- Transformation may not be defined for each value of the initial variable (ex : log(0))
-

With GLM, the transformation is applied to the mean of the variable

$$\log\left(E\left(Y_i\right)\right)=\mu+\alpha_i$$

Linear Models vs Generalized Linear Models

ANOVA $Y_{ik} = \mu + \alpha_i + \epsilon_{ik}$

Linear models

Linear regression $Y_{ik} = \beta_0 + \beta_i X_i + \epsilon_{ik}$

GLM models extend the linear models to variables that do not follow a normal distribution

The GLM **linear predictor** $\nu_i = \beta_0 + \beta_i X_i$

Generalized linear models

Link function describes how the mean depends on the linear predictor

$$\nu_i = g(\mu_i)$$

Structure of the errors

A variance function that describe how the variance depends on the mean

$$var(Y_i) = V(\mu)$$

Type of variables to model : count variables

For 1 trial, the event follows a Bernouilli law X can take two values : 0 (fail) or 1 (success)

P(X=1) = p
P(X=0) = 1-p

For n trials, the number of success in the n-sample in a count variable

$$Z = \Sigma_{i=1}^{n} X_i$$

n-sample with n unknown and may be infinite : **Poisson distribution**

Fixed n-sample, sampling with replacement : **Binomial distribution**

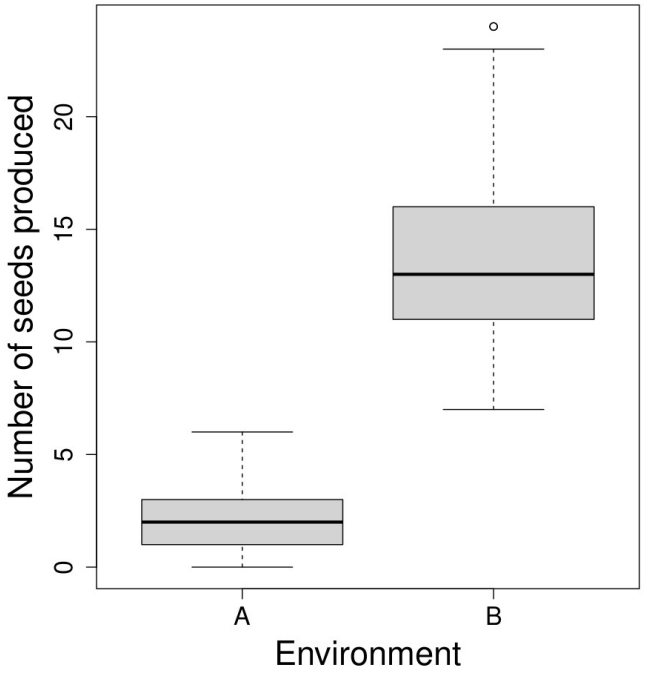
Fixed n-sample, sampling without replacement (probabilities change after each trial) :
Hypergeometric distribution

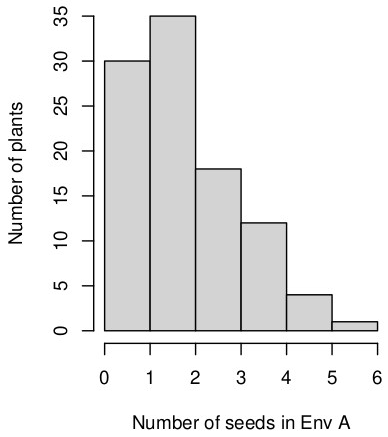
Stop afet r fails (or success) : **Negative Binomial distribution**
(case in PCR amplification procedure, a maximum can be reached)

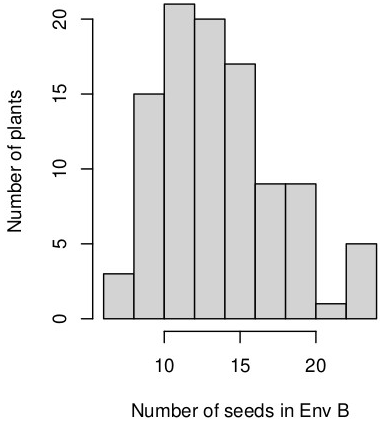
Type of variables to model : count variables

Y is the ramdom variable representing the number of seeds produced by a plant

$$Y_i \sim Poisson(\lambda_i)$$







Poisson with λ parameter

(equal to the mean **and** the variance, variance will increase with fitted values !)

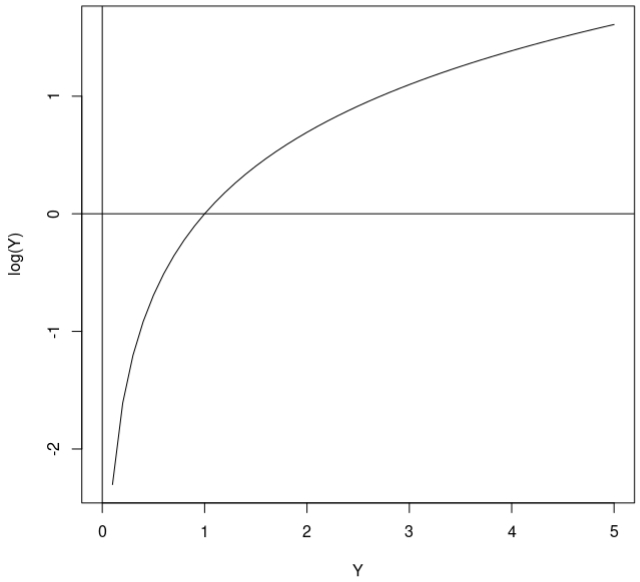
Poisson distribution $Y_i \sim Poisson(\lambda_i)$

The GLM **linear predictor** $\nu_i = \beta_0 + \beta_i X_i$

Link function describes how the **mean** depends on the linear predictor

$$\nu_i = g(\mu_i)$$

$$g(\mu_i) = \log(\mu_i)$$



Log transformation
allows to extend the range from [0;+∞] to]-∞ ; +∞[

Structure of the errors
A variance function that describe how the variance depends on the mean

$$var(Y_i) = V(\mu)$$

$$Y_i \sim Poisson(\lambda_i)$$

$$E(Y_i) = \lambda_i \qquad var(Y_i) = \lambda_i$$

So, the variance function is

$$V(\mu_i) = \mu_i$$

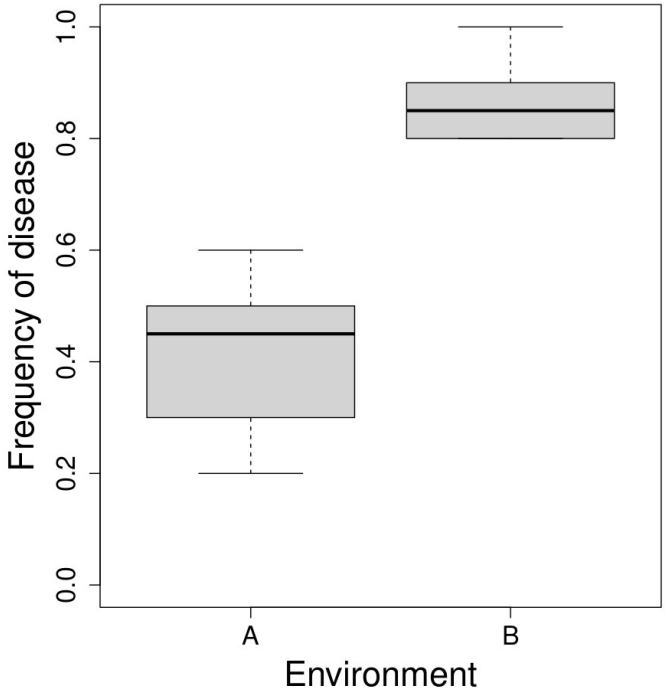
Type of variables to model : frequencies

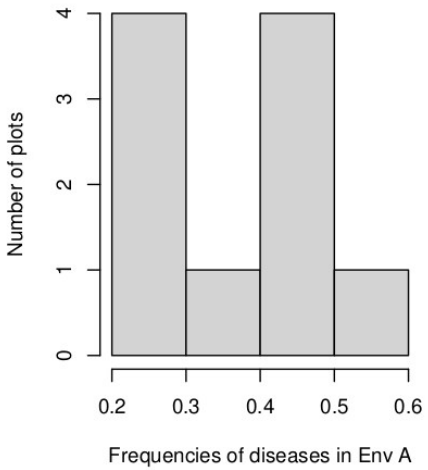
Y is the random variable that describe the number of plants that germinate among 100

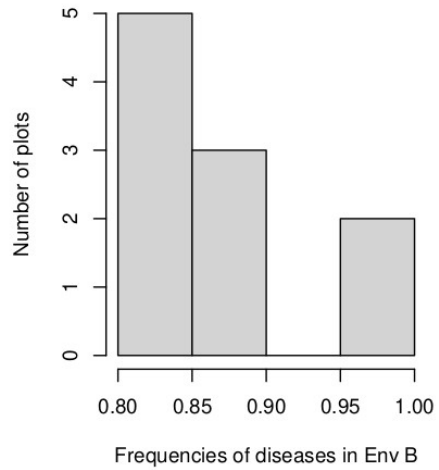
$$Y_i \sim Binomial\left(n=100, p_i\right)$$

$\frac{Y_i}{n}$ is an estimator of p_i

H0 : $p_A = p_B$ the germination rate is the same in both environments
H1 : $p_A \neq p_B$ the germination rate is different in both environments







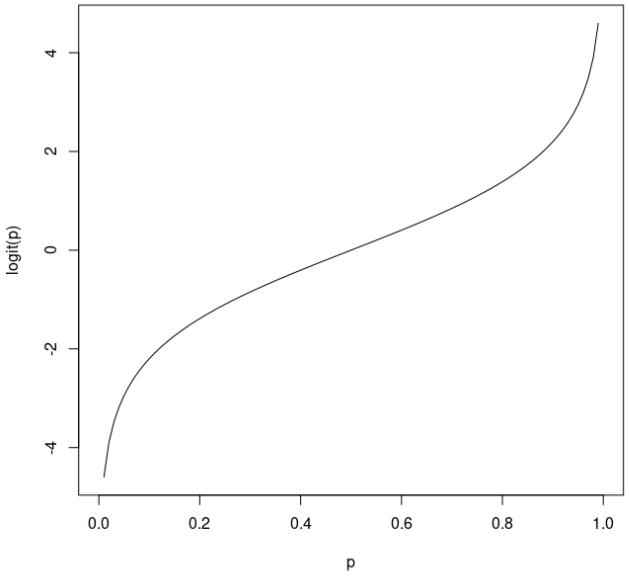
Binomial distributions $Y_i \sim Binomial(n_i, p_i)$

The GLM **linear predictor** $\nu_i = \beta_0 + \beta_i X_i$

Link function describes how the **mean** depends on the linear predictor

$$\nu_i = g(\mu_i)$$

$$g(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$$



Logit transformation
allows to extend the range from [0;1] to]-∞ ; +∞[

Structure of the errors
A variance function that describe how the variance depends on the mean

$$Y_i \sim Binomial(n_i, p_i)$$

$$E(Y_i/n_i) = p_i$$

$$var(Y_i/n_i) = \frac{1}{n_i} p_i(1-p_i)$$

So, the variance function is

$$V(\mu_i) = \mu_i(1-\mu_i)$$

Type of variables to model : count variables

Fixed n-sample, sampling with replacement : **Binomial distribution**

n-sample with n unknown and may be infinite : **Poisson distribution**

Fixed n-sample, sampling without replacement (probabilities change after each trial) : **Hypergeometric distribution**

Stop afet r fails (or success) : **Negative Binomial distribution**
(case in PCR amplification procedure, a maximum can be reached)

Law	parameters	Expectancy	Variance
Binomial	$\mathcal{B}(n,p)$	$n.p$	$n.p.(1-p)$
Poisson	$\mathcal{P}(\lambda = n.p)$	λ	λ
Hypergeometrical	$\mathcal{H}(n,p,A)$	$n.p$	$n.p.(1-p)\frac{A-n}{A-1}$
Negative binomial	$\mathcal{NB}(r,1-p))$	$r\frac{p}{1-p}$	$r\frac{p}{(1-p)^2}$

Examples of generalized linear models

The exponential family functions available in R are

- binomial(link = "logit")
- poisson(link = "log")
- gaussian(link = "identity")

Type of response and errors variables	Law of response variable and errors	Response modeled	Link function	
Quanti. continous] -inf ; + inf [	Gaussian	Variable itself	identity	$g(\mu)=\mu$
Count Integer [0 ; +inf]	Poisson	Log of the mean	log	$g(\mu)=\log(\mu)$
Binary (0/1) Integer [0;1]	Binomial	Log of the chance ratio	logit	$g(\mu)=\log\left(\frac{\mu}{1-\mu}\right)$

But also :

- Gamma(link = "inverse")
- inverse.gaussian(link = "1/mu 2 ")

GLM interpretation

In R output :

- **Coefficients** table indicate the significance of the effects of the model

- **Deviances** allows to evaluate **the quality of a model**, its ability of the model to predict the response variable
  - Null model (wihout any explanatory variable, only the intercept)
  - Residual deviance (low if the model is able to perform good prediction)

$$\chi^2_{model} \sim \chi^2_{p\,ddl}$$

Chi^2_{model}= Null model - Residual deviance

→ follows a Chi^2 law at *p* degrees of freedom (where p = number of predictor variables)
→ a pvalue can be calculated and indicates the significance of the model

- **AIC** can be used to **compare several models** (low AIC for best models) : AIC = 2K – 2ln(L)

→ K is the number of parameters of the model
→ ln(L) is the log-likelihood of the model

Modelling of RNAseq counts

1) Variable to model : Count data Y

2) Distribution law and its parameters

Follows a negative binomial distribution characterized by a mean λ and a variance ϕ (indices to determine)

$$NB(\lambda, \phi)$$

3) List of the effects impacting the gene count

The log of the mean $\log(\lambda)$ will be modelled

Describe each of the effects