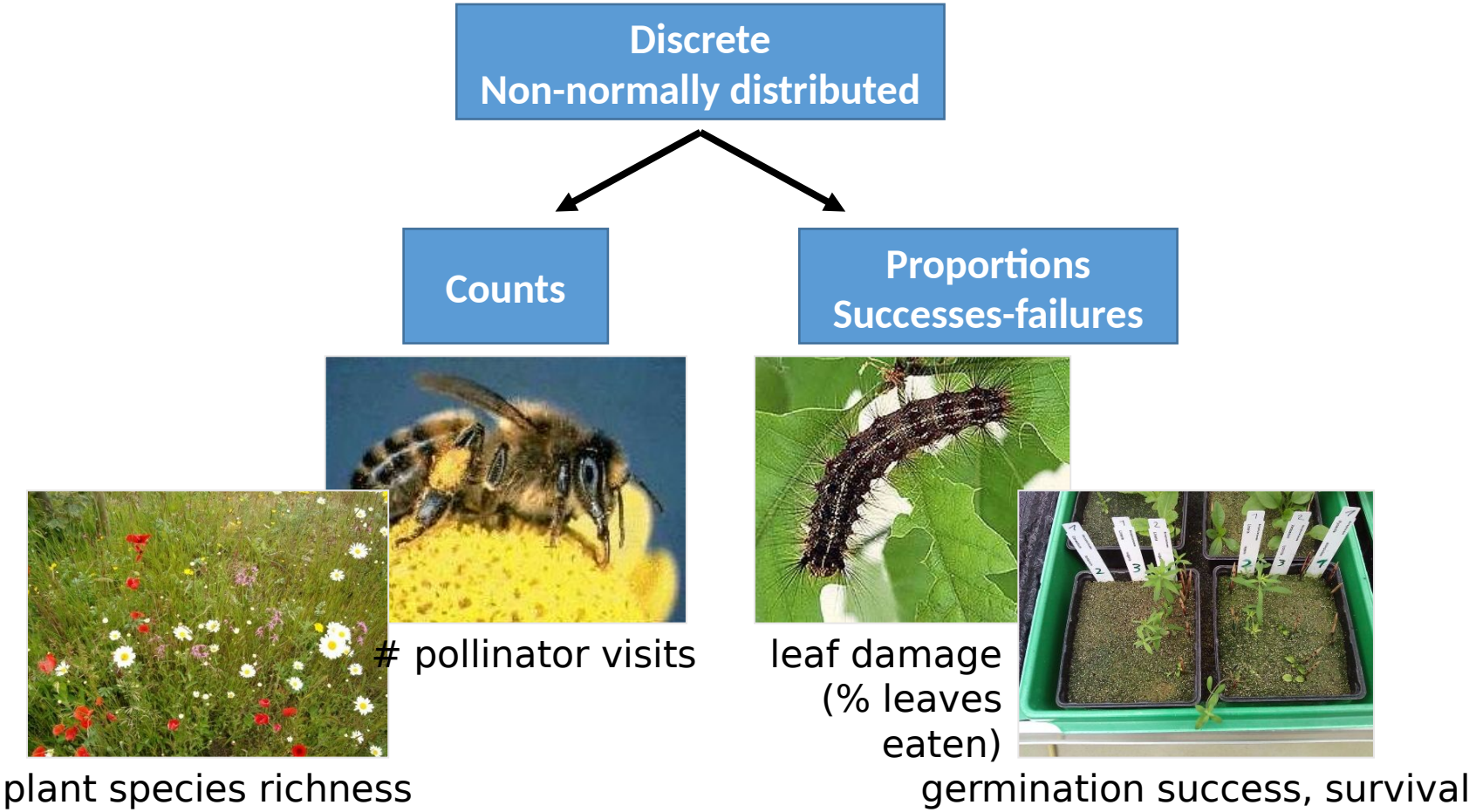
Introduction to statistics in R

Generalised linear models

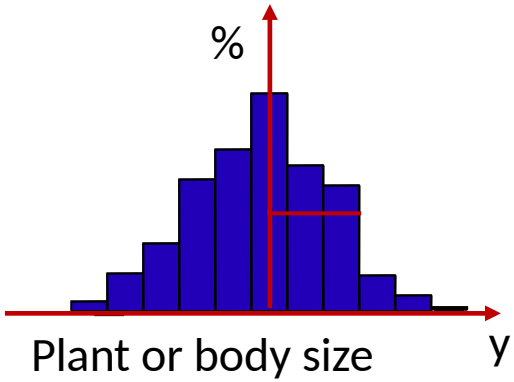
on-normal data in ecology

- Non-normal data often falls into two categories:

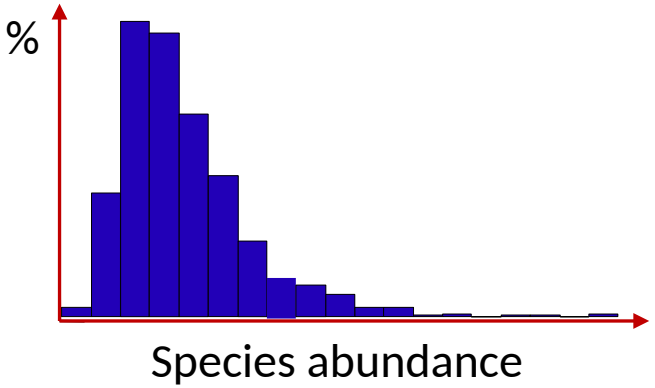


Data distributions

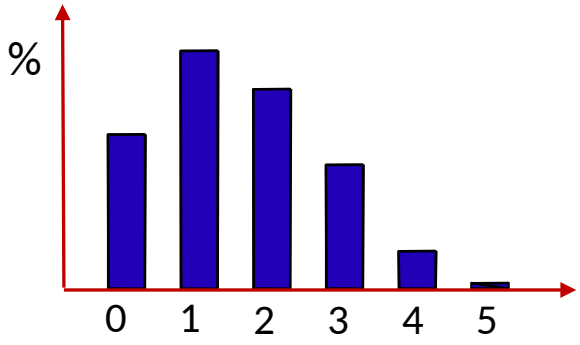
Normal distribution



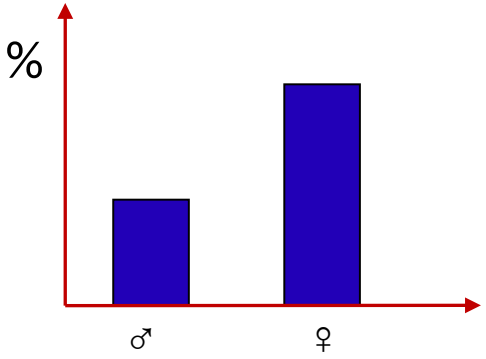
Log-Normal distribution



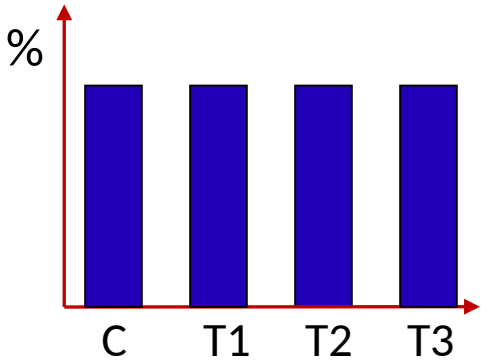
Poisson distribution



Binomial distribution



Uniform distribution



Differences compared to linear model

1. A new method to estimate parameters
   Maximum likelihood

2. A new test statistics
   Likelihood ratios or deviance (difference between log-likelihoods)

3. A variance function which specifies how the variance **changes with the fitted values**

4. A link function which transforms the response variable

Data transformation or glm?

- In some situations a response variable can be transformed to improve linearity and homogeneity of variance so that a linear model can be applied.

- Drawbacks
  - response variable changes
  - transformation must simulateneously improve linearity and homogeneity of variance

Generalised linear model GLM

Data are may come from different type of distributions

$$Y_i = f(a + b \cdot X_i) + \varepsilon_i$$

$$\varepsilon_i \sim Poisson(\lambda)$$

$$\varepsilon_i \sim Binom(\pi)$$

$$\varepsilon_i \sim Gamma(a, b)$$

In R these distributions are called families

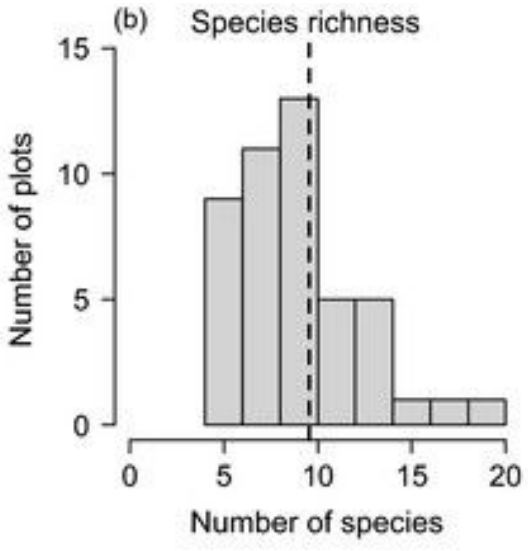
`gaussian`	For normal distribution
`binomial` `quasibinomial`	For binomial distribution
`poisson` `quasipoisson`	For Poisson distribution
`Gamma`	For gamma distribution

Poisson distribution

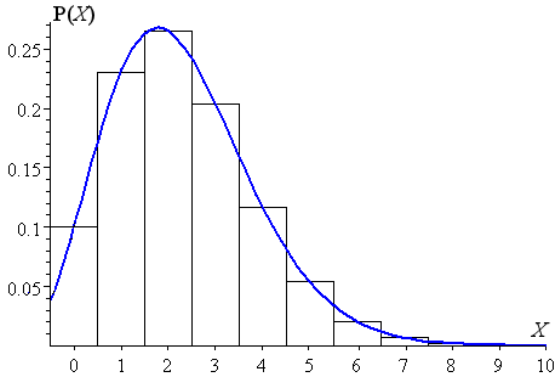
- How real data may look like:

Counts

plant species richness



Follows a **Poisson distribution**



Myers et al. (2015) J. Ecol. 103:1291-1299

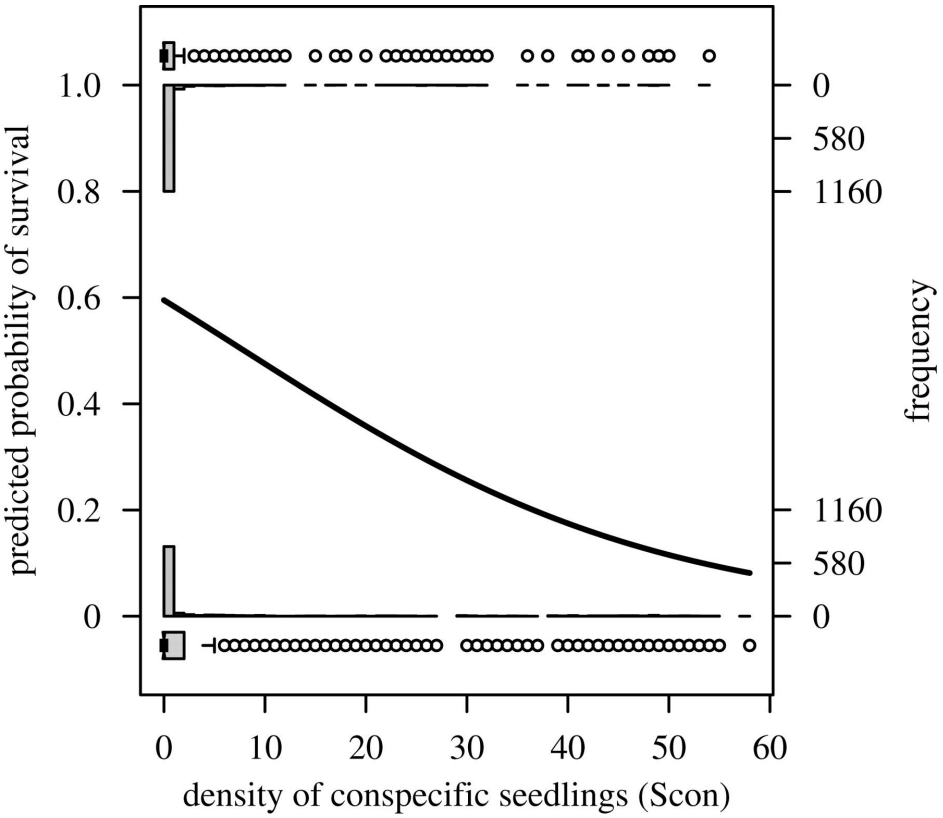
Binomial distribution

- How real data may look like:

Proportions
Successes-failures

germination success, survival

Follows a **binomial distribution**

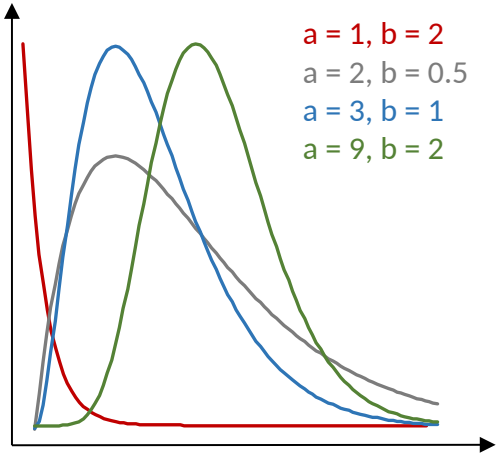


Queenborough et al. (2009) Proc. Roy. Soc. B 276:4197-4205

Gamma distribution

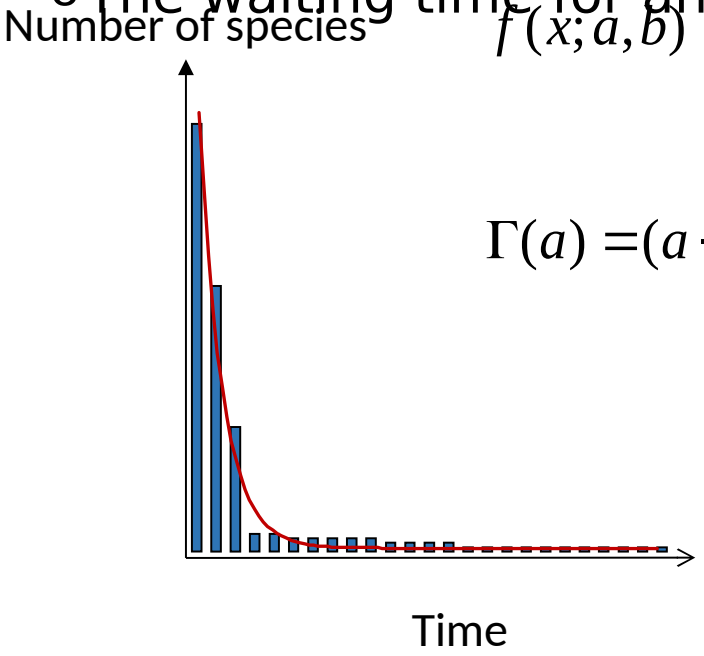
The Gamma distribution is a very general and flexible discriptions of many events, but the parameters *a* and *b* are difficult to interpret biologically

- The searching time to find species in an area
- The waiting time for an event



$$f(x;a,b) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx}$$

$$\Gamma(a) = (a-1)!$$

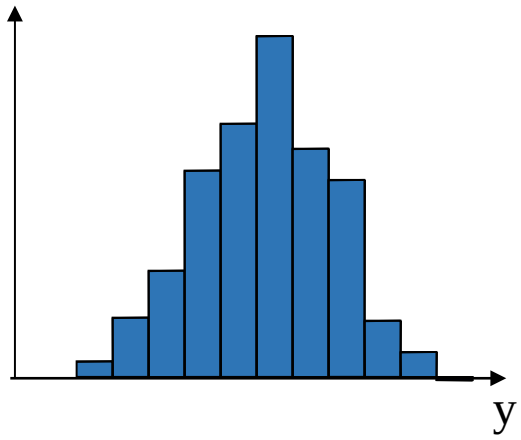


Difference with normal distribution

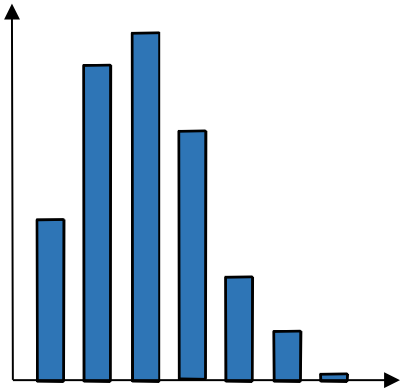
The probability distributions are often asymmetric, in contrast to a Normal distribution

The probability decreases differently from the peak (= mode)

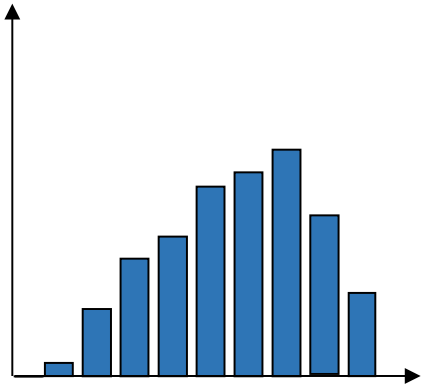
Normal



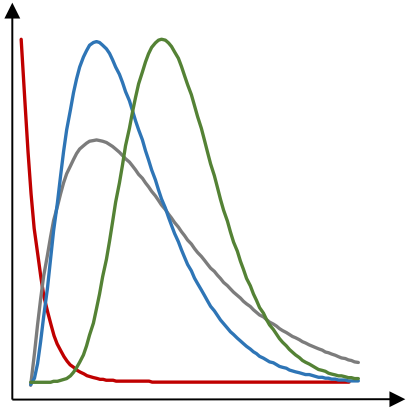
Poisson



Binomial



Gamma



The link function

The link function specifies how the parameters are related to the response variable. For most applications, use the standard link function.

The standard link functions are:

Normal data	identity	$y = a + b \cdot x$
Poisson data	log	$y = \log(a + b \cdot x)$
Binomial data	logit	$\log\left(\frac{p}{1 - p}\right) = a + b \cdot x$
Gamma	reciprocal	$y = \frac{1}{a + b \cdot x}$

Generalised linear model GLM in R

- The family argument specifies the error distribution and link function. See ?family for more information
  - binomial(link = "logit")
  - gaussian(link = "identity")
  - poisson(link = "log")

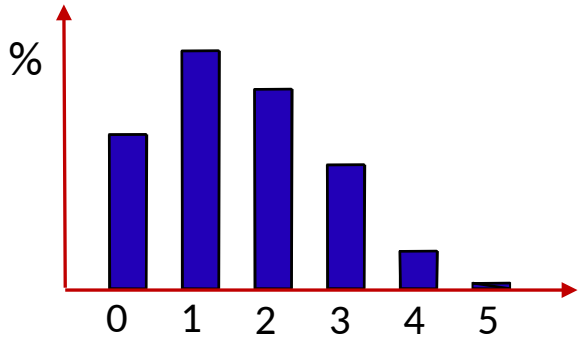
Maximum likelihood

- given the data,
- and given our choice of model,
- what values of the parameters of that model make the observed data most likely?

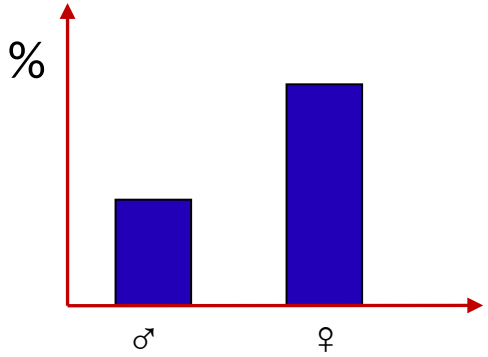
▭ probability

- Relaxing the assumption of normally distributed residuals

- Allowing other frequency distributions (Poisson, binomial, etc.)



Number of leaves per plant



Sex of captured insects

Model assumptions

- Independence of observations
- No overdispersion (Poisson, binomial): happens when there is more variation in the data than expected based on the given distribution
- (No pattern when plotting studentized residuals)

Differences in the model summary

```
> summary(glm(Plant_biomass~LUI,data=dat))

Call:
glm(formula = Plant_biomass ~ LUI, data = dat)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-487.38  -173.59   -30.62   158.95   816.71

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   415.62      59.28   7.011 1.46e-10 ***
LUI           147.64      34.23   4.314 3.30e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 70282.01)

    Null deviance: 9812009  on 122  degrees of freedom
Residual deviance: 8504123  on 121  degrees of freedom
AIC: 1725.8

Number of Fisher Scoring iterations: 2
```

```
> summary(lm(Plant_biomass~LUI,data=dat))

Call:
lm(formula = Plant_biomass ~ LUI, data = dat)

Residuals:
    Min      1Q  Median      3Q     Max
-487.38 -173.59  -30.62  158.95  816.71

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   415.62      59.28   7.011 1.46e-10 ***
LUI           147.64      34.23   4.314 3.30e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 265.1 on 121 degrees of freedom
Multiple R-squared:  0.1333,    Adjusted R-squared:  0.1261
F-statistic: 18.61 on 1 and 121 DF,  p-value: 3.296e-05
```

Contribution of each point to likelihood

Differences in the model summary

```
> summary(glm(Plant_biomass~LUI,data=dat))

Call:
glm(formula = Plant_biomass ~ LUI, data = dat)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-487.38  -173.59   -30.62   158.95   816.71

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   415.62      59.28   7.011 1.46e-10 ***
LUI           147.64      34.23   4.314 3.30e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 70282.01)

    Null deviance: 9812009  on 122  degrees of freedom
Residual deviance: 8504123  on 121  degrees of freedom
AIC: 1725.8

Number of Fisher Scoring iterations: 2
```

```
> summary(lm(Plant_biomass~LUI,data=dat))

Call:
lm(formula = Plant_biomass ~ LUI, data = dat)

Residuals:
    Min      1Q  Median      3Q     Max
-487.38 -173.59  -30.62  158.95  816.71

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   415.62      59.28   7.011 1.46e-10 ***
LUI           147.64      34.23   4.314 3.30e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 265.1 on 121 degrees of freedom
Multiple R-squared:  0.1333,    Adjusted R-squared:  0.1261
F-statistic: 18.61 on 1 and 121 DF,  p-value: 3.296e-05
```

The dispersion parameter for the gaussian family corresponds to the residual variance.

Differences in the model summary

No R-squared but Null and Residual deviance.

```
> summary(glm(Plant_biomass~LUI,data=dat))

Call:
glm(formula = Plant_biomass ~ LUI, data = dat)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-487.38  -173.59   -30.62   158.95   816.71

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   415.62      59.28   7.011 1.46e-10 ***
LUI           147.64      34.23   4.314 3.30e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 70282.01)

    Null deviance: 9812009  on 122  degrees of freedom
Residual deviance: 8504123  on 121  degrees of freedom
AIC: 1725.8

Number of Fisher Scoring iterations: 2
```

```
> summary(lm(Plant_biomass~LUI,data=dat))

Call:
lm(formula = Plant_biomass ~ LUI, data = dat)

Residuals:
    Min      1Q  Median      3Q     Max
-487.38 -173.59  -30.62  158.95  816.71

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   415.62      59.28   7.011 1.46e-10 ***
LUI           147.64      34.23   4.314 3.30e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 265.1 on 121 degrees of freedom
Multiple R-squared:  0.1333,    Adjusted R-squared:  0.1261
F-statistic: 18.61 on 1 and 121 DF,  p-value: 3.296e-05
```

Differences in the model summary

No R-squared but Null and Residual deviance.

```
> summary(glm(Plant_biomass~LUI,data=dat))

Call:
glm(formula = Plant_biomass ~ LUI, data = dat)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-487.38  -173.59   -30.62   158.95   816.71

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   415.62      59.28   7.011 1.46e-10 ***
LUI           147.64      34.23   4.314 3.30e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 70282.01)

    Null deviance: 9812009  on 122  degrees of freedom
Residual deviance: 8504123  on 121  degrees of freedom
AIC: 1725.8

Number of Fisher Scoring iterations: 2
```

```
> summary(lm(Plant_biomass~LUI,data=dat))

Call:
lm(formula = Plant_biomass ~ LUI, data = dat)

Residuals:
    Min      1Q  Median      3Q     Max
-487.38 -173.59  -30.62  158.95  816.71

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   415.62      59.28   7.011 1.46e-10 ***
LUI           147.64      34.23   4.314 3.30e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 265.1 on 121 degrees of freedom
Multiple R-squared:  0.1333,    Adjusted R-squared:  0.1261
F-statistic: 18.61 on 1 and 121 DF,  p-value: 3.296e-05
```

The residual deviance is like the residual sum of squares in a linear regression.

The smaller the better.

Differences in the model summary

```
> summary(lm(Plant_biomass~LUI,data=dat))

Call:
lm(formula = Plant_biomass ~ LUI, data = dat)

Residuals:
    Min      1Q  Median      3Q     Max
-487.38 -173.59  -30.62  158.95  816.71

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   415.62      59.28   7.011 1.46e-10 ***
LUI           147.64      34.23   4.314 3.30e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 265.1 on 121 degrees of freedom
Multiple R-squared:  0.1333,    Adjusted R-squared:  0.1261
F-statistic: 18.61 on 1 and 121 DF,  p-value: 3.296e-05
```

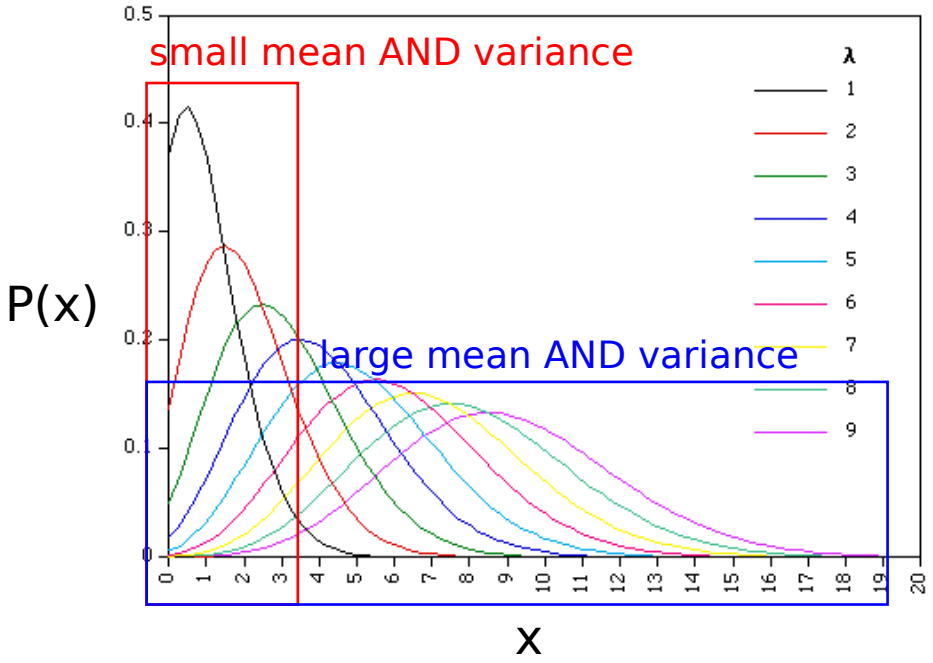
```
> summary(glm(Plant_biomass~LUI,data=dat))

Call:
glm(formula = Plant_biomass ~ LUI, data = dat)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-487.38  -173.59   -30.62   158.95   816.71

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   415.62      59.28   7.011 1.46e-10 ***
LUI           147.64      34.23   4.314 3.30e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 70282.01)

    Null deviance: 9812009  on 122  degrees of freedom
Residual deviance: 8504123  on 121  degrees of freedom
AIC: 1725.8

Number of Fisher Scoring iterations: 2
```

Akaike’s Information Criterion (AIC)
see model selection

The Poisson distribution

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

- Characteristics of a Poisson distribution:
  - **The variance equals the mean** (1 parameter=**λ**),
  - As the mean increases, the distribution gets closer to a normal distribution.



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…).

$$\underbrace{some\ function\ of\ y} = \alpha + \beta x_i$$

$$\ln(y_i) = \alpha + \beta x_i$$

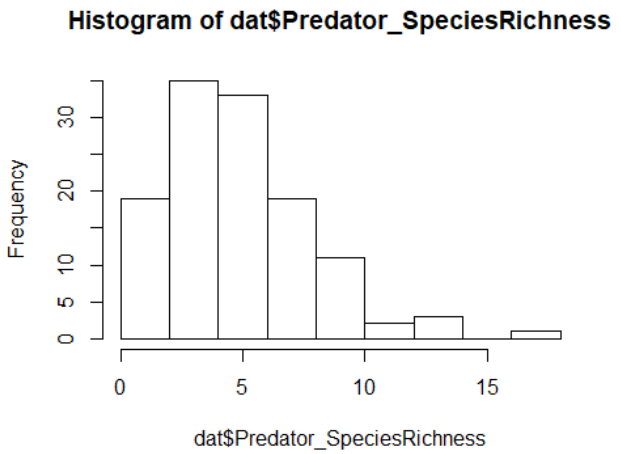
$$y_i = e^{\alpha + \beta x_i}$$

where y_i follows a Poisson distribution.

- Because y is modelled as an exponential, it is always positive (which is useful, as we can't have negative counts!).

Poisson GLM – example

```
> m10 <- glm(Predator_SpeciesRichness~Region + LUI + Herbivore_biomass,data=dat, family="poisson")
> summary(m10)

Call:
glm(formula = Predator_SpeciesRichness ~ Region + LUI + Herbivore_biomass,
    family = "poisson", data = dat)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8060  -0.8338  -0.0949   0.5261   3.3501

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)        1.595e+00  1.065e-01  14.970  < 2e-16 ***
RegionSCH         -7.161e-02  9.720e-02  -0.737   0.4613
RegionALB         -2.087e-01  1.012e-01  -2.062   0.0392 *
LUI               -6.012e-02  5.675e-02  -1.059   0.2894
Herbivore_biomass  2.235e-05  3.068e-06   7.283 3.27e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 197.37  on 122  degrees of freedom
Residual deviance: 149.32  on 118  degrees of freedom
AIC: 577.15

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**.

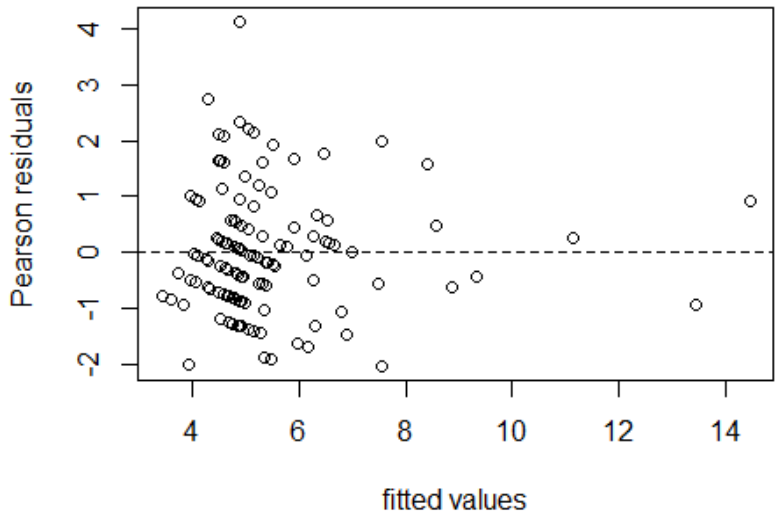
- For non-Gaussian GLMs, we use **Pearson residuals**:

$$Pearson\ residuals = \frac{Y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}}$$

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

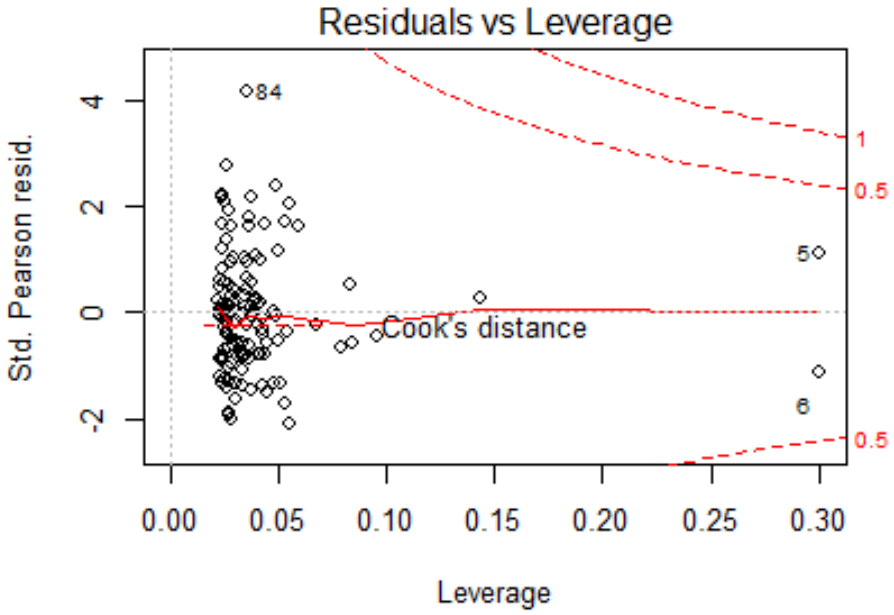
Poisson GLM model validation

```
> E2 <- resid(m10, type="pearson")
> F2 <- fitted(m10, type="response")
> plot(x=F2, y=E2, xlab="fitted values", ylab="Pearson residuals")
> abline(h=0, lty=2)
```



No patterns should be visible when we plot the Pearson residuals against the fitted values.

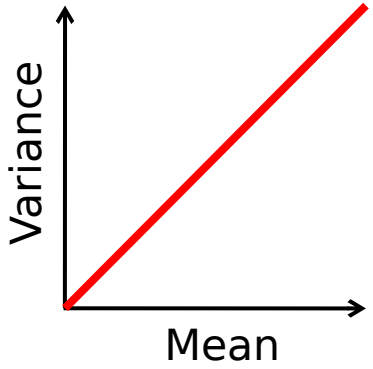
Poisson GLM model validation

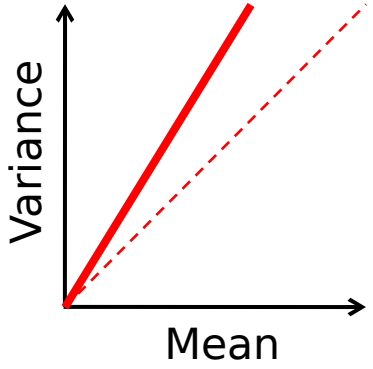


And we might still have influential points (like in a linear regression).

Poisson GLM and overdispersion

- In a Poisson GLM, remember, **the mean should equal the variance** ($\mu=\sigma^2$).
- Dispersion is characterized by the **dispersion parameter, ρ**. ρ should be 1 if $\mu=\sigma^2$.
- **Often in ecological data, ρ>1**, i.e. the variance exceeds the mean. This is called **overdispersion**.
- Overdispersion can be thought of as **extra variation in the response** that cannot be captured by a Poisson GLM.





Poisson GLM and overdispersion

- We can check to see if our response variable is overdispersed, using the GLM summary:

```
> m10 <- glm(Predator_SpeciesRichness~Region + LUI + Herbivore_biomass,data=dat, family="poisson")
> summary(m10)

Call:
glm(formula = Predator_SpeciesRichness ~ Region + LUI + Herbivore_biomass,
    family = "poisson", data = dat)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8060  -0.8338  -0.0949   0.5261   3.3501

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)        1.595e+00  1.065e-01  14.970  < 2e-16 ***
RegionSCH         -7.161e-02  9.720e-02  -0.737   0.4613
RegionALB         -2.087e-01  1.012e-01  -2.062   0.0392 *
LUI               -6.012e-02  5.675e-02  -1.059   0.2894
Herbivore_biomass  2.235e-05  3.068e-06   7.283 3.27e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 197.37  on 122  degrees of freedom
Residual deviance: 149.32  on 118  degrees of freedom
AIC: 577.15

Number of Fisher Scoring iterations: 4
```

The residual deviance (unexplained variation in the response) should be ≈ residual degrees of freedom.

isson GLM and overdispersion

- If residual deviance > residual degrees of freedom, then ρ > 1,
  i.e. y is overdispersed. **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).

- **If the dispersion estimate is really large (>1.5)**, we can do 2 things:

  - Use a **'quasi-poisson' GLM – family="quasipoisson"**
    It calculates ρ based on our mean and variance,
    still applying the Poisson distribution

    $$variance = \rho \, . \mu$$

  - Use a **negative binomial distribution**, where we estimate the variance as:

    $$variance = \mu + \frac{\mu^2}{k}$$

    where k is an estimated dispersion parameter

    `glm.nb() in R`

An overdispersed model

```
> mod11 <- glm(Herbivore_SpeciesRichness~Plant_biomass,data=dat, family="poisson")
> summary(mod11)

Call:
glm(formula = Herbivore_SpeciesRichness ~ Plant_biomass, family = "poisson",
    data = dat)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.0706  -1.6421  -0.1246   1.3683   3.4645

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    3.640e+00  3.987e-02  91.286  < 2e-16 ***
Plant_biomass -3.471e-04  5.846e-05  -5.937 2.91e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 443.74  on 122  degrees of freedom
Residual deviance: 408.21  on 121  degrees of freedom
AIC: 1052

Number of Fisher Scoring iterations: 4
```

An overdispersed model

Use negative binomial model

```
> library(MASS)
> mod11bis <- glm.nb(Herbivore_SpeciesRichness~Plant_biomass,data=dat)
> summary(mod11bis)

Call:
glm.nb(formula = Herbivore_SpeciesRichness ~ Plant_biomass, data = dat,
    init.theta = 13.00294581, link = log)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.52676  -0.91845  -0.05423   0.72938   1.80570

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    3.6328176  0.0744284  48.810  < 2e-16 ***
Plant_biomass -0.0003364  0.0001062  -3.166  0.00155 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(13.0029) family taken to be 1)

    Null deviance: 134.53  on 122  degrees of freedom
Residual deviance: 124.20  on 121  degrees of freedom
AIC: 916.16

Number of Fisher Scoring iterations: 1
```

Binomial GLM

- **Binomial GLMs** also involve **transformation of a linear predictor**, to obtain predicted values of y. We most often the **logit link function**.

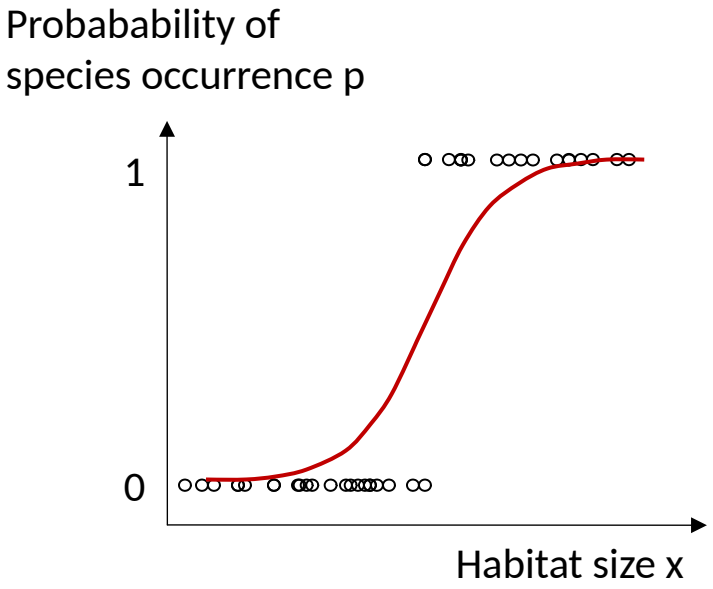
- Probability is bounded by 0 and 1.

$$\underbrace{some\ function\ of\ y} = \alpha + \beta x$$

$$\log(odds) = \alpha + \beta x$$

$$\Leftrightarrow \quad \text{logit}(y_i) = \log\left(\frac{p}{1-p}\right) = \alpha + \beta x$$

$$y_i = \frac{e^{\alpha+\beta x}}{1 + e^{\alpha+\beta x}}$$



Binomial GLM with proportional data

- With **proportional data**, we need to **represent the data as the number of 'successes' and the number of 'failures'**, using cbind().

- 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:

$$\text{Variance} = \boldsymbol{\rho} np(1-p)$$

In R write: `glm(cbind(success,failure) ~ x1 +x2 + …)`

Binomial GLM – example

```
> m12 <- glm(cbind(Herbivore_SpeciesRichness,Predator_SpeciesRichness)~ Region + Fstd + Gstd,
+            family="binomial", data=dat)
> summary(m12)

Call:
glm(formula = cbind(Herbivore_SpeciesRichness, Predator_SpeciesRichness) ~
    Region + Fstd + Gstd, family = "binomial", data = dat)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7220  -0.6335   0.2156   0.7826   2.5456

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.75783    0.08164  21.532   <2e-16 ***
RegionSCH   -0.03873    0.10559  -0.367    0.714
RegionALB    0.14280    0.10446   1.367    0.172
Fstd        -0.03079    0.02945  -1.045    0.296
Gstd        -0.03498    0.02818  -1.241    0.215
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 140.40  on 122  degrees of freedom
Residual deviance: 135.13  on 118  degrees of freedom
AIC: 542.73

Number of Fisher Scoring iterations: 4
```

Also check for overdispersion ⇨ quasibinomial

What to present to other researchers

- Clear hypotheses and questions
- Sampling design
- Any change to the initial dataset
- Statistical model
- Package used
- Results
  - Estimates + std errors
  - t and pvalues
  - R-squared
  - Degrees of freedom
  - Plots with raw / modelled data