Basic statistics and statistical analyses

Emily Haeuser

Eva Malecore

Ecology Konstanz

TReND

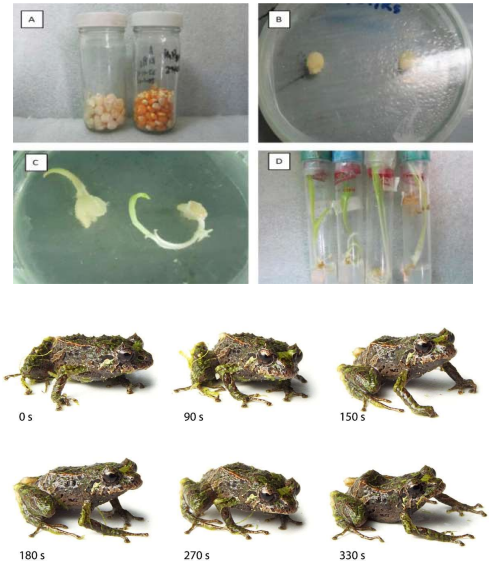
Institute for Health Metrics and Evaluation

The Company of Biologists

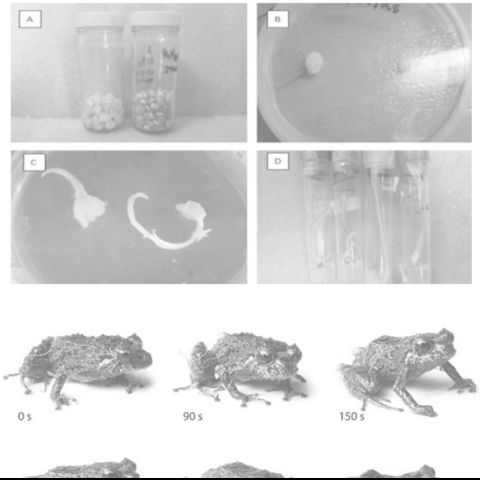
International Max Planck Research School for Organismal Biology

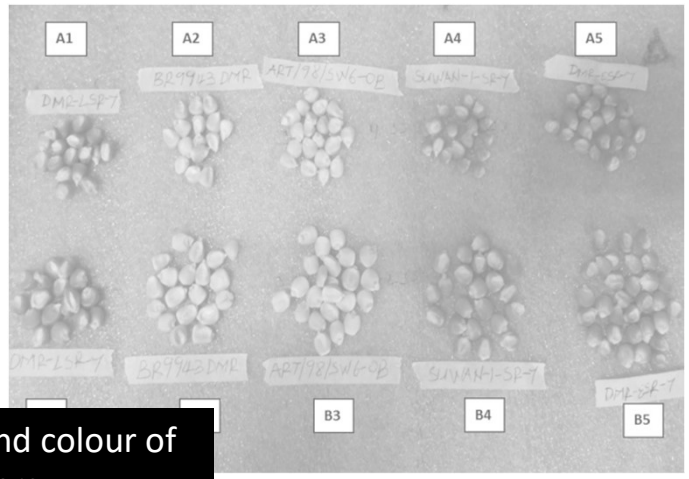
Basis for statistics

Why do we need statistics?

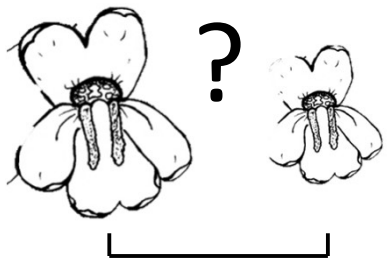


Why do we need statistics?





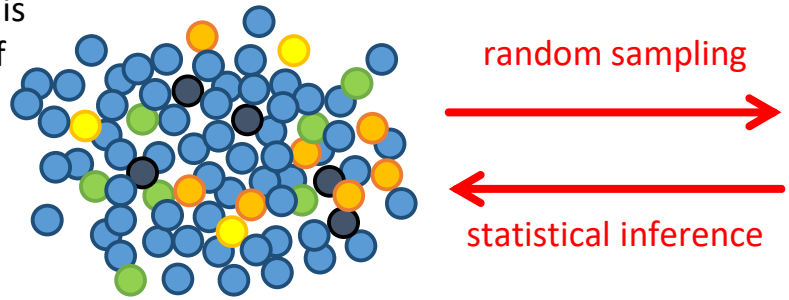
Variation in flower size, shape and colour of *Mimulus guttatus* flowers



Why do we need statistics?

- In everything we measure or observe, there is variation.
  - How to make sense of this variation?
  - Is it just random variation, or is there true underlying difference?
- We can never measure everything, so we take a sample.
  - We calculate an effect or a difference based on this sample,
  - We use statistics to make inferences from this sample.

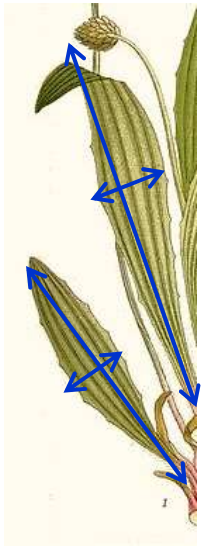
The **population** is the collection of all individuals about which we want to make a scientific statement.

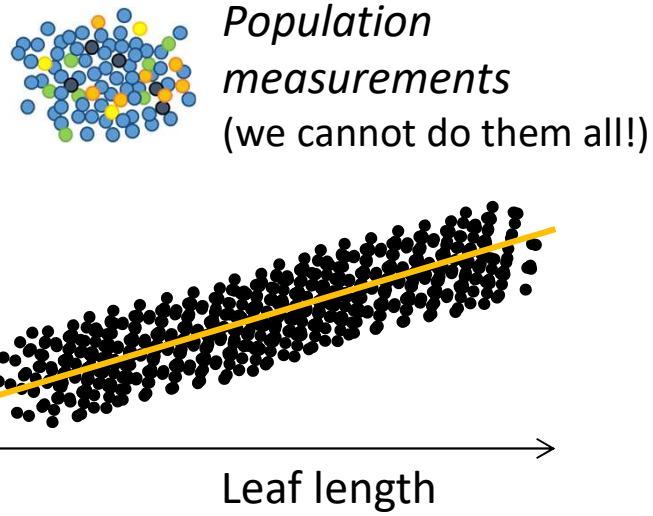


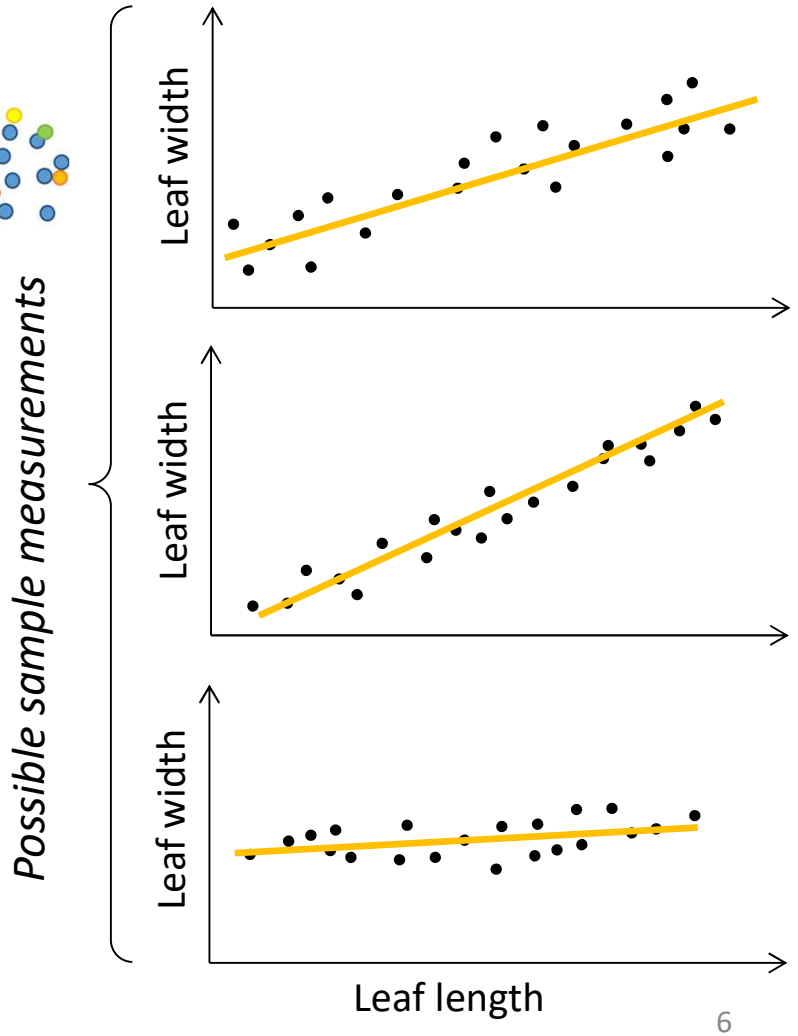
The **sample** is the subset of individuals that we actually measure. It should be representative of the population.

Sampling error

- It refers to the fact that the sample may (by chance) not be representative of the population.

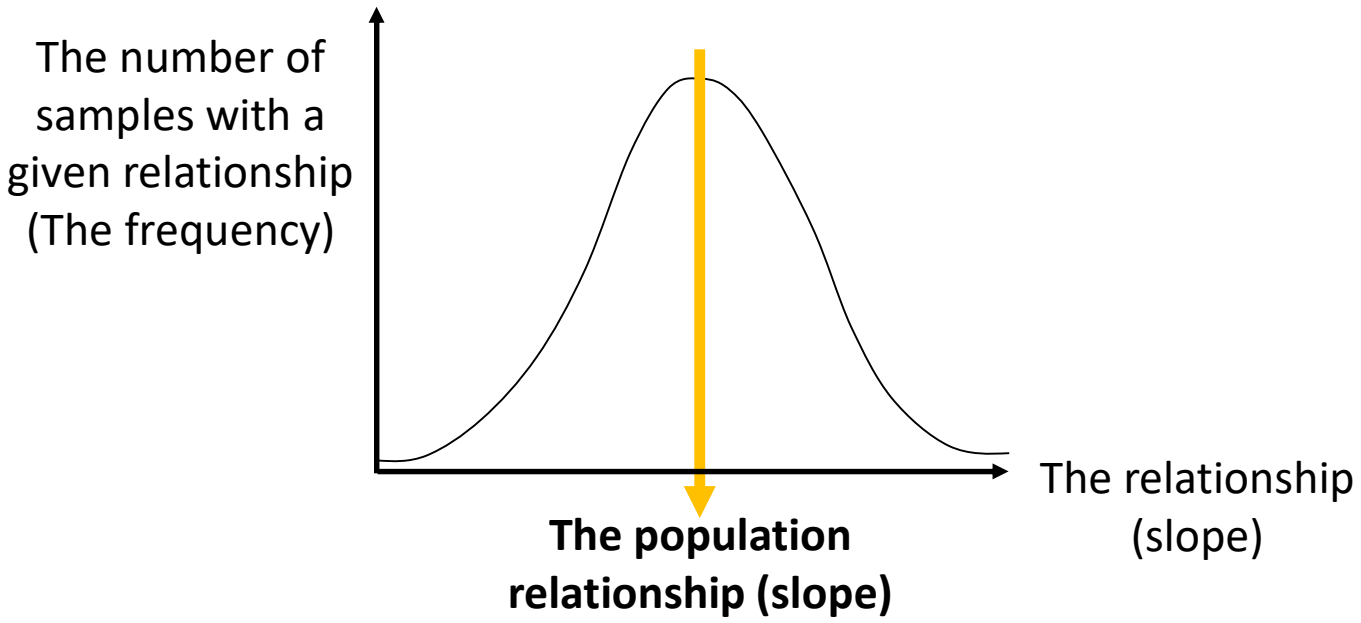






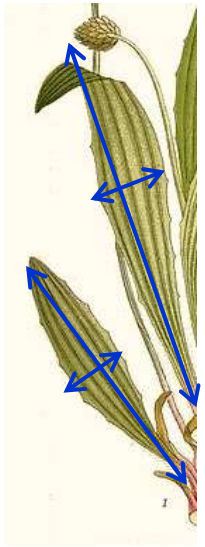
Sampling error

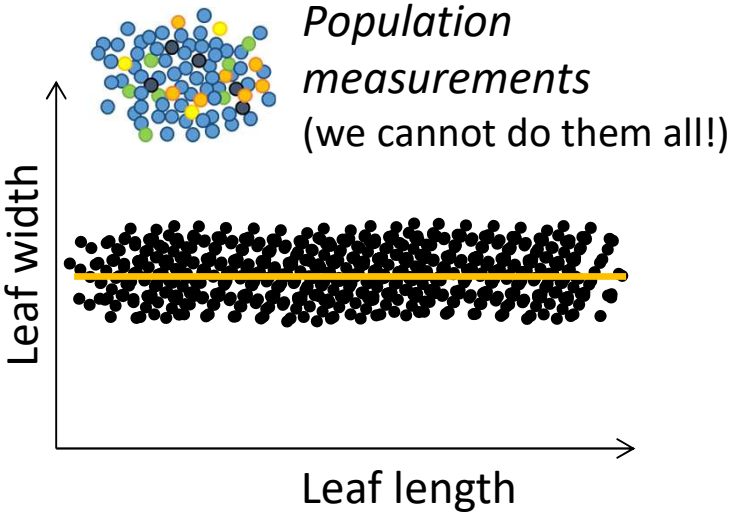
- But if we repeated the experiment many times (i.e. if we had many samples), most of the time we would expect to get a positive relationship close to the population relationship.

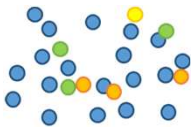


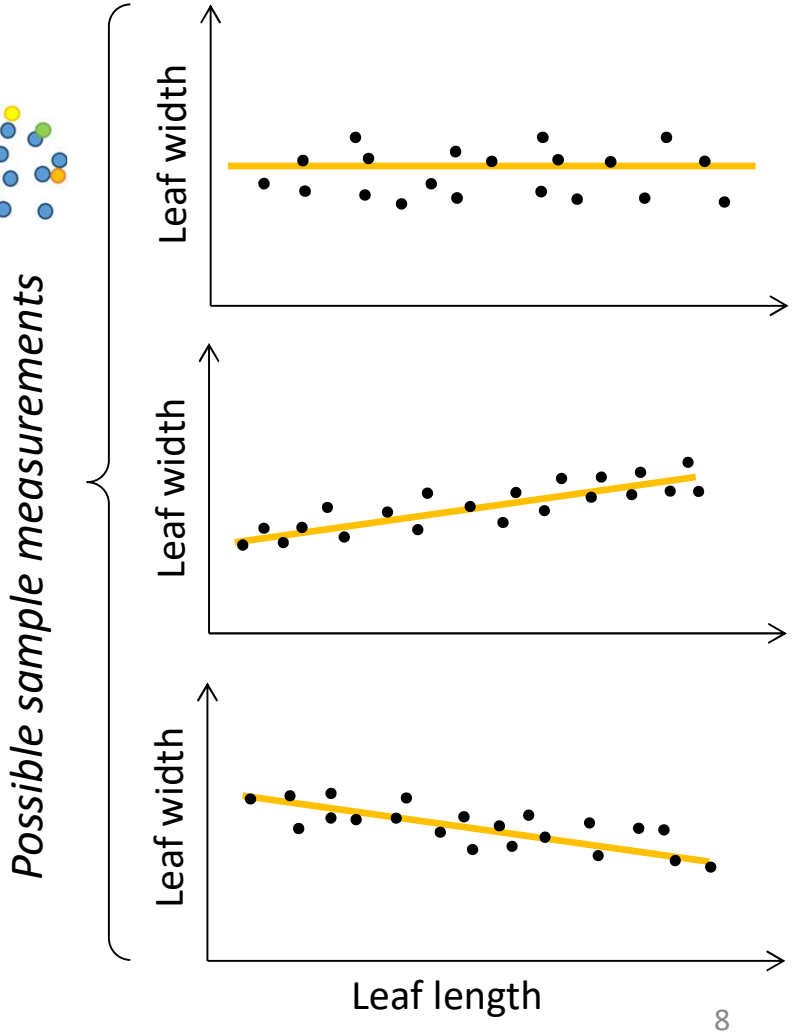
Sampling error

- Similarly…



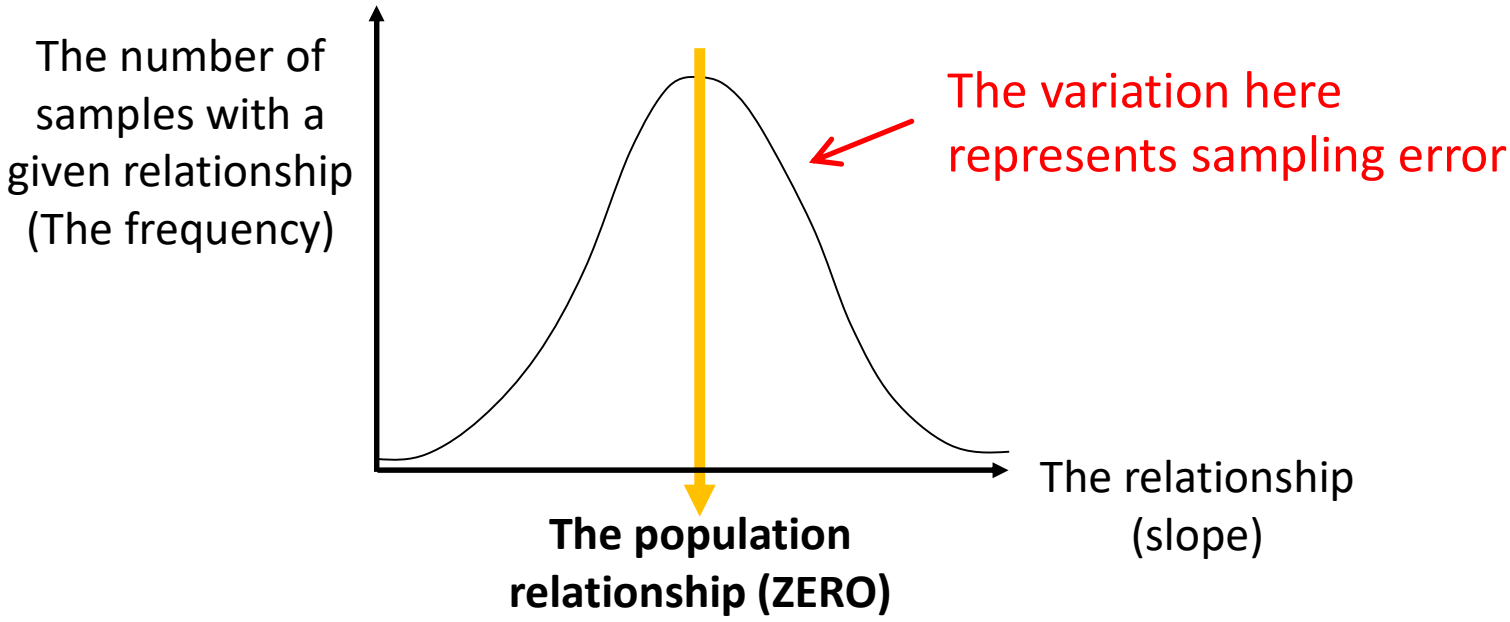






Sampling error

- But again, if we repeated the experiment many times, most of the time we would expect to get no relationship.



- This would be the **distribution** of possible relationships which we would get from our samples **when there is no real relationship**.

Hypothesis testing

- In statistics, we say we do 'statistical tests' because we ***test* hypotheses**.

- We examine two opposing hypotheses:
  - The **null hypothesis** (called 'H_0') describes what we expect to find if the underlying mechanism we are interested in does NOT operate
  - The **alternative hypothesis** (called 'H_1') describes what we expect to find if the underlying mechanism we are interested in indeed operates

Example:

Null hypothesis H_0	Alternative hypothesis H_1
There is no difference in biomass between treatment and control	There is a difference in biomass between treatment and control (nondirectional hypothesis)
	Biomass is higher in treatment than in control (directional hypothesis)

Statistical test

- We collect data and we use statistics to **test whether we can reject H_0** in favour of H_1 (i.e. the data support H_1).

- Specifically, we calculate the probability of finding the pattern we have in our data if H_0 was true (called '**p-value**'):
  - When the pattern in our data is so improbable given H_0 that it cannot have arisen through mere chance (i.e. **p-value very low**), we **reject H_0** and we say there is a **significant difference**
  - Alternatively, when it is not too improbable that the pattern in our data has arisen through mere chance (i.e. **p-value not so low**), we **cannot reject H_0** and we say there is **no significant difference**.

Statistical test

- As ‘very low’ and ‘not so low’ are very subjective, we need to agree on a **criterion** to decide how low the p-value must be to reject H_0.

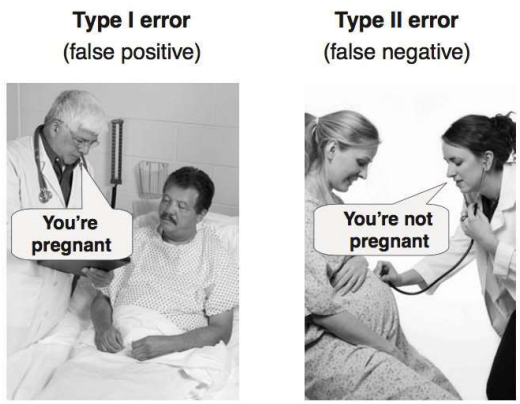
- This critical probability is called **significance level**, denoted **α**. It represents the risk of wrongly rejecting H_0 while it is true.
  - Most **common** is to use **α=0.05** (5%).
  - Sometimes, people use α=0.01 (more stringent) or α=0.1 (less stringent).
  - The significance level must be clearly decided before doing the analysis!

Type I and Type II errors

- We never know 'the truth'! Statistics only allow to make inferences based on the data sampled. There is always a risk you draw the wrong conclusion:

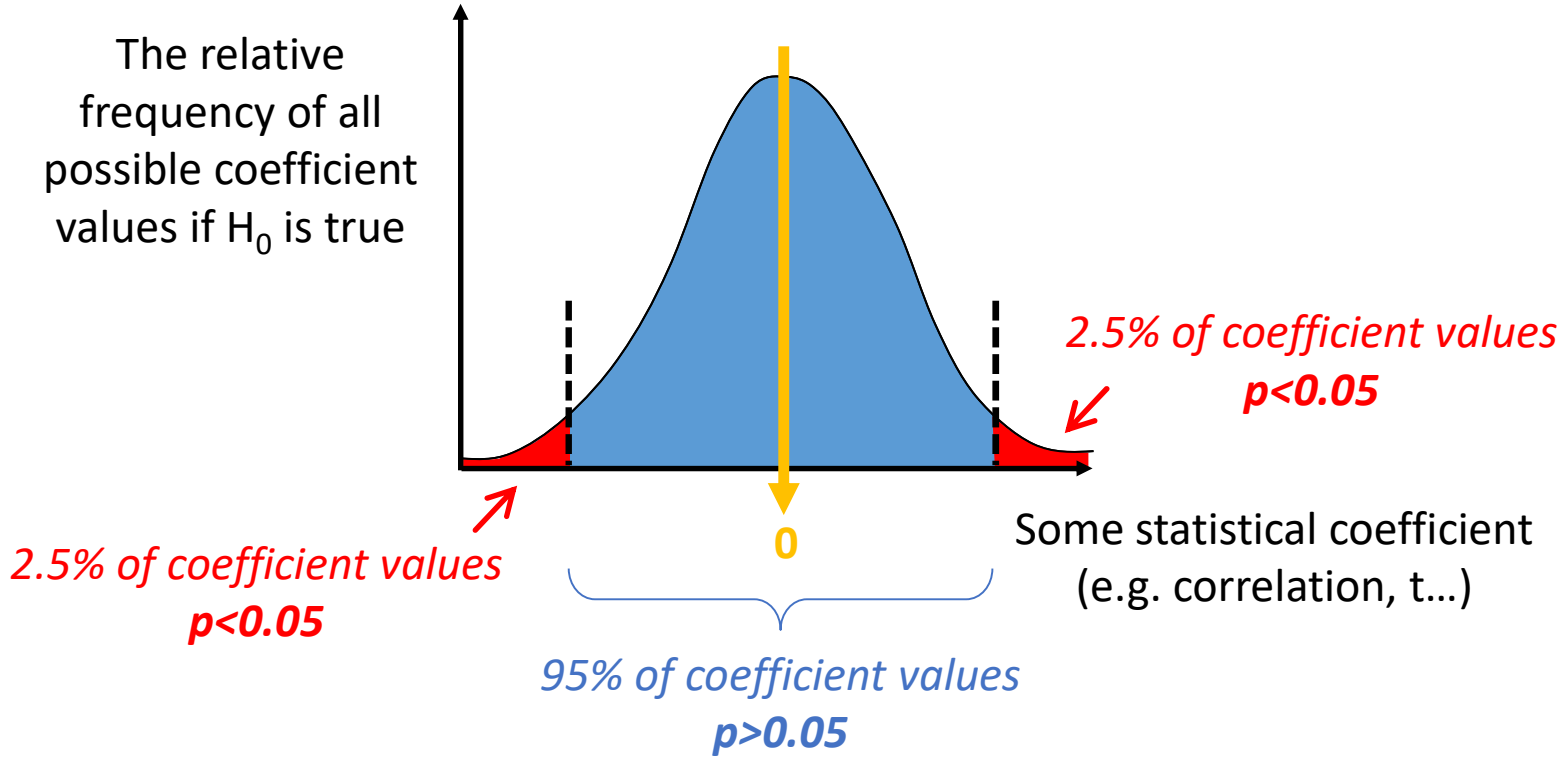
		In reality:	
		H_0 is true	**H_0 is false**
Outcome of statistical test is to:	**Reject H_0** ($p\text{-value}<\alpha$)	Type I error (false positive) Probability=α	Correct conclusion
	Not reject H_0 ($p\text{-value}>\alpha$)	Correct conclusion	Type II error (false negative)

Example:
H0: "You're not pregnant."



The famous p-value

- It tells us the probability of obtaining a given statistical coefficient just by chance based on our sample, when there is in reality no relationship in the true population (i.e. H_0 is true).



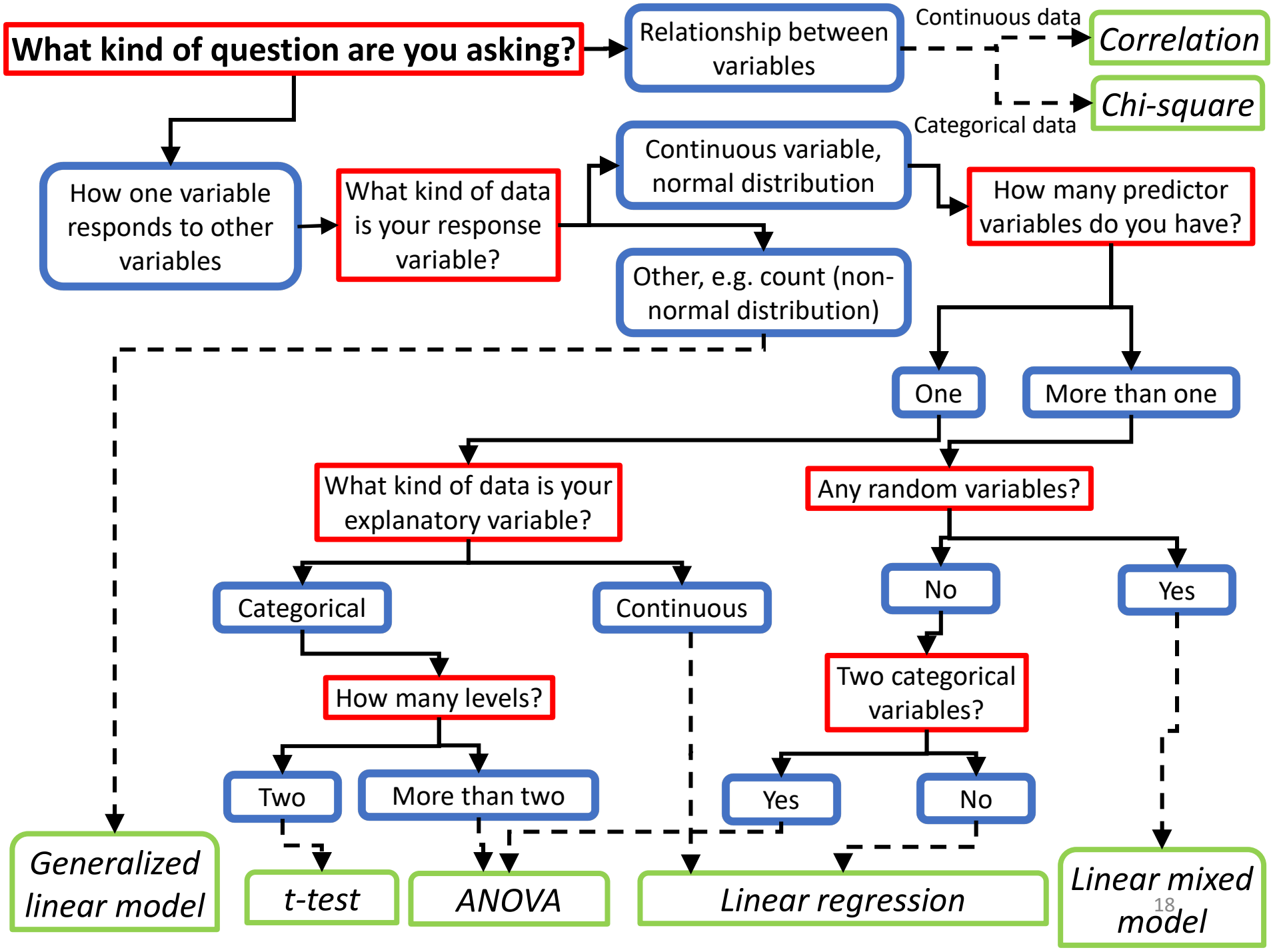
Performing statistical analyses

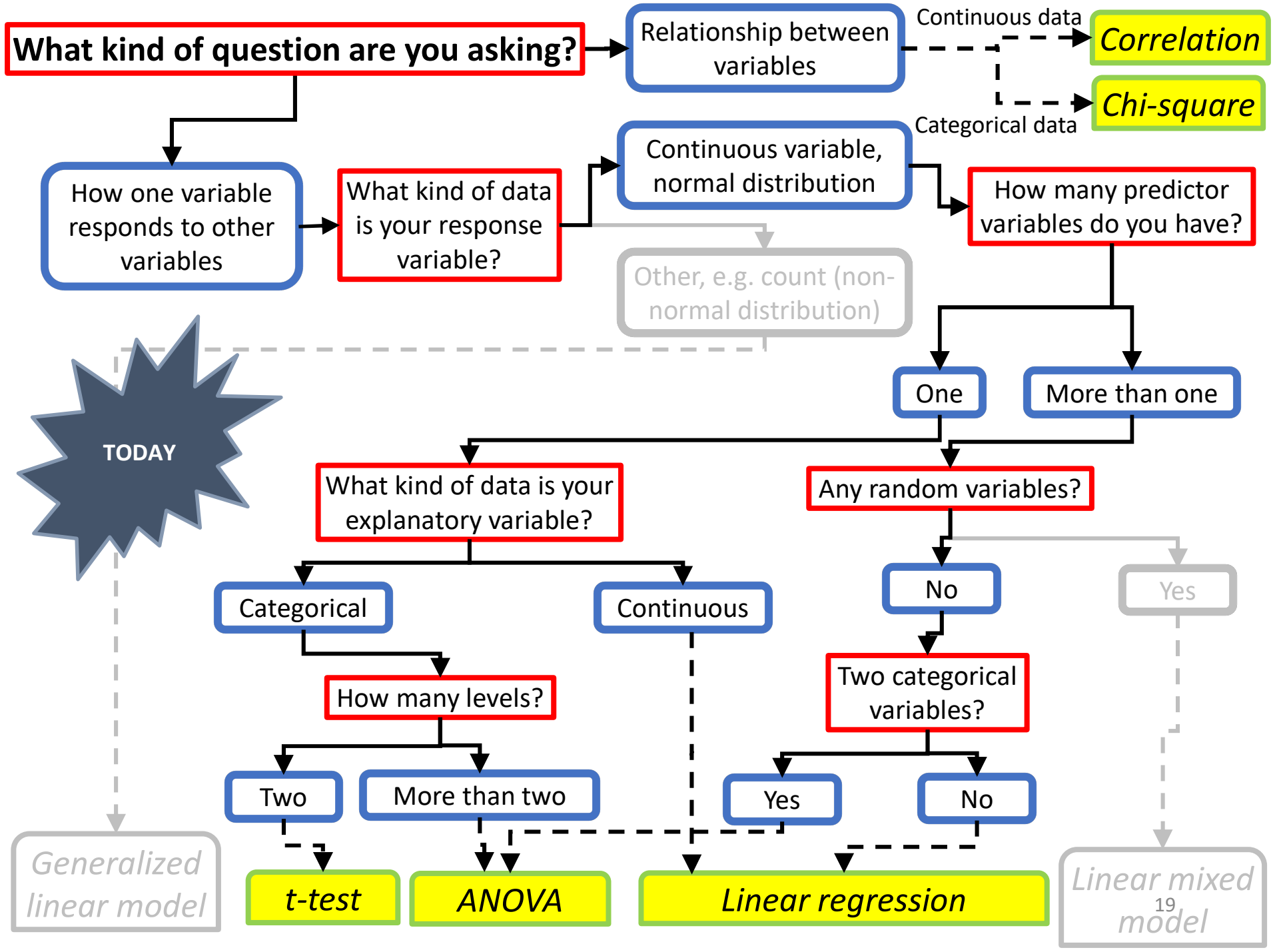
How to select the right analysis?

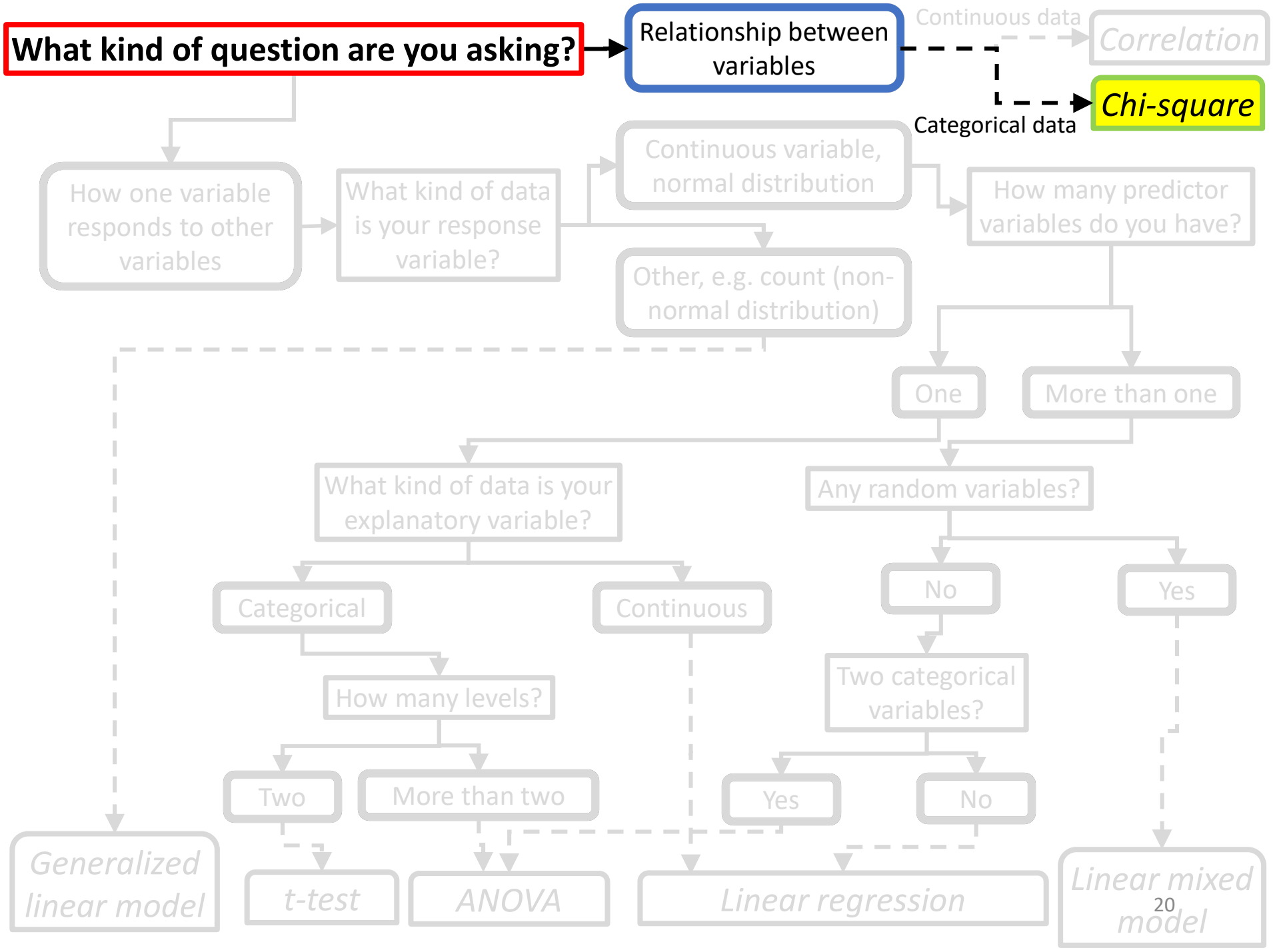
1. What is your **hypothesis**?
2. What is your **response** (a.k.a. **dependent**) variable?
   Do you even have one?
3. What kind of data is it, how is it **distributed**?
   Use graphics to explore distribution shape and data outliers
4. What are your **explanatory** (a.k.a. **independent**) variables?
   How many do you have? Are they continuous, categorical, etc.?
5. Are your **data points grouped** in some way?
   Do you need to include any random variables?

Warning!

As you will see shortly, **implementation** of analysis in R can be relatively simple. This makes R very powerful but also dangerous. Understanding what these analyses are doing is important for selecting the right test, and correctly interpreting and explaining your results.


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

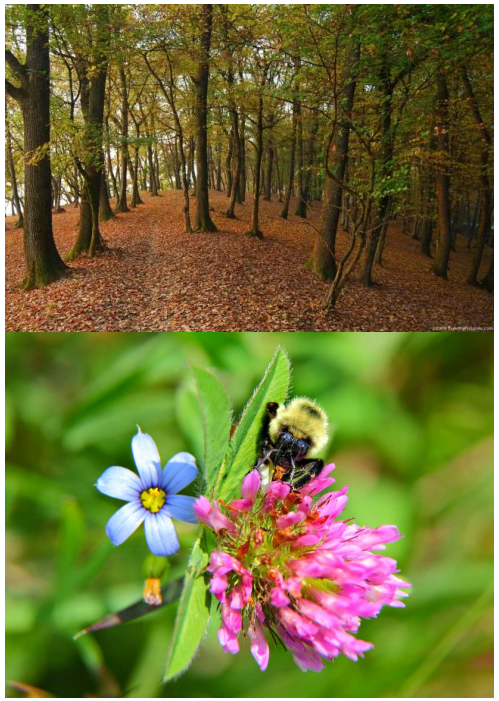

What kind of question are you asking?
Relationship between variables
Continuous data
Correlation
Categorical data
Chi-square
How one variable responds to other variables
What kind of data is your response variable?
Continuous variable, normal distribution
Other, e.g. count (non-normal distribution)
How many predictor variables do you have?
TODAY
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?
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

Chi-square test

- The **Chi-square test of independence** evaluates whether there is a significant association between the categories of **two categorical variables**.

- For example:
  - Do two tree species tend to grow in similar numbers in different soil types?
  - Do flowers of different colors attract similar types of pollinators?

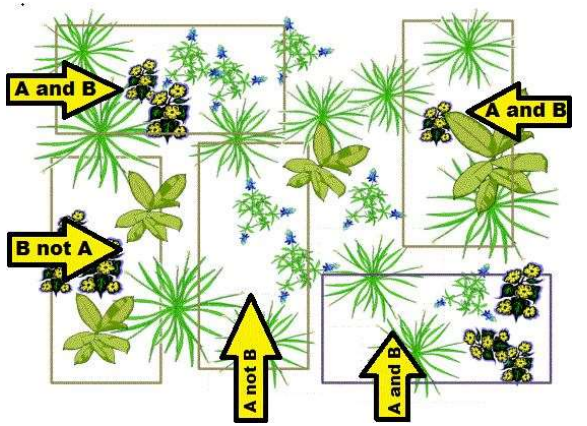


Chi-square test

- Chi-square tests are based on comparing **observed** and **expected** values: does the observed data differ from what we would expect if there were no association between groups.

- For observed values, we construct **contingency tables**.
  - Example: is there an association between the presence and absence of species A and B in 100 vegetation plots?

Observed values	Species A	Species B
Present	69	55
Absent	31	45



Chi-square test

- From these tables we can **derive** the expected values.
  - For our example:

Observed values	Species A	Species B
Present	69	55
Absent	31	45

124 *All presences*

76 *All absences*

200 *Total observations*

All species_A obs (69+31)	?
All obs (69+31+55+45)	All presences (69+55)

(69+31)*(69+55)/(69+55+31+45) = 62

All species_A obs (69+31)	?
All obs (69+31+55+45)	All absences (31+45)

(69+31)*(31+45)/(69+55+31+45) = 38

Expected values	Species A	Species B
Present	62	62
Absent	38	38

(Same for species B)

Chi-square test

- The formula for Chi-square is:

$$X^2 = \sum \frac{(obs - exp)^2}{exp}$$

  - For our example:

$$X^2 = \frac{(69-62)^2}{62} + \frac{(55-62)^2}{62} + \frac{(31-38)^2}{38} + \frac{(45-38)^2}{38} = 3.16$$

Observed values	Species A	Species B
Present	69	55
Absent	31	45

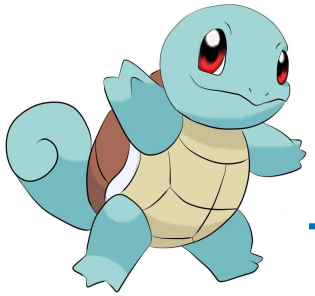
Expected values	Species A	Species B
Present	62	62
Absent	38	38

Chi-square test

- But what does that value of 3.16 mean? Is there a significant association?

- The calculated X^2 value is then compared to the "critical X^2 value" in a chi-square distribution table, based on **degrees of freedom**, to give us a p-value.
  - Degrees of freedom tell us about our **power** to detect significant relationships in our data, which depends on our sample size and analytical complexity
  - Our degrees of freedom are based on the number of levels within our two variables: df=(nlevels$_{Var1}$-1)*(nlevels$_{Var2}$-1)
  - In our example: (2-1)*(2-1)=1*1=1 degree of freedom

- Alternatively, we can do it in R…



Chi-square test in R

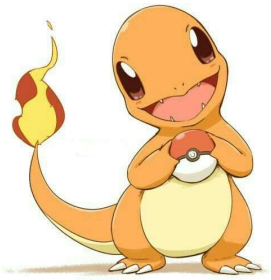
Chi-square test

You can let R do the whole process for you!
chisq.test(Data$Variable1, Data$Variable2)

Expected output:

```
Pearson's Chi-squared test

data:  Data
X-squared =XXX, df = XXX, p-value = XXX
```



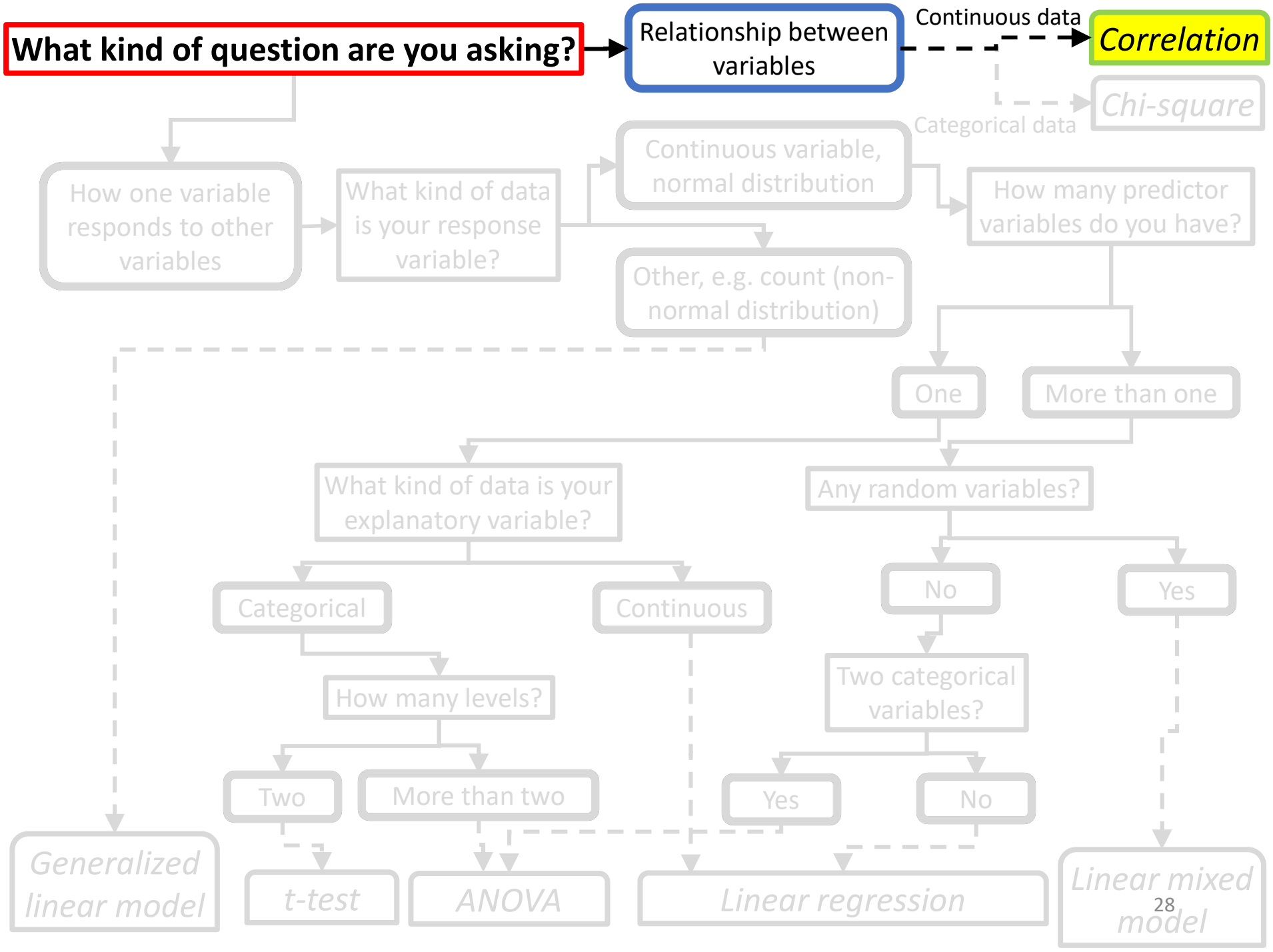
Time for an exercise

With our leaf traits dataset, we want to know the relationship between:

- Deciduousness (deciduous/evergreen) and leaf type (needle/broadleaf)

Steps:

1) Load ‘leaftraits.txt’ (if not already loaded)
2) Examine file structure
3) Run chi-square test for the appropriate variables


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

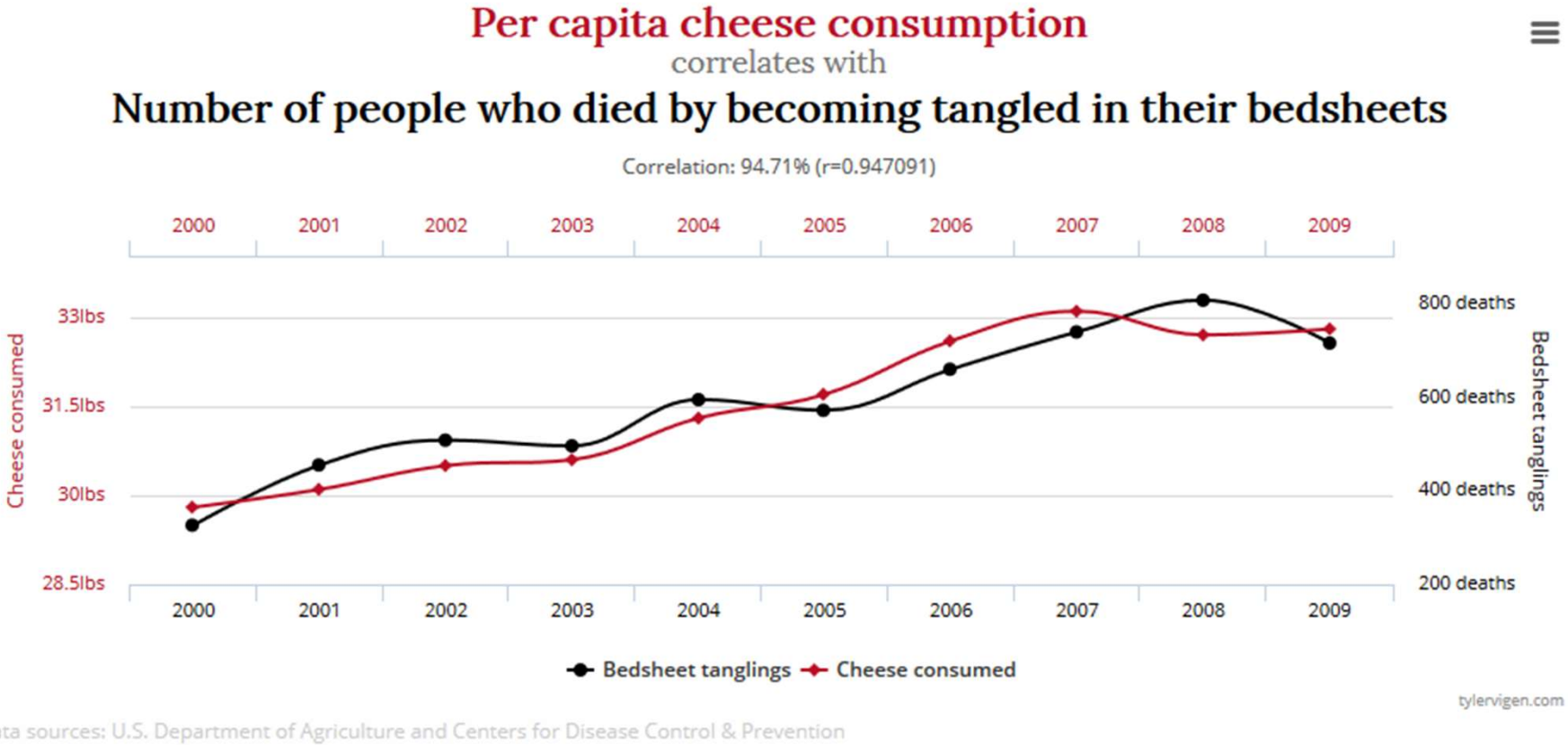
Correlation

- Correlation tells us about the presence and strength of a relationship between two continuous (numerical) variables.

- For example:
  - Is there a relationship between flower size and visitation rate by pollinators?
  - Is there a relationship between light levels and leaf area?
  - How strong is the relationship between seed weight and dispersal distance?

Correlation

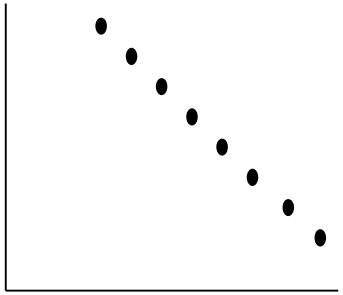
- Correlation does not measure causation or directional effects!



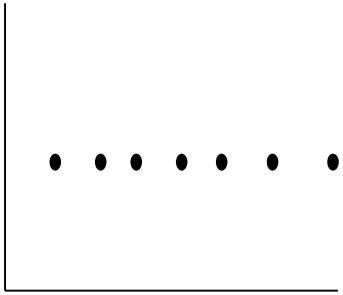
http://tylervigen.com

Correlation

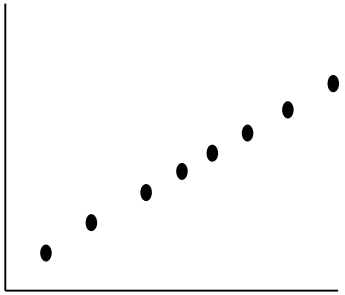
- Correlation tells us **how much y changes in relation to x**.
- It is captured by a **correlation coefficient**, which ranges between -1 (perfect negative) to +1 (perfect positive), while 0 means there is no correlation.
- Most correlations lie somewhere between extremes (there are no perfect correlations, due to residual variation).



-1 =
Perfect negative relationship between y and x



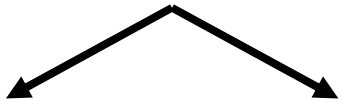
0 =
No relationship between y and x



+1 =
Perfect positive relationship between y and x

Correlation coefficient

- There are two correlation coefficients most often used:



Pearson's Product Moment Correlation (Pearson for short)

Best for normally-distributed data

Assumes a linear relationship between variables

Sensitive to outliers

Spearmans's Rank Correlation (Spearman for short)

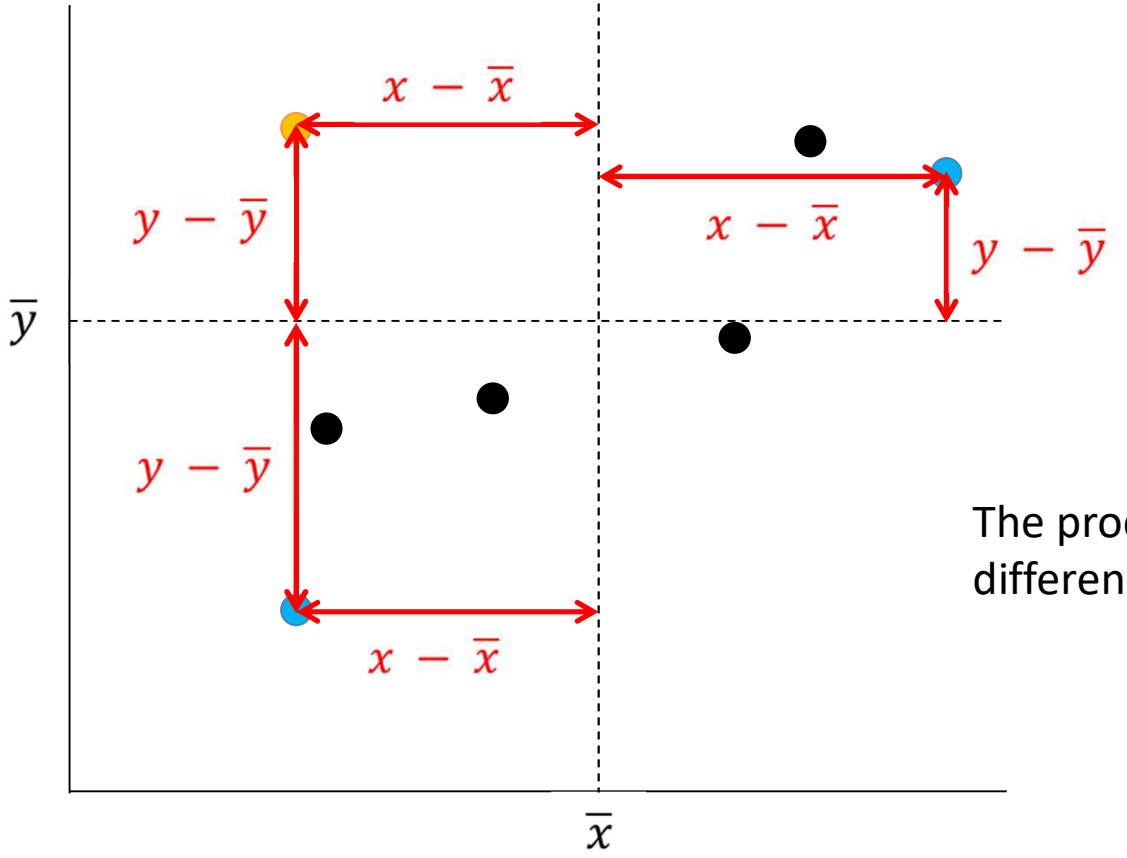
Used with non-normal data (distribution-free)

Insensitive to outliers

Pearson's correlation coefficient

The product of this x-y difference will be negative, thus reducing the sum and the coefficient strength

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$



The product of these x-y differences will be positive

Pearson's correlation test

- When we do a correlation test, our set of hypotheses is:
  - $\boldsymbol{H_0}$**: r = 0**
  - H_1: r ≠ 0

- To perform a correlation test, a first step is to **convert our correlation coefficient (r) into a t-value** in order to account for sample size:

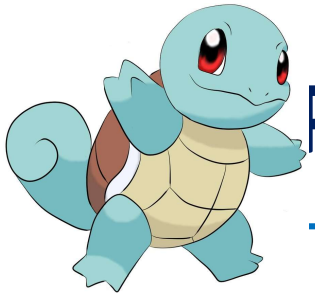
$$t = \frac{r}{\sqrt{\frac{1 - r^2}{n - 2}}}$$

Pearson's correlation test

- The t-value is what we call the '**test statistic**'.
  - In general, a test statistic summarizes the sample data in one value
  - This test statistic is then compared to a known probability distribution expected under H_0 (here, the t distribution)
  - This **allows to calculate a probability, i.e. the p-value**

- Note that the X^2 in the Chi-square test was also a test statistic…

- We will see how that works in more detail with the t-test, but for now let us focus on the interpretation of the test output.



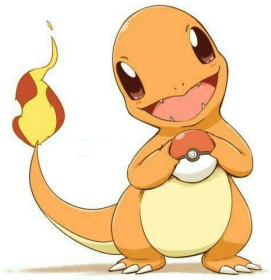
Pearson's correlation test in R

Pearson' correlation test

In R, the Pearson's correlation can be called like so:
cor.test(Data$Variable1, Data$Variable2)

Expected output:

```
Pearson's product-moment correlation
data:  Data$Variable1 and Data$Variable2
t = XXX, df = XXX, p-value = XXX
alternative hypothesis: true correlation <is not> equal to 0
95 percent confidence interval:
XXX XXX
sample estimates:
cor
XXX
```



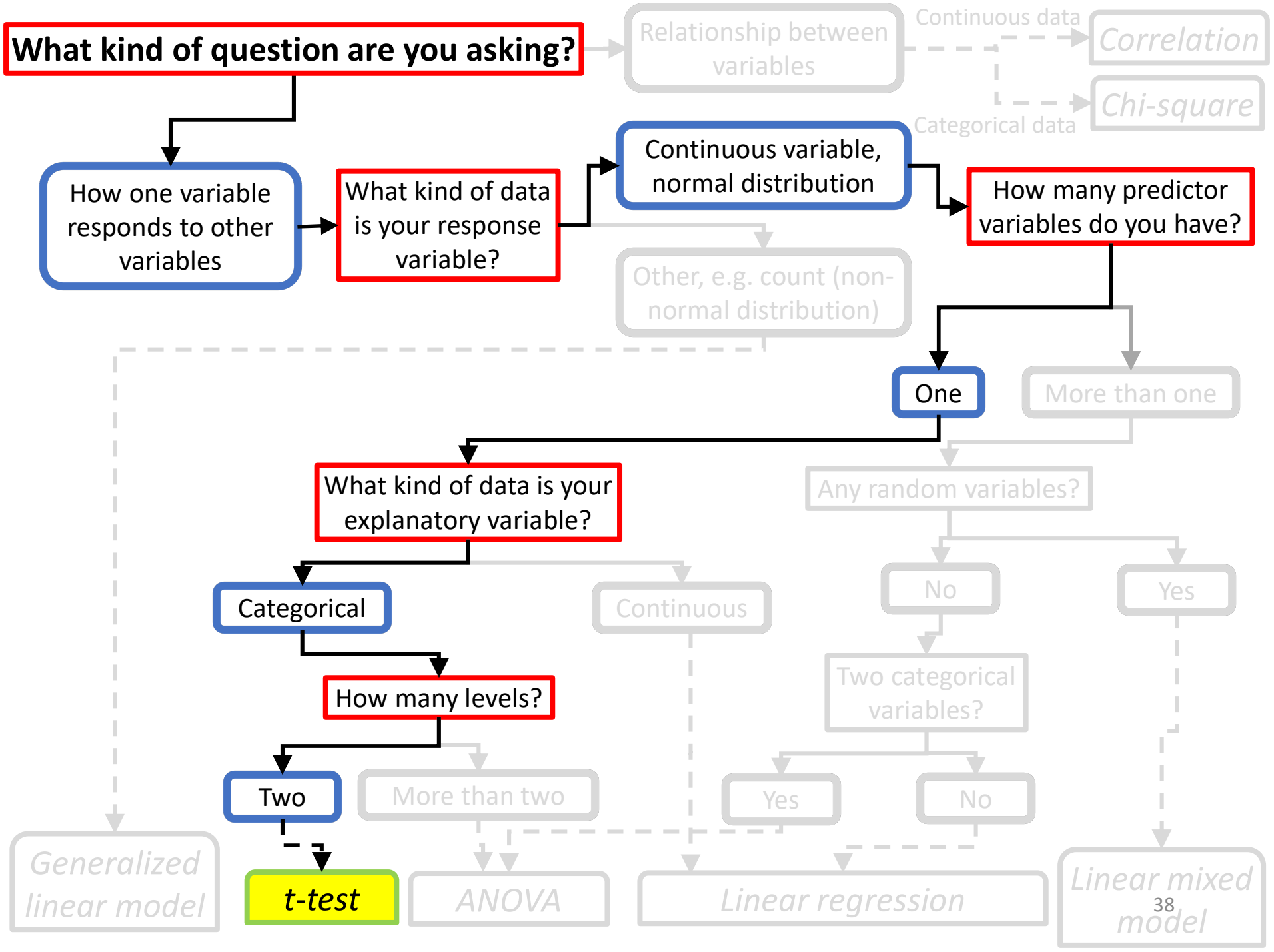
Time for an exercise

With our leaf traits dataset, we want to know what the strength of correlation is, between:

- Leaf N content and leaf P content
- Leaf N content and Leaf LMA
- Leaf P content and Leaf LMA

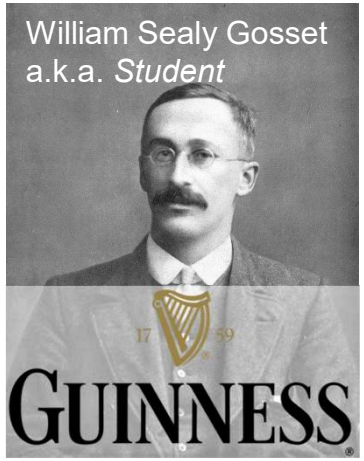
Steps:

1) Load ‘leaftraits.txt’
2) Examine file structure
3) Run correlation tests for the appropriate pairs of variables


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

Student's t-test

- With t-tests, we compare values of a response variable among two different groups, to know if they differ on average.



- There are several versions of the Student's t-test:
  - **One sample**
    to compare the mean from a random sample with a particular target value
  - **Two samples**
    to compare the means from two samples with each other
    - **Independent samples**
      the values in one sample say nothing about the values of the other sample
    - **Paired samples**
      the values in one sample affect the values in the other sample (either repeated measures or samples paired by design)

Student's two-sample t-test

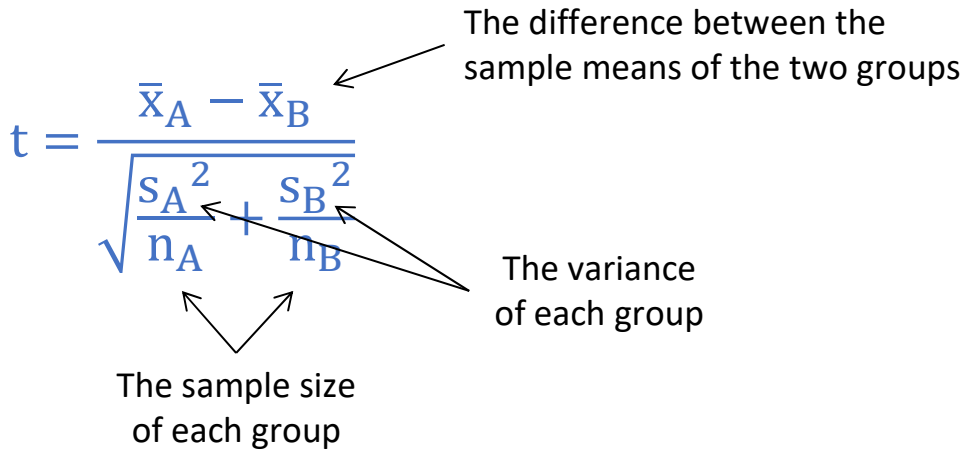
- What do we do when we compare two sample means?
  - We calculate the difference between the two sample means
  - We **test whether the mean difference is different from zero**

- Our set of hypotheses is:
  - $\boldsymbol{H_0: \mu_A - \mu_B = 0}$
  - H_1: $\mu_A - \mu_B \neq 0$ (or $\mu_A - \mu_B > 0$ or $\mu_A - \mu_B < 0$, depending on our research question)

- Because we work with samples, we need to convert the mean difference into a **standardized value** that accounts:
  - For the **size** of each sample
  - For the **variance** in each sample

Student's two-sample t-test

- The standardized value obtained is called the **t-value**:

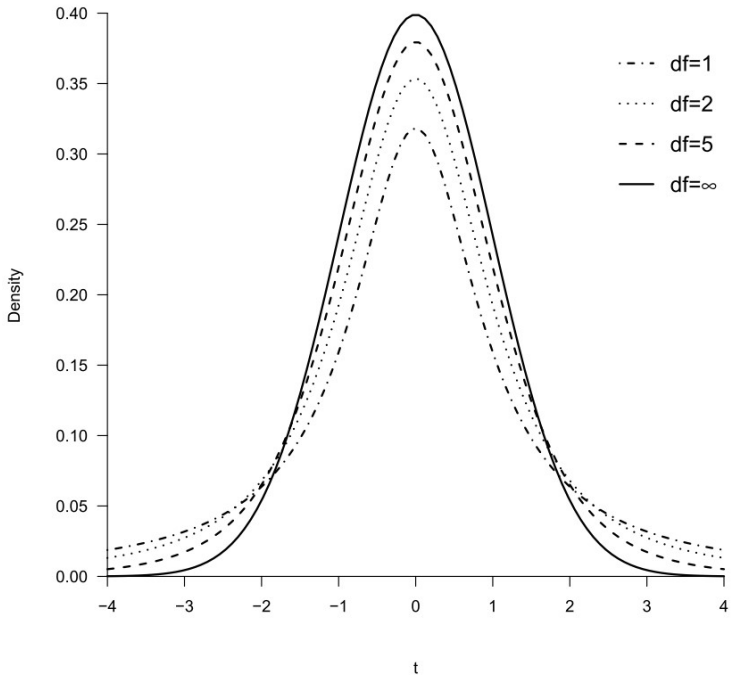


- We then **compare our sample t-value to the t distribution**.

The t distribution

- The **t distribution** represents the **range of t-values you could get under H_0** (i.e. when $\mu_A - \mu_B = 0$), if you drew multiple random samples.
- It is also a probability density function, so the area under the curve is a probability.

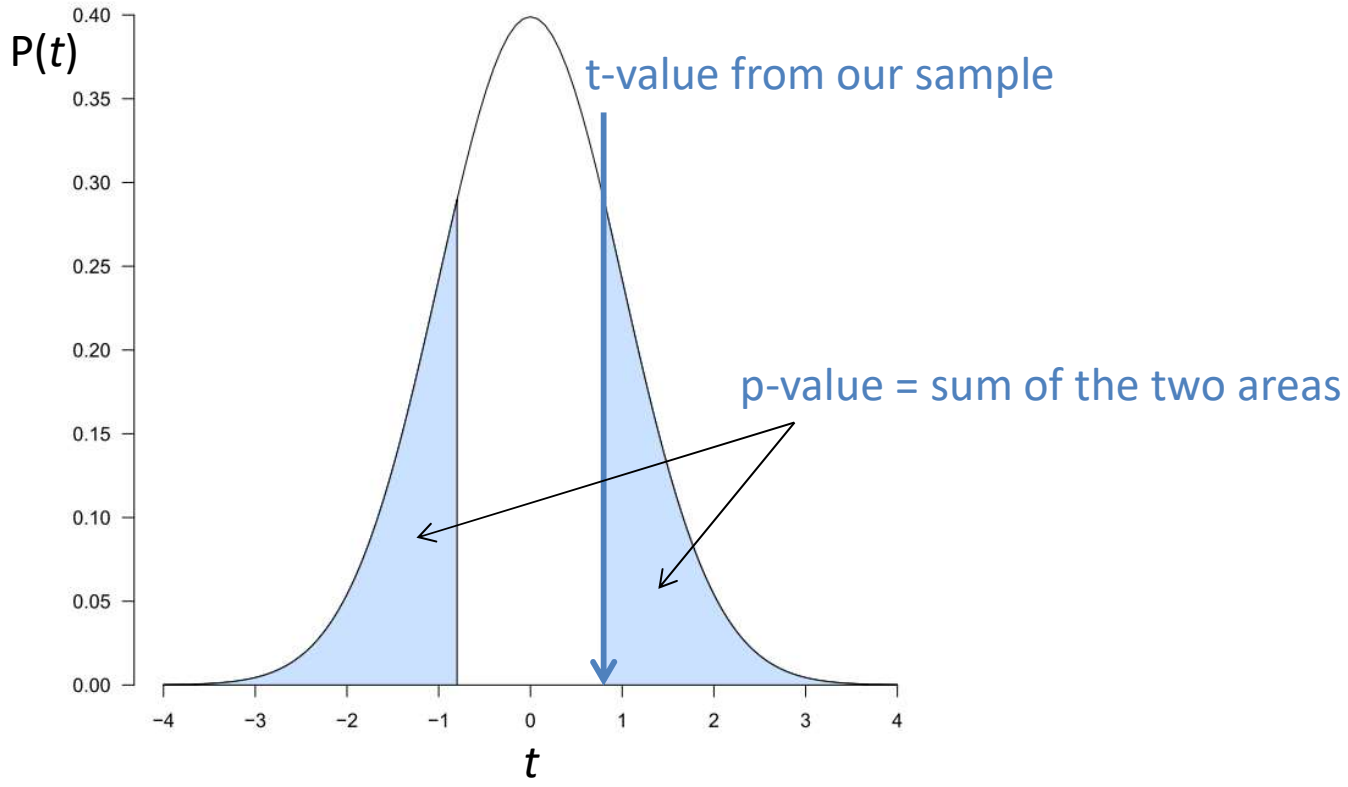
- The shape of the t distribution depends on the '**degrees of freedom**' (df)
  - With equal sample size, df=2n-2
  - With unequal sample size, df=n_A+n_B-2



Student's two-sample t-test

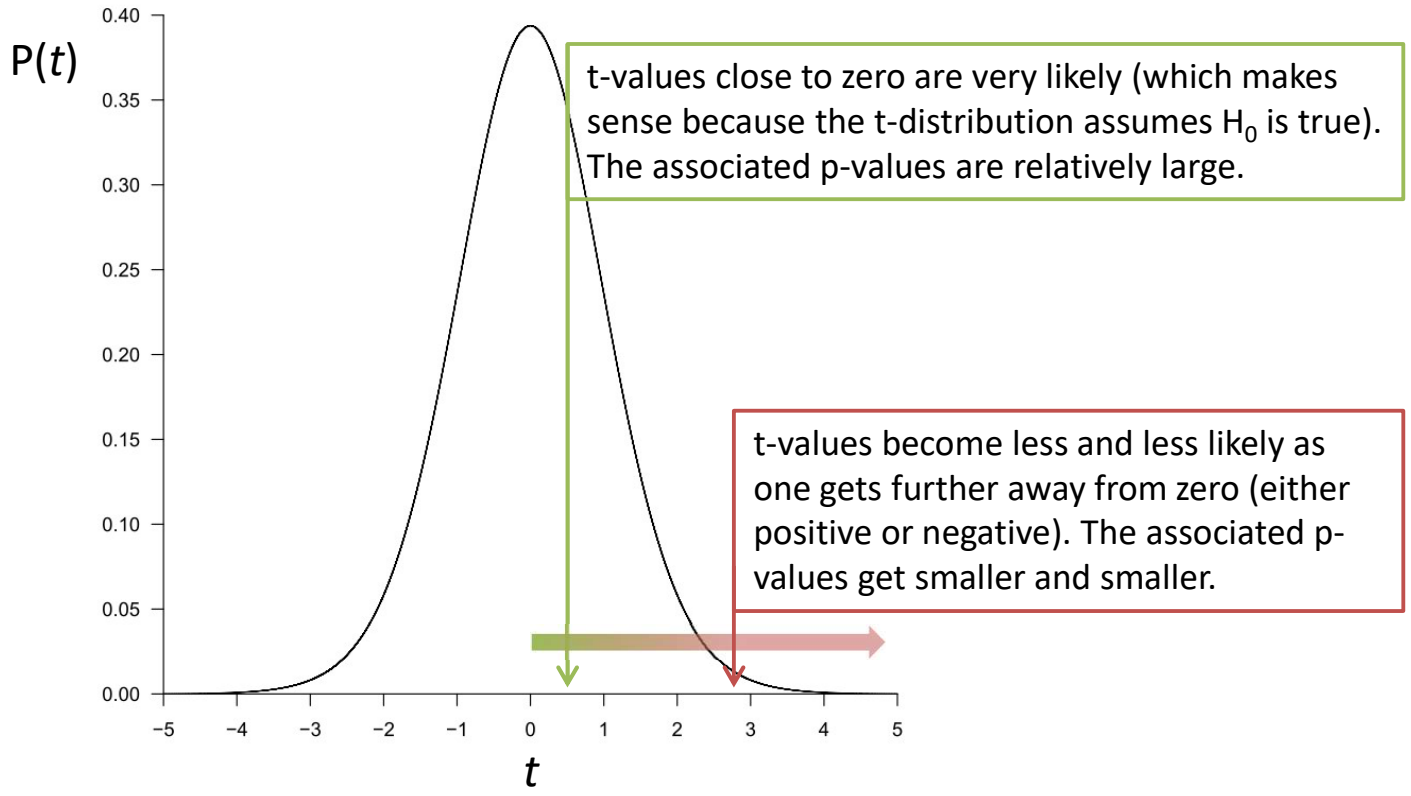
- For a given t-value, **the area under the curve gives us the p-value**.

If H_0 was true, we should most often obtain a t-value close to 0.



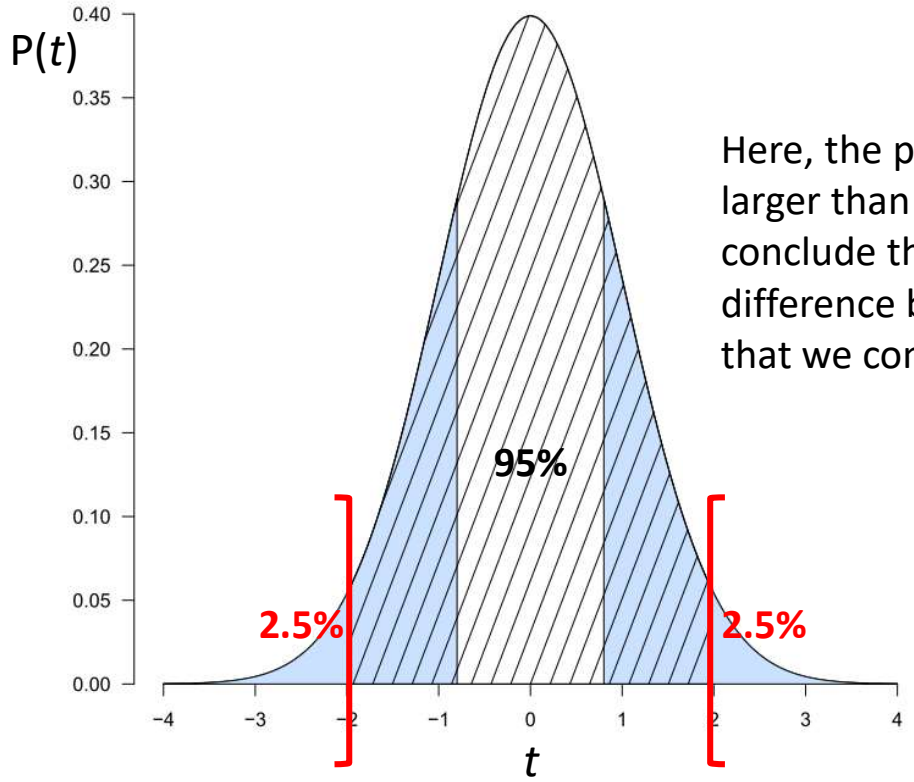
Student's two-sample t-test

- The larger the t-value, the smaller the p-value. They are just two ways of looking at the same thing: how extreme our data are relative to H_0.



Student's two-sample t-test

- Finally, we must **compare the p-value to our significance level** (α=0.05)
  - If p-value > 0.05, we cannot reject H_0, the difference is not significant
  - If p-value < 0.05, we can reject H_0, the difference is significant

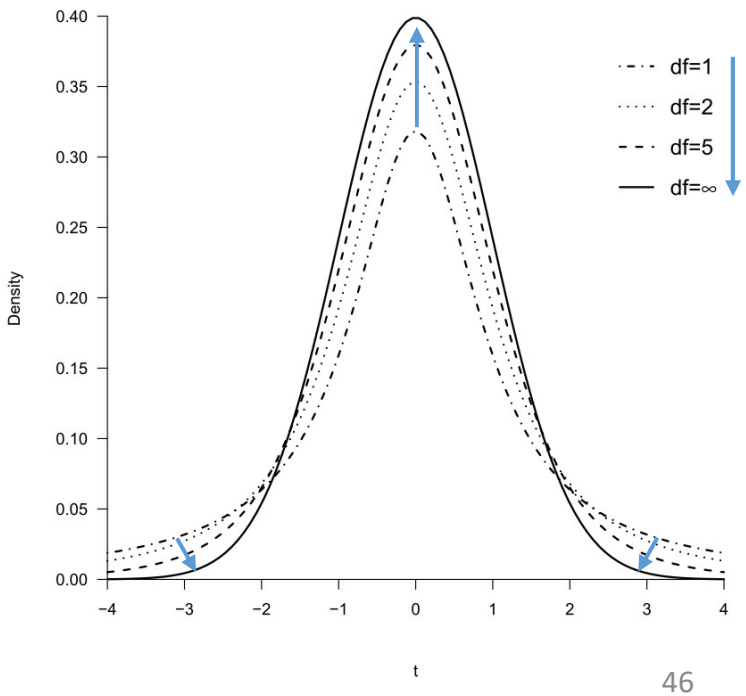


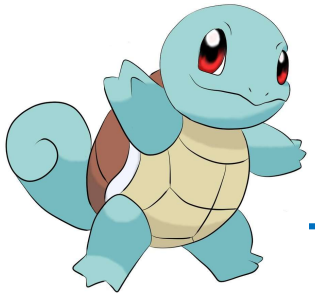
Here, the p-value (area in blue) is larger than the critical 0.05. we conclude that there is no significant difference between the two groups that we compared.

Degrees of freedom

- How do degrees of freedom affect the t-test?
  - **As df increase**, the distribution becomes higher and narrower.
  - So, the range of t-values for which p-value<0.05 becomes larger.
  - This means we have a **greater capacity to detect a mean difference as significant**.

- Remember that degrees of freedom depend on sample size!





Student's t-test in R

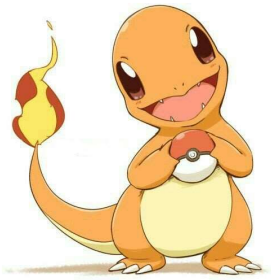
Student's t-test

R calculates the t-value and performs the t-test by calling:
t.test(Data$Response~Data$Explanatory)

Expected output:

```
            Welch Two Sample t-test

data:  Data$Response by Data$Explanatory
t = XXX, df = XXX, p-value = XXX
alternative hypothesis: true difference in means <is not> equal to 0
95 percent confidence interval:
 XXX XXX
sample estimates:
mean in group Expl_level1 mean in group Expl_level2
 XXX XXX
```



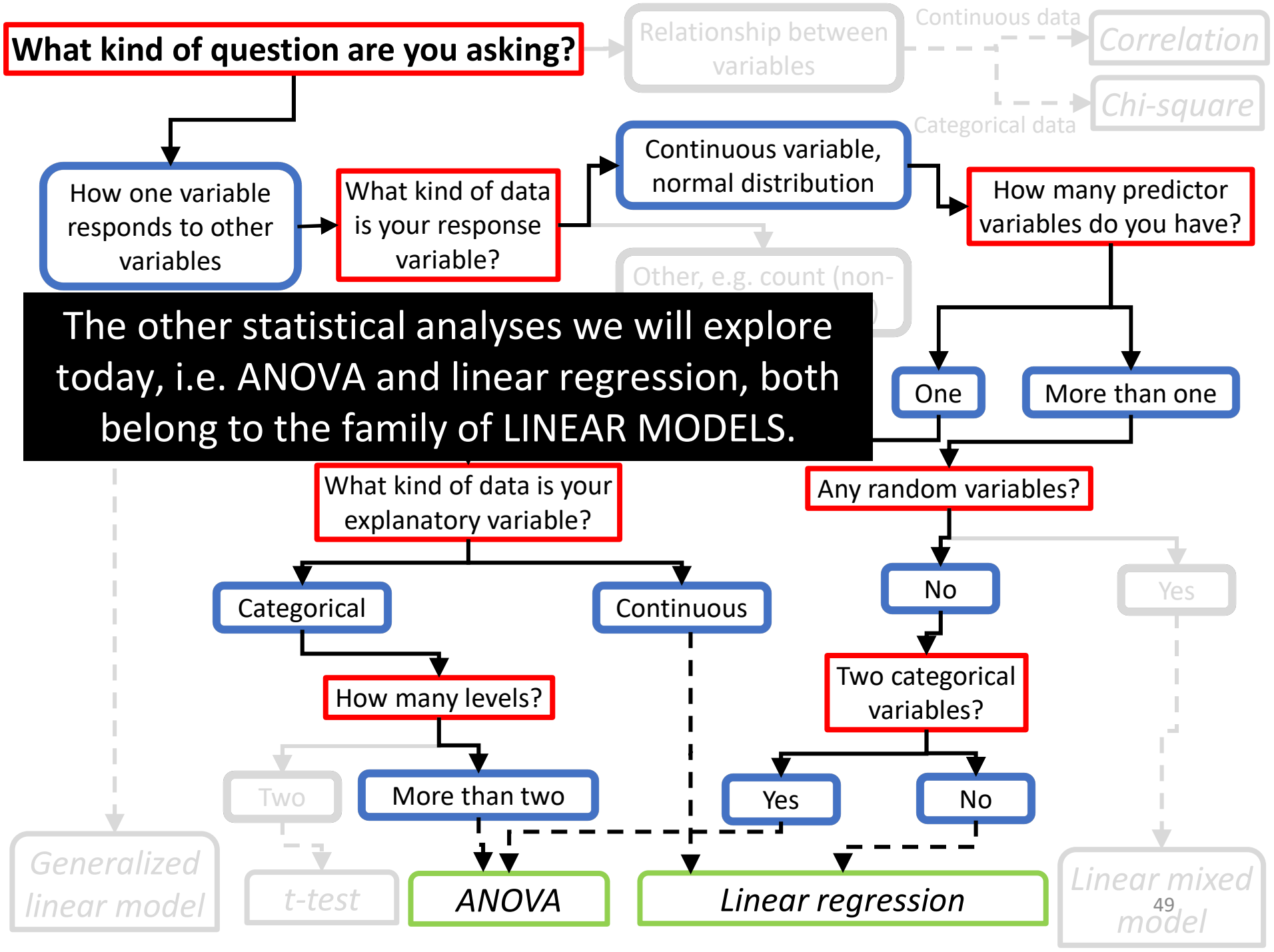
Time for an exercise

With our leaf traits dataset, we want to know the relationship (response) between:

- N-fixing ability (yes/no) and leaf N content

Steps:

1) Load ‘leaftraits.txt’
2) Examine file structure
3) Run t-test for the appropriate variables


What kind of question are you asking?
Relationship between variables
Continuous data
Correlation
Chi-square
Categorical data
How one variable responds to other variables
What kind of data is your response variable?
Continuous variable, normal distribution
Other, e.g. count (non-
How many predictor variables do you have?
The other statistical analyses we will explore today, i.e. ANOVA and linear regression, both belong to the family of LINEAR MODELS.
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
49

Linear models

- A linear model describes how one variable responds (or varies) with one or more other variables.

- It can be written as:

ε_i: error, i.e. unexplained variation (**residuals**)

$$Y_i = \alpha + \beta * X_i + \varepsilon_i$$

Y_i : response variable (dependent variable)

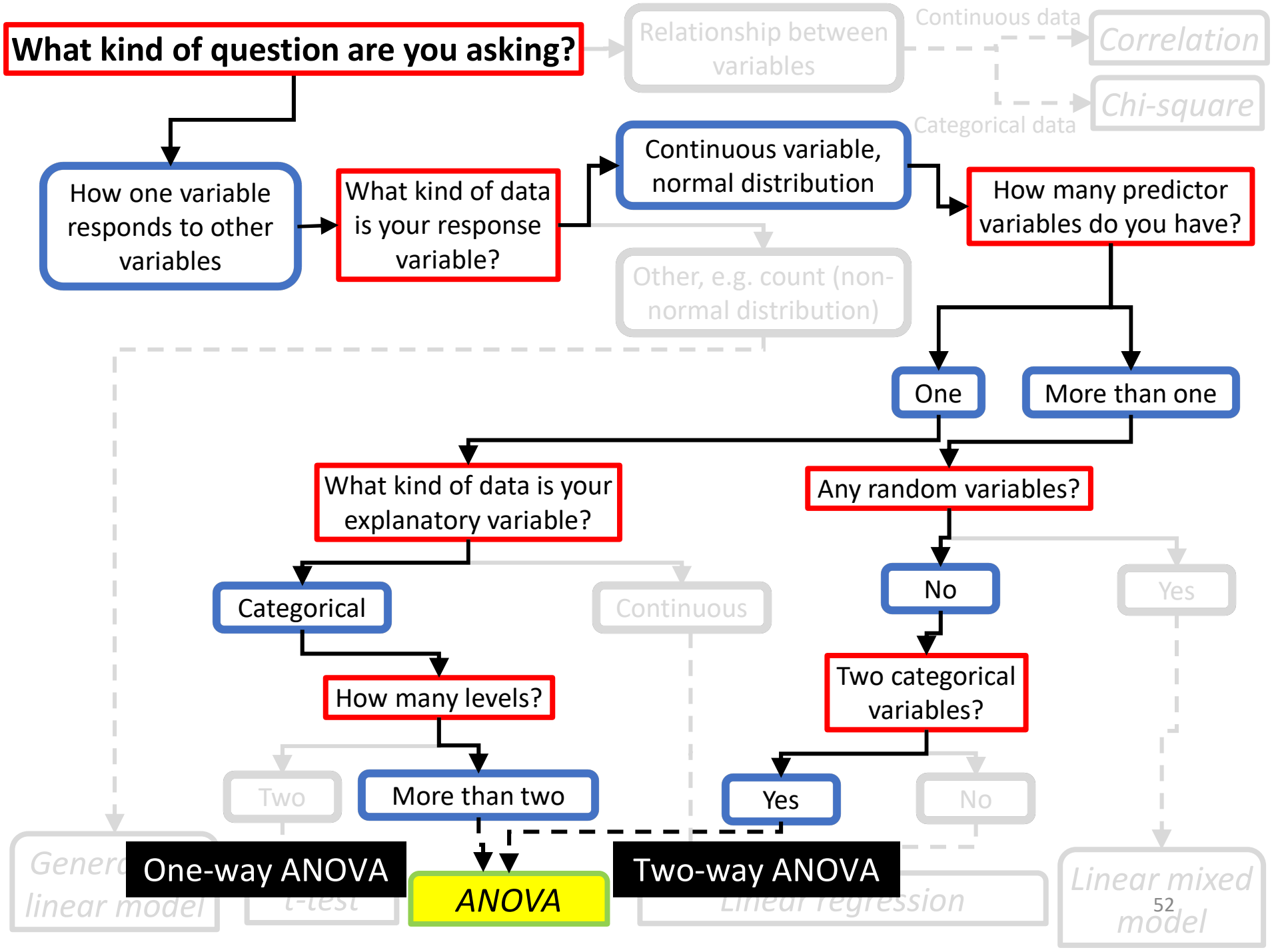
X_i : explanatory variable (independent variable)

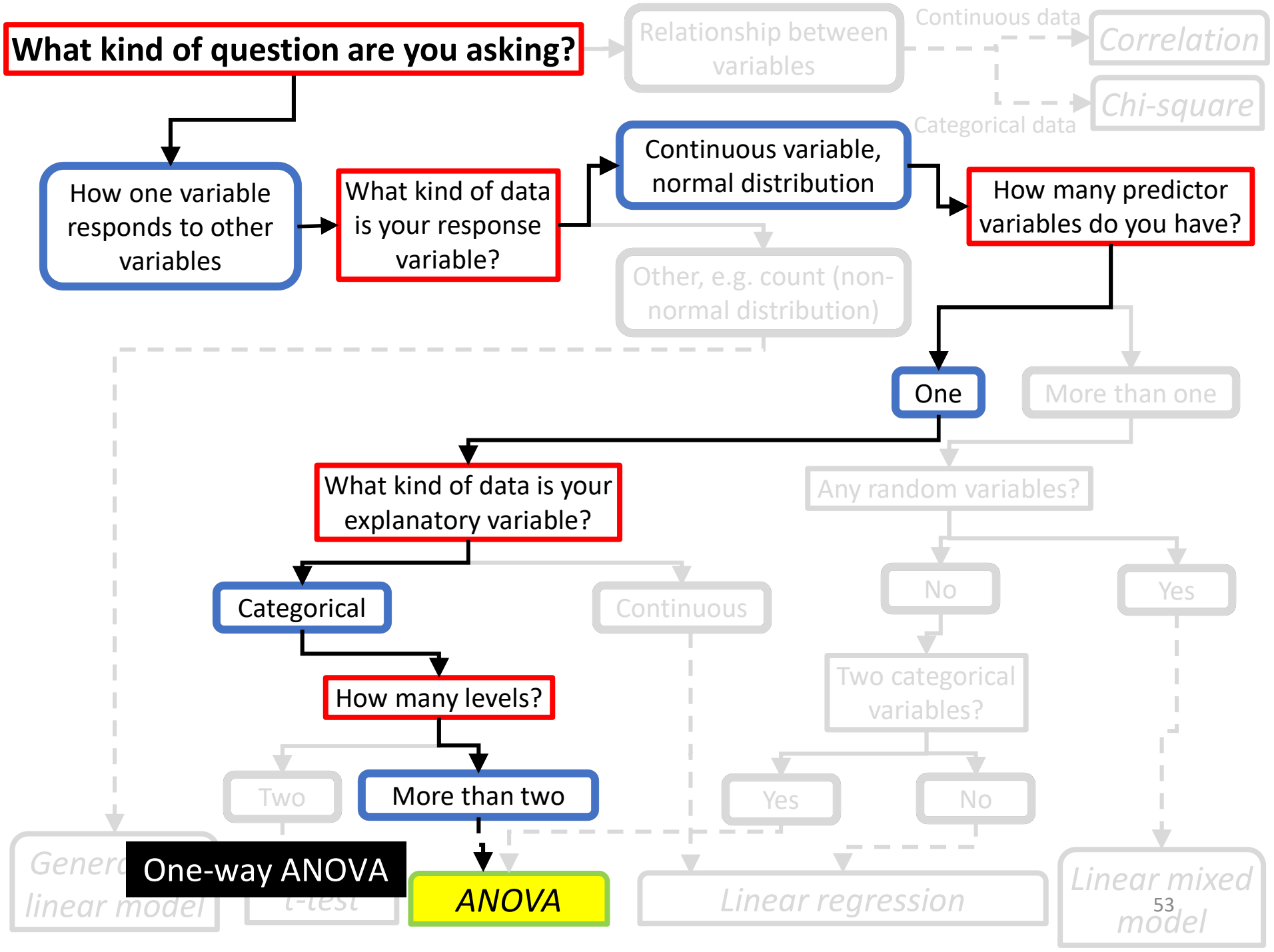
α : the **intercept** (value when X=0)

β : the **slope/strength of the X-Y relationship** (value by which Y changes with X)

Linear models: assumptions

- Linear models imply several **assumptions**:
  - **Normality of residuals**:
    for each x value, y values should be distributed normally, i.e. $\varepsilon_i \sim N(0, \sigma^2)$
  - **Homogeneity of variance**:
    across x values, y values should maintain equal variance
  - **Independence**:
    one sample value should not depend on another sample value

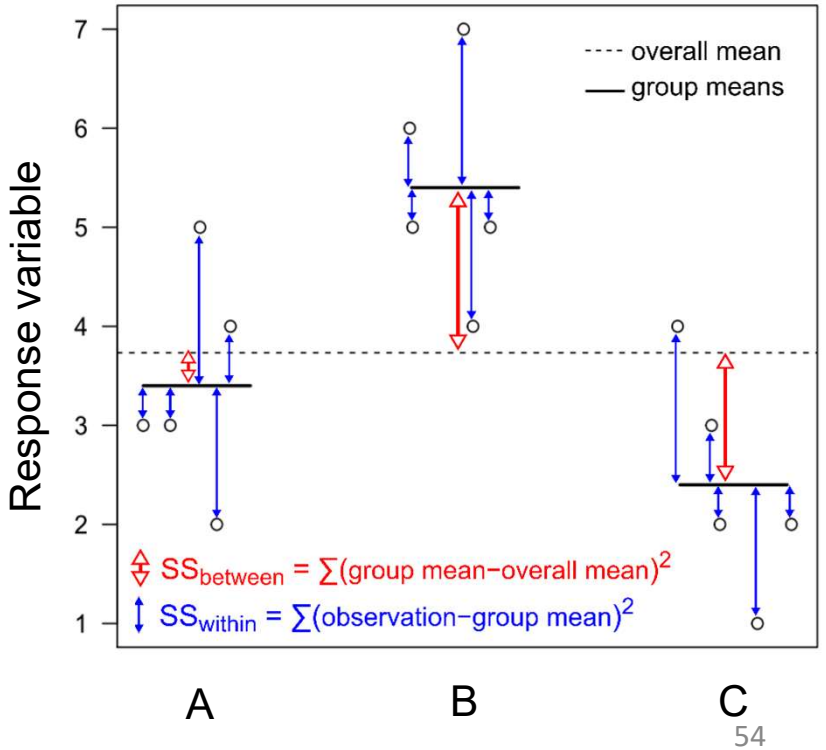

What kind of question are you asking?
Relationship between variables
Continuous data
Correlation
Chi-square
Categorical data
How one variable responds to other variables
What kind of data is your response variable?
Continuous variable, normal distribution
Other, e.g. count (non-normal distribution)
How 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
One-way ANOVA
ANOVA
Two-way ANOVA
Linear mixed model


What kind of question are you asking?
Relationship between variables
Continuous data
Correlation
Chi-square
Categorical data
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
One-way ANOVA
ANOVA
Linear regression
Linear mixed model

One-way ANOVA

- ANOVA works by separating the variation in your data into two types:
  - Variation between groups ('explained')
  - Variation within groups ('unexplained' or 'residual')

- The ratio of variation between groups to variation within groups is called **F-ratio**.



One-way ANOVA

	df	SS	MS	F
Model	$df_{between}$	$SS_{between} = \sum_{j=1}^{K} n_j(\bar{x}_j - \bar{x}_{overall})^2$	$MS_{between} = \frac{SS_{between}}{df_{between}}$	$\frac{MS_{between}}{MS_{within}}$
Residuals	df_{within}	$SS_{within} = \sum_{j=1}^{K}\sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2$	$MS_{within} = \frac{SS_{within}}{df_{within}}$	

$df_{between}$ = number of groups – 1

df_{within} = number of observations – 1 – df_{betwe}

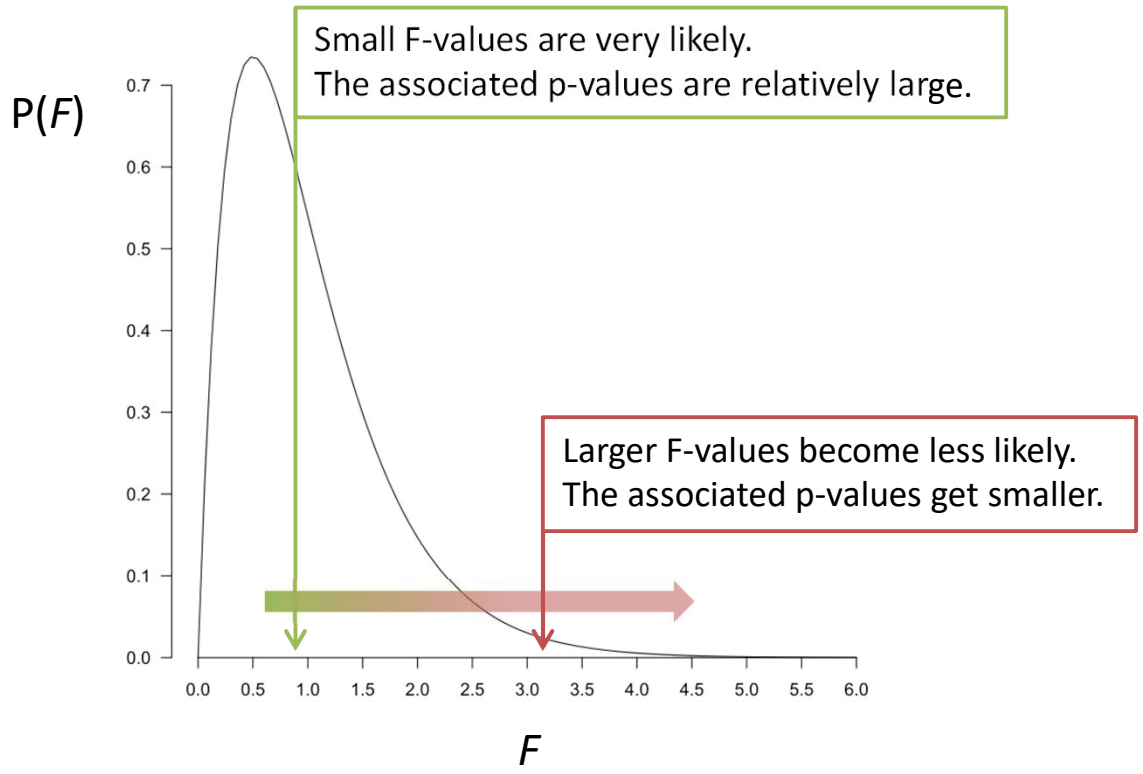
F-ratio=

$$\frac{\text{Explained variation}}{\text{Unexplained variation}}$$

Note that the F-ratio can only be positive!

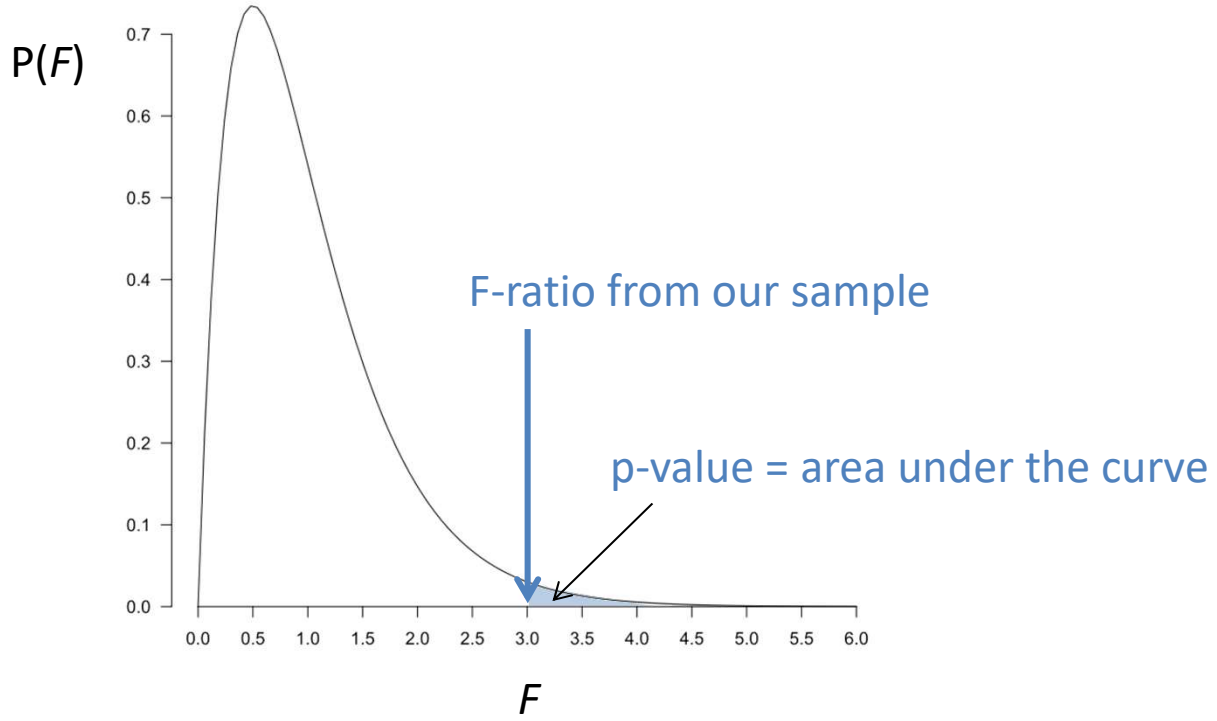
One-way ANOVA

- The **F-ratio** (or F-value) is the **test statistic** of the ANOVA.
- We **compare our sample F-ratio to the F distribution**, i.e. the expected distribution of F-values under H_0.



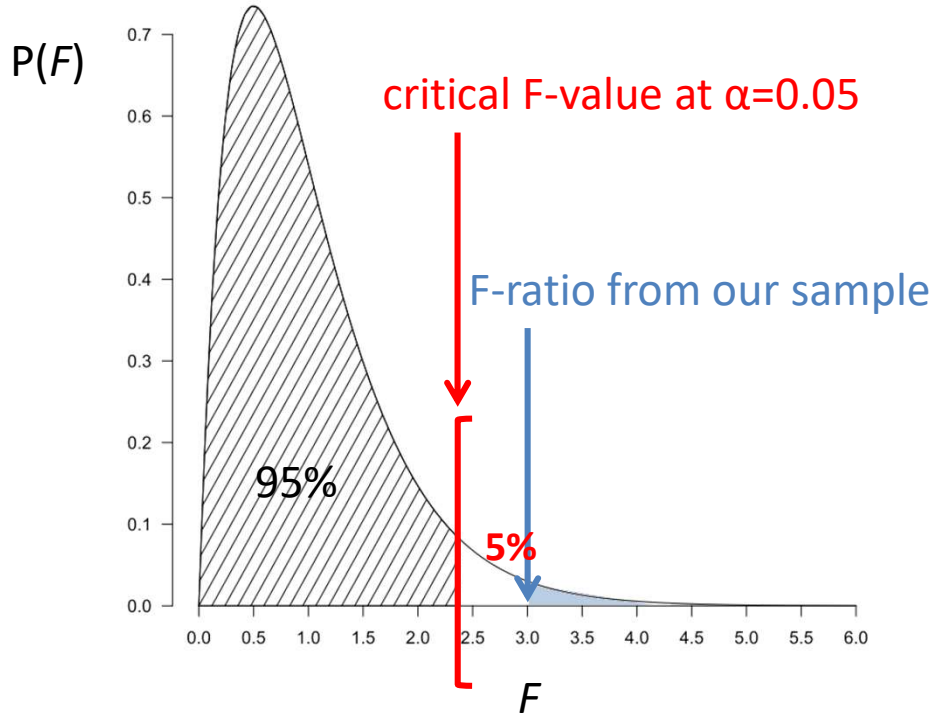
One-way ANOVA

- For a given F-ratio, **the area under the curve gives us the p-value**.



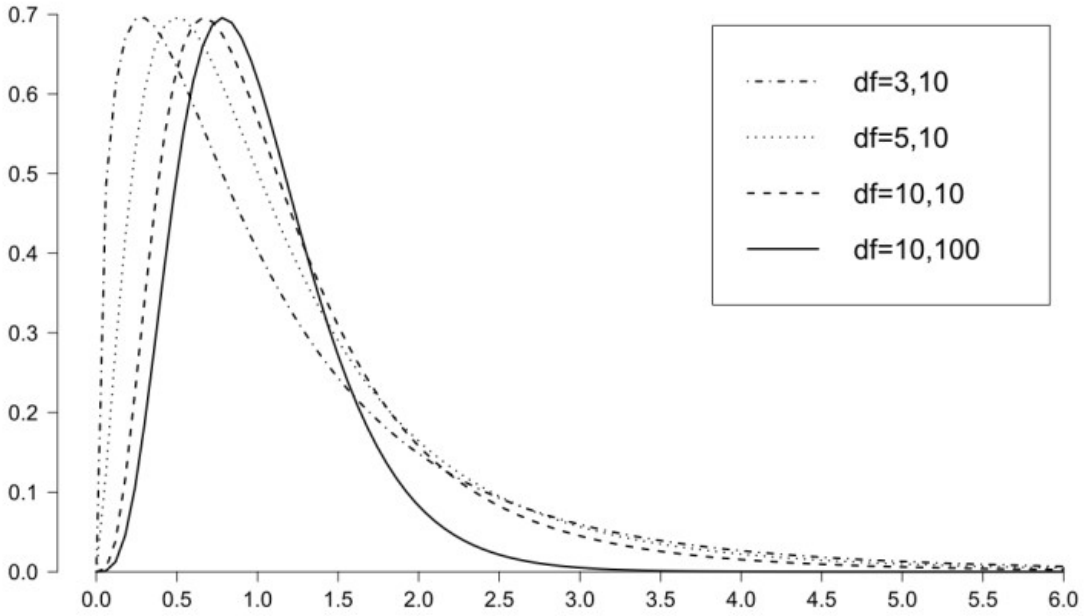
One-way ANOVA

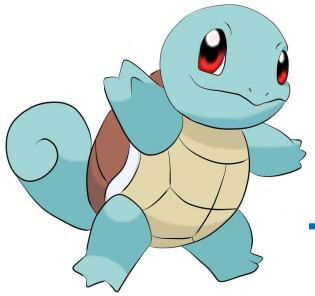
- Again, we **compare the p-value to our significance level** (α=0.05)
  - If p-value > 0.05, we cannot reject H_0, the difference is not significant
  - If p-value < 0.05, we can reject H_0, the difference is significant



One-way ANOVA

- Like the t distribution, the shape of the F distribution varies with the **degrees of freedom**, which depend on the **number of groups**, and on the **sample size**.





One-way ANOVA in R

One-way ANOVA

Doing an ANOVA in R requires two steps:

First, we create a model that you can examine further.
a1<-aov(Data$Response~Data$Explanatory)

Second, we get the ANOVA table, which tells us whether our response variable significantly differs or not among the different groups.
anova(a1)

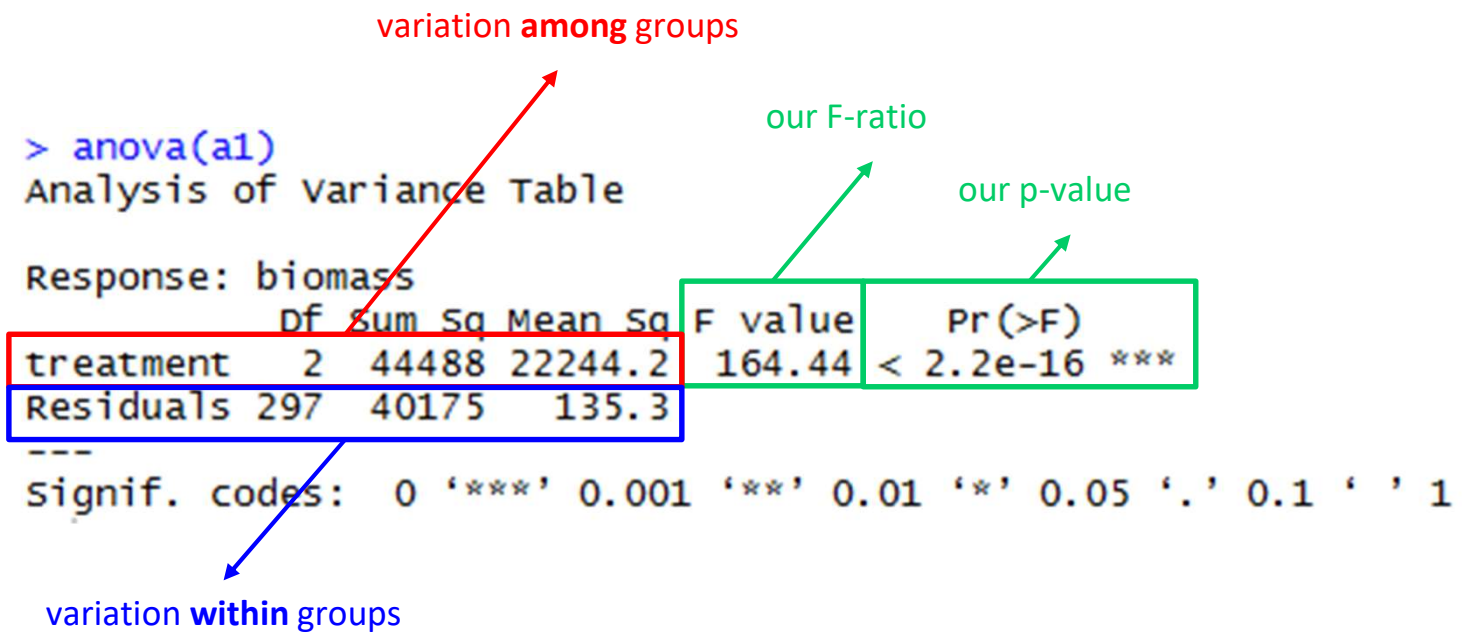
Example of one-way ANOVA

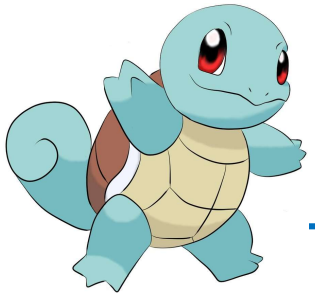
- Response to fertilization by an invasive and a native species
- In an experiment, we have grown two species:
  - Solidago canadensis
  - Eupatorium cannabinum
- At three levels of fertilization (50 plants each per species):
  - Control (A)
  - Fertilization (B)
  - Fertilization High (C)

- The total biomass was measured at the end of the experiment. Does it depend on fertilization level?

Example of one-way ANOVA

Let's go through the one-way ANOVA table





Model validation in R

ANOVA: checking model assumptions

Both assumptions (normality of residuals, variance homogeneity) can be checked using plots:

par(mfrow=c(2,2))
plot(a1)

The variance homogeneity assumption can also be tested statistically, using a Bartlett test:

bartlett.test(Data$Response,Data$Explanatory)

*Note that here, we **want** p-value > 0.05 (this means the groups **do not** have significantly different variances!)*

Expected output:

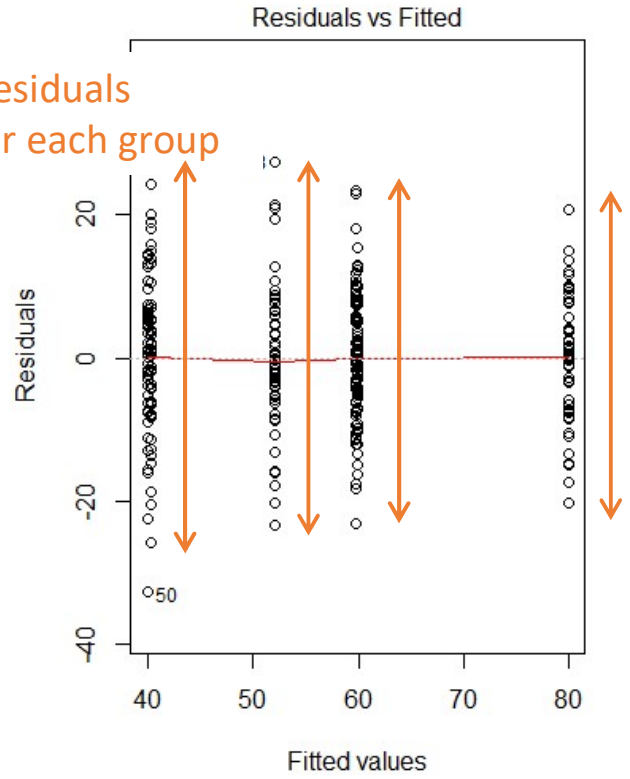
```
        Bartlett test of homogeneity of variances

data:  Data$Response by Data$Explanatory
Bartlett's K-squared = XXX, df = XXX, p-value = XXX
```

Model validation in R

Plots of residuals

The spread of the residuals should be similar for each group



The plot of residuals vs. fitted values is useful to check for variance heterogeneity.

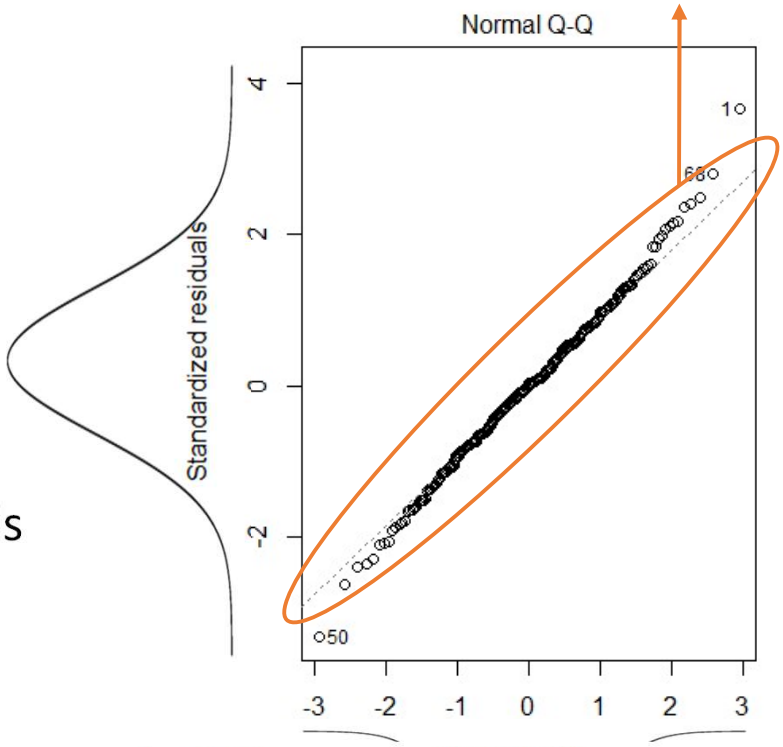
Model validation in R

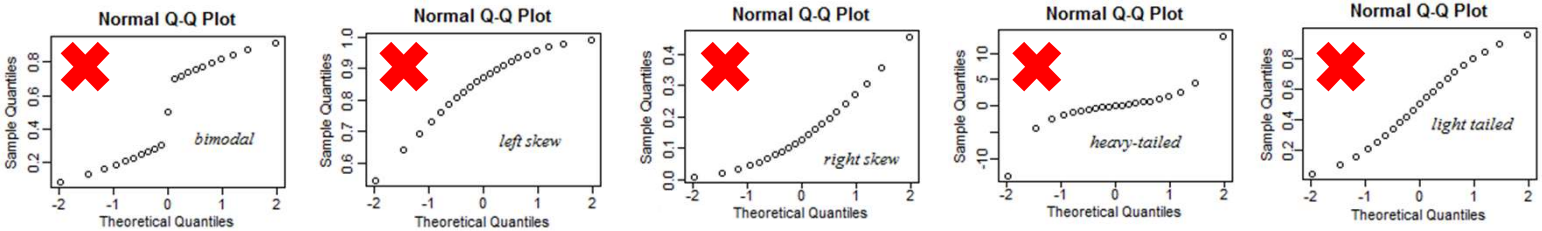
Plots of residuals

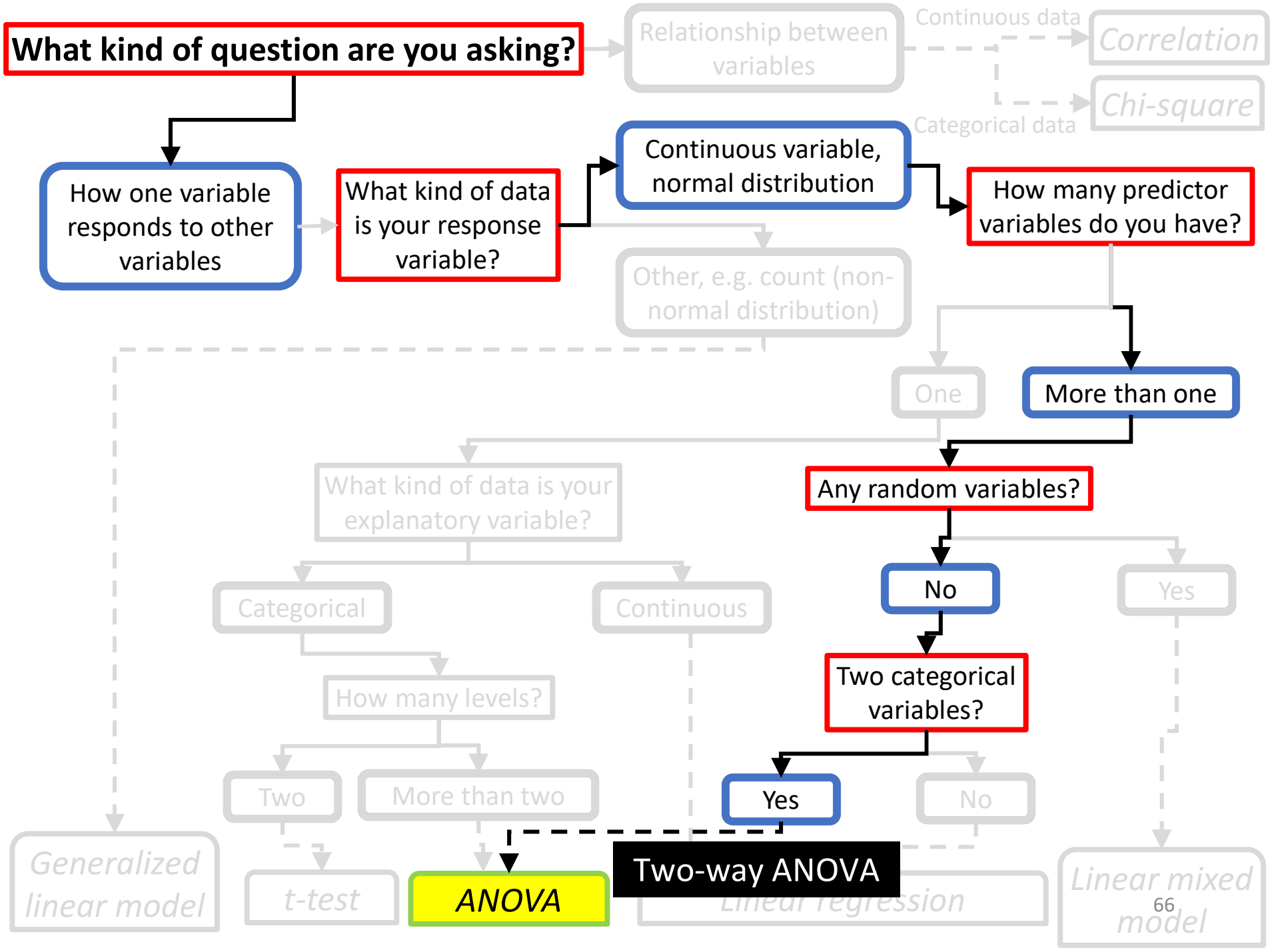
The so-called Q-Q plot is useful to check for normality of residuals.

'Q' stands for 'Quantile'. The Q-Q plot plots quantiles from the distribution of the residuals against quantiles from a normal distribution N(0,1). If the residuals are normally distributed, the dots should lie on a 1:1 line.

Residuals should lie on a 1:1 line



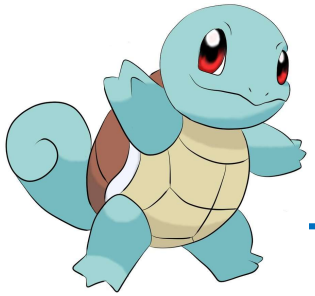



What kind of question are you asking?
Relationship between variables
Continuous data
Correlation
Categorical data
Chi-square
How one variable responds to other variables
What kind of data is your response variable?
Continuous variable, normal distribution
Other, e.g. count (non-normal distribution)
How 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
Two-way ANOVA
Linear regression
Linear mixed model

Two-way ANOVA

- What if we have two categorical explanatory variables, and both could explain variation in our data?

- In our fertilization example, we could expect that biomass does not only depend on fertilization level but also differs overall between the two species.



Two-way ANOVA in R

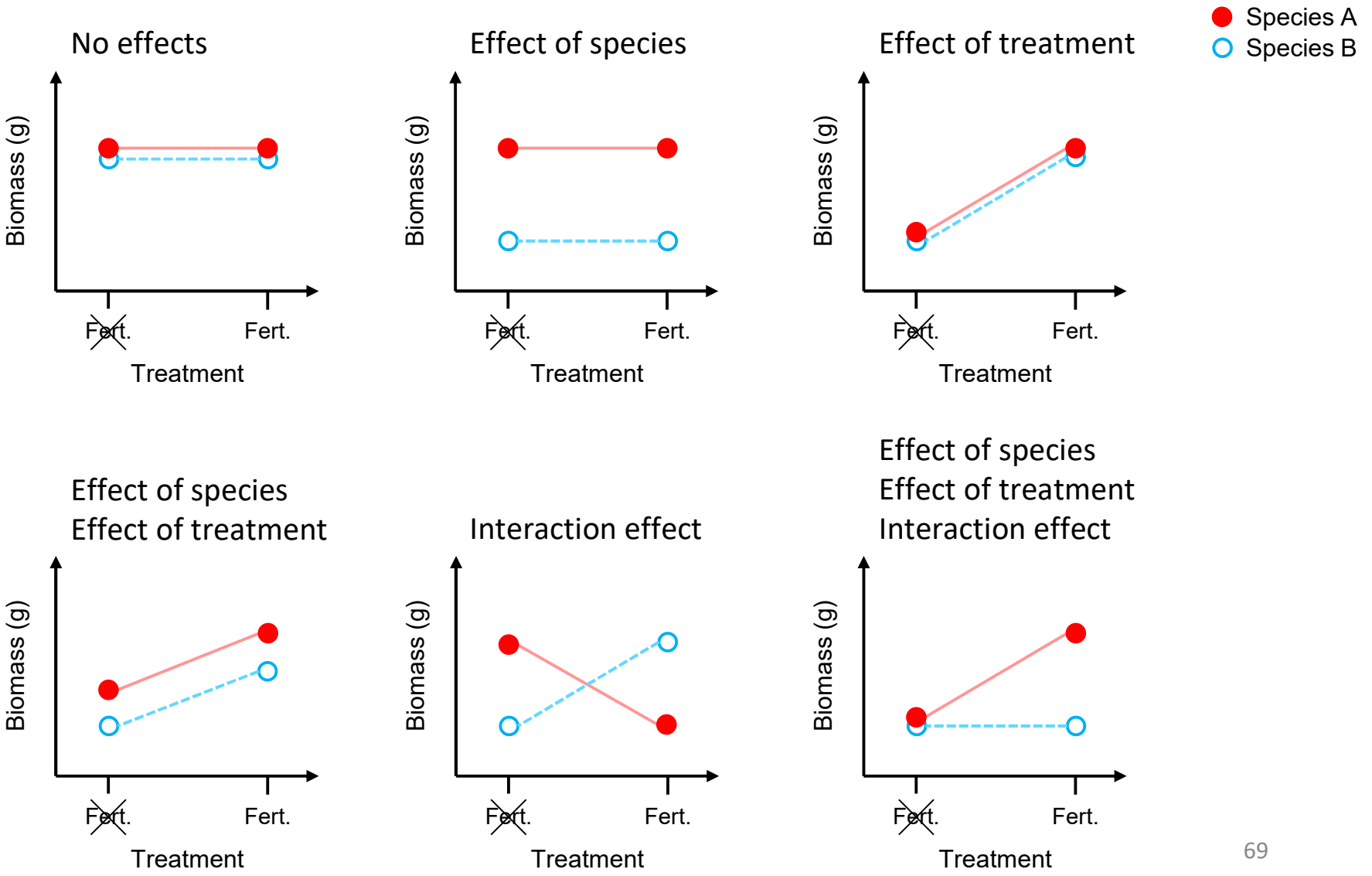
Two-way ANOVA with additive effects (as opposed to interactive)

Same procedure as before:

First, we create a model that you can examine further.
a2<-aov(Data$Response~Data$Explanatory1+Data$Explanatory2)

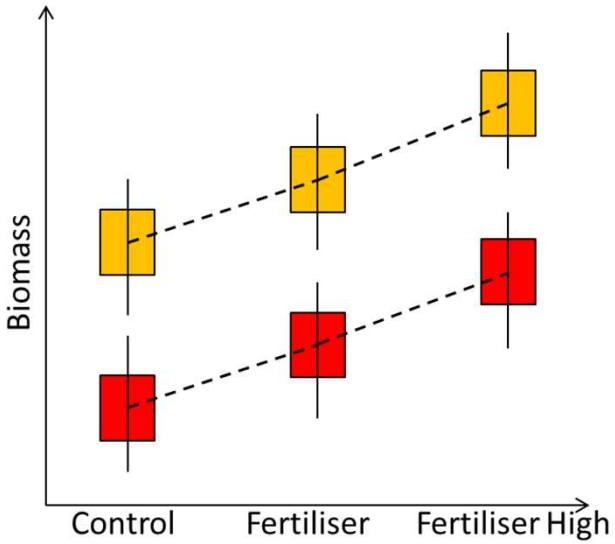
Second, we get the ANOVA table.
anova(a2)

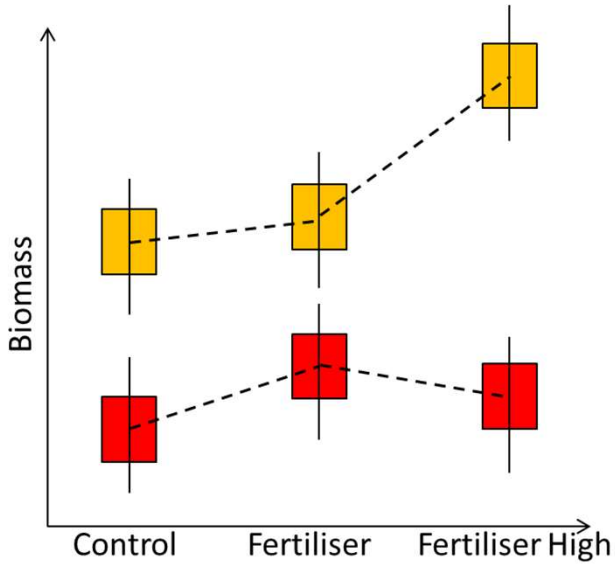
The notion of interaction

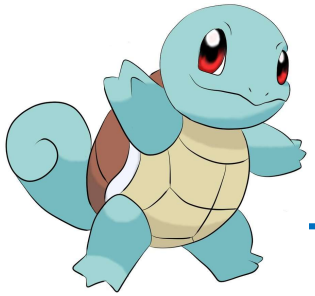


Two-way ANOVA

- So, in our example:
  - Maybe both species respond in a similar way to fertilization…
  - Or, maybe they respond differently?
  - So, do the two variables **interact** to affect the response (biomass)?







Two-way ANOVA in R

Two-way ANOVA with interactive effects

Same procedure as before:

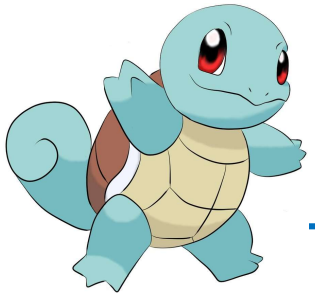
First, we create a model that you can examine further.

*a2int<-aov(Data$Response~Data$Explanatory1*Data$Explanatory2)*

Second, we get the ANOVA table.

anova(a2int)

The asterisk indicates the testing of an interaction between two variables.



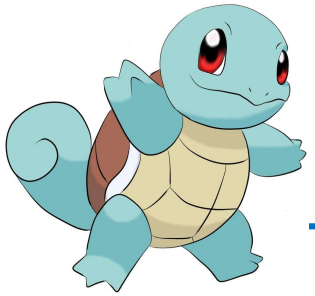
Two-way ANOVA in R

Let's go through the two-way ANOVA table

```
> anova(a2)
Analysis of Variance Table

Response: biomass
                  Df Sum Sq Mean Sq F value    Pr(>F)
treatment          2  44488 22244.2 227.684 < 2.2e-16 ***
species            1   6492  6492.4  66.454 1.043e-14 ***
treatment:species  2   4960  2479.8  25.383 6.777e-11 ***
Residuals        294  28723    97.7
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

There is a significant interaction,
i.e. the two species respond significantly differently to fertilization.

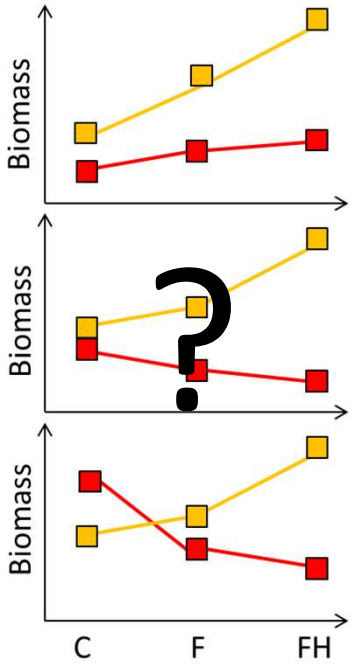


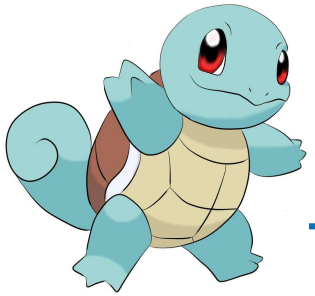
Two-way ANOVA in R

Post-hoc testing

When the interaction is significant, to see HOW the two species differ in their responses to the fertilization treatment, we need to compare differences between groups, using pairwise comparisons.

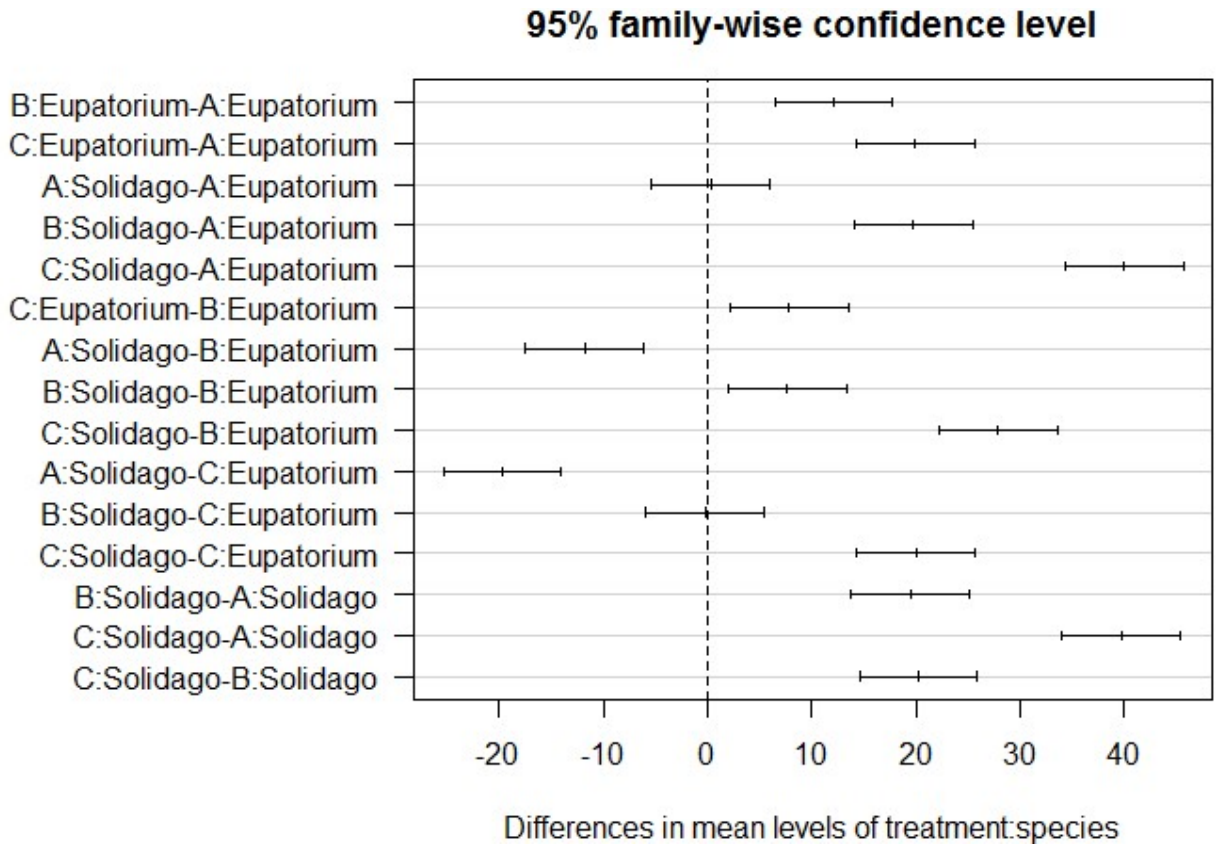
We can use Tukey's HSD (Honest Significant Difference) tests:
TukeyHSD(a2int)
plot(TukeyHSD(a2int), las=1)

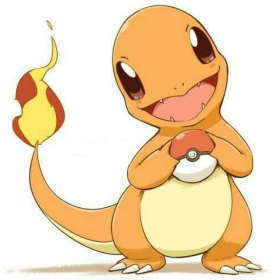




Two-way ANOVA in R

Results of Tukey'HSD post-hoc tests





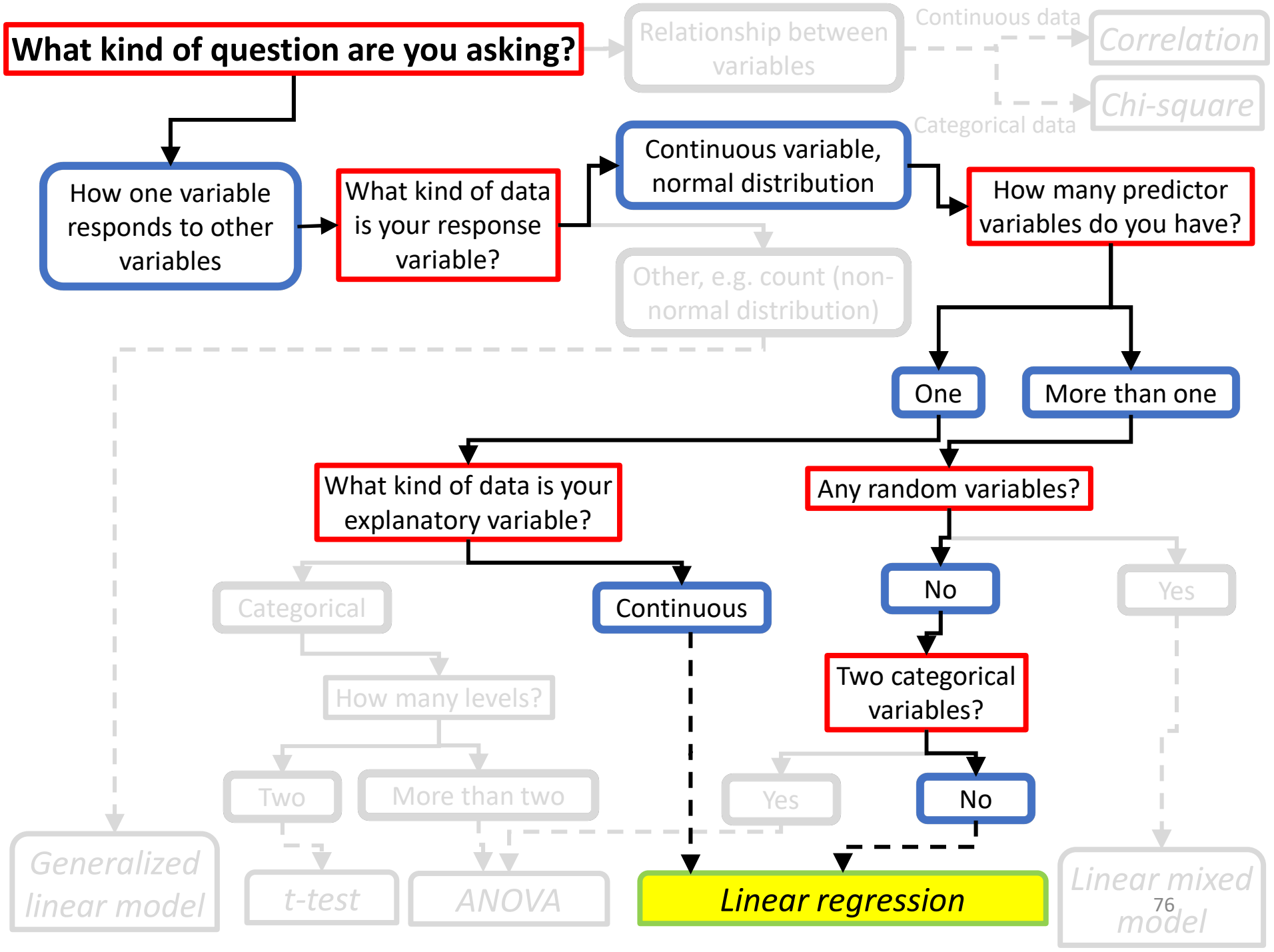
Time for an exercise

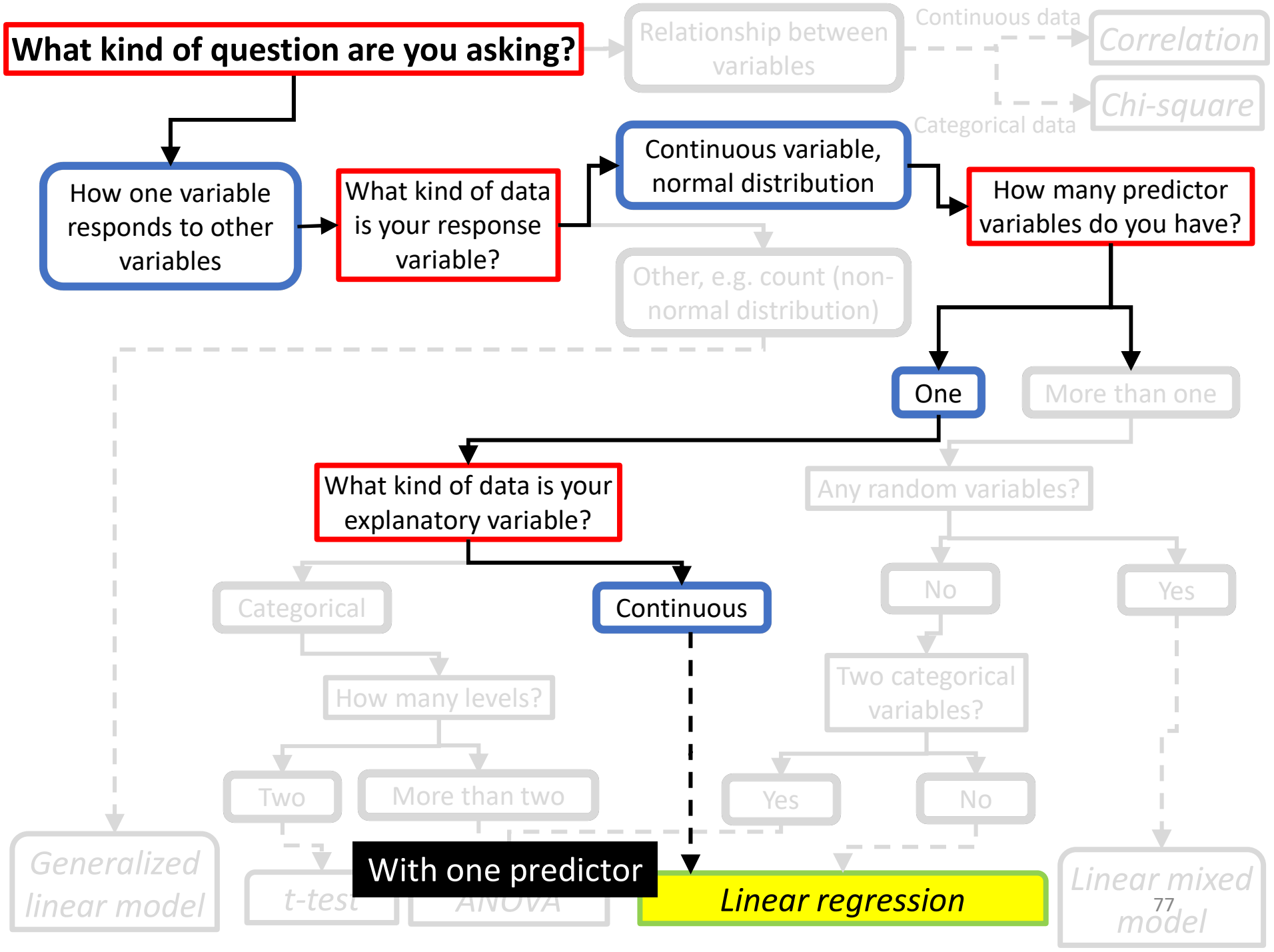
With our leaf traits dataset, we want to know the relationship (response) between:

- Leaf LMA and plant growth form
- Leaf lifespan and both N-fixing capacity (yes/no) and deciduousness (deciduous/evergreen)

Steps:

1) Load 'leaftraits.txt'
2) Examine file structure
3) Run ANOVA
4) Compare groups
5) Check model assumptions


What kind of question are you asking?
Relationship between variables
Continuous data
Correlation
Chi-square
Categorical data
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?
Relationship between variables
Continuous data
Correlation
Chi-square
Categorical data
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
With one predictor
Linear regression
Linear mixed model

Linear regression

- Remember the linear model:

ε_i: error, i.e. unexplained variation (**residuals**)

$$Y_i = \alpha + \beta * X_i + \varepsilon_i$$

Y_i : response variable (dependent variable)

X_i : explanatory variable (independent variable)

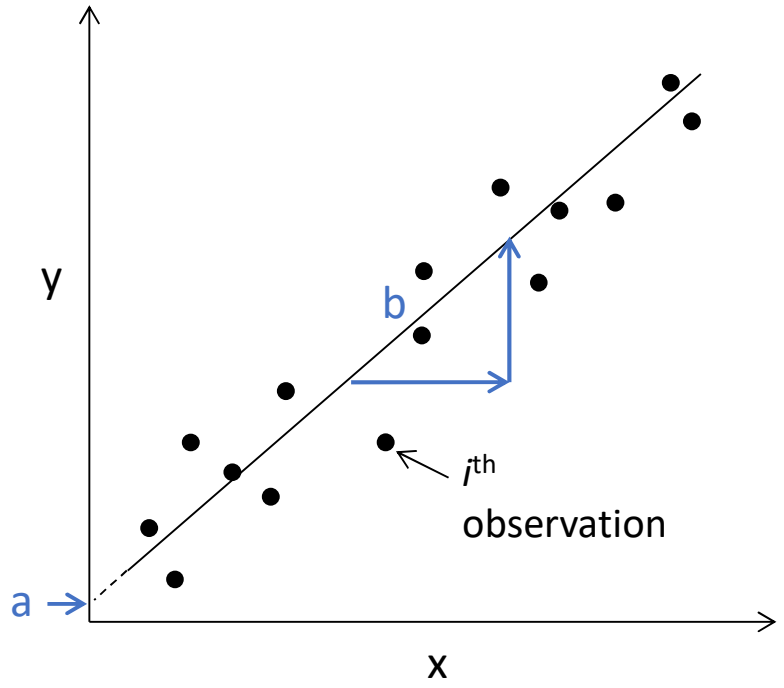
α : the **intercept** (value when X=0)

β : the **slope/strength of the X-Y relationship** (value by which Y changes with X)

Linear regression

- The relationship between two continuous variables looks like:

- When we do a linear regression, we calculate a and b (estimates of α and β)
  - Using the method of **Ordinary Least Squares** (OLS)
  - Which minimizes the difference between observed data and the predicted line ('Residual Sum of Squares')

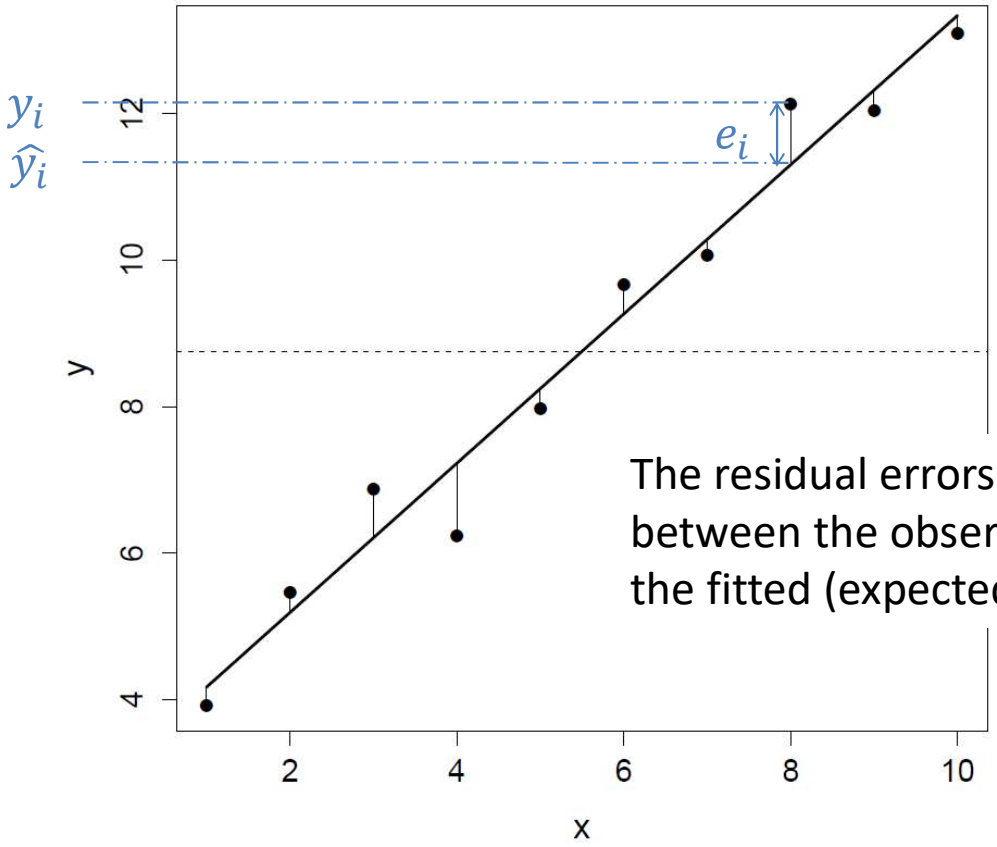


- Ultimately, what we care about is the relationship between our X and Y variables, i.e., **is b significantly different from 0?**
- Again, we do this by calculating an **F-ratio**.

Linear regression

$$SS_{residual} = \sum_i {e_i}^2 = \sum_i (y_i - \hat{y}_i)^2$$

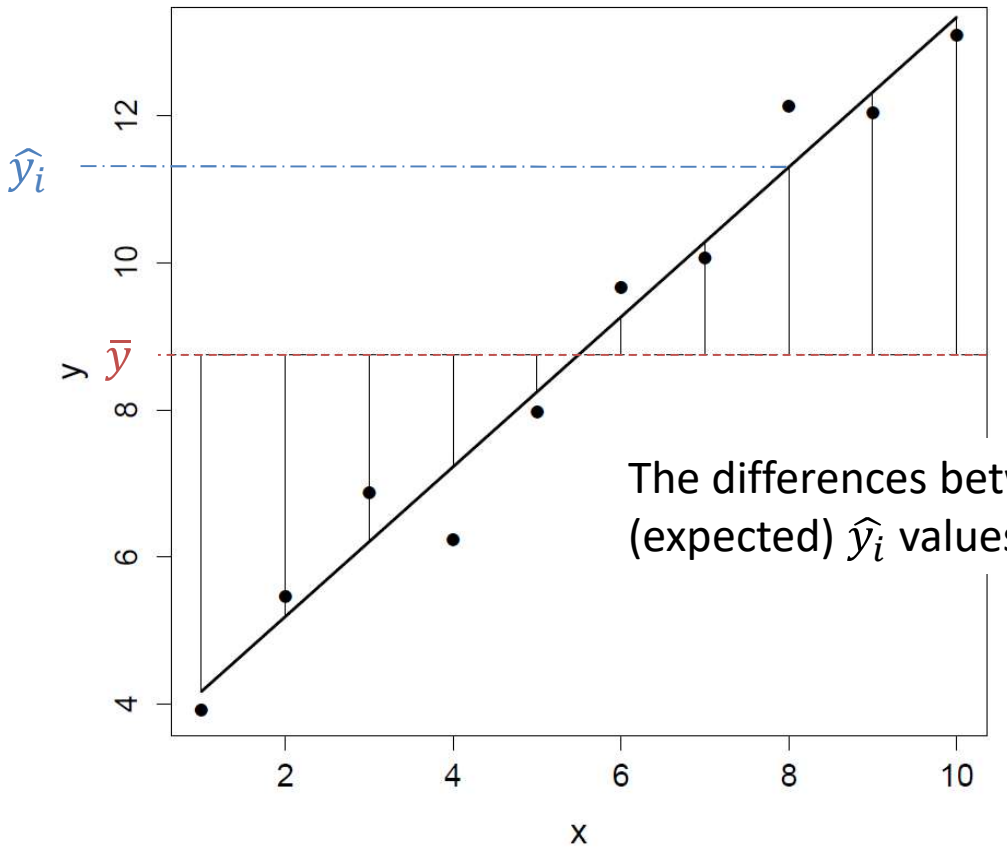
$\hat{y}_i$ = fitted values, e_i = errors (residuals)



The residual errors are the differences between the observed y_i values and the fitted (expected) $\hat{y}_i$ values

Linear regression

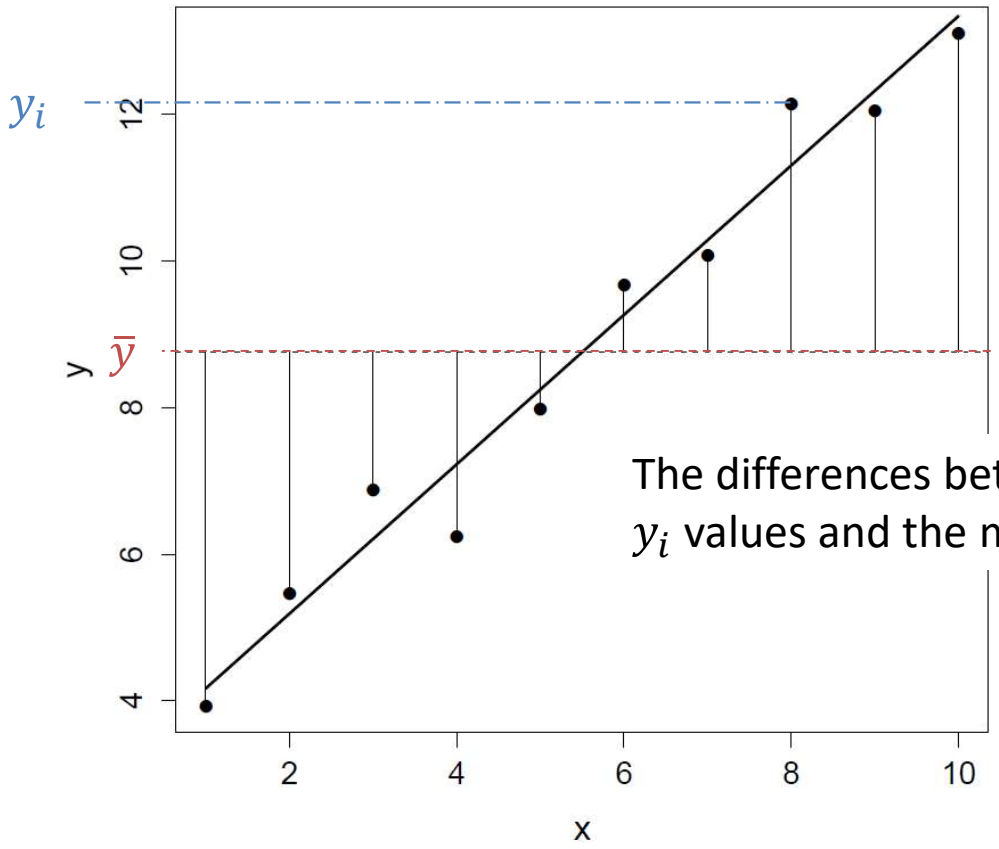
$$SS_{regression} = \sum_i (\hat{y}_i - \bar{y})^2$$



The differences between the fitted (expected) $\hat{y}_i$ values and the mean $\bar{y}$

Linear regression

$$SS_{total} = \sum_i (y_i - \bar{y})^2$$



The differences between the observed y_i values and the mean $\bar{y}$

Linear regression

	df	SS	MS	F
Model	$df_{regression}$	$SS_{regression} = \sum_i (\hat{y}_i - \bar{y})^2$	$MS_{regression} = \frac{SS_{regression}}{df_{regression}}$	$\frac{MS_{regression}}{MS_{residual}}$
Residuals	$df_{residual}$	$SS_{residual} = \sum (y_i - \hat{y}_i)^2$	$MS_{residual} = \frac{SS_{residual}}{df_{residual}}$	

$df_{regression}$ = number of parameters estimated – 1

$df_{residual}$ = number of observations – 1 – $df_{regression}$

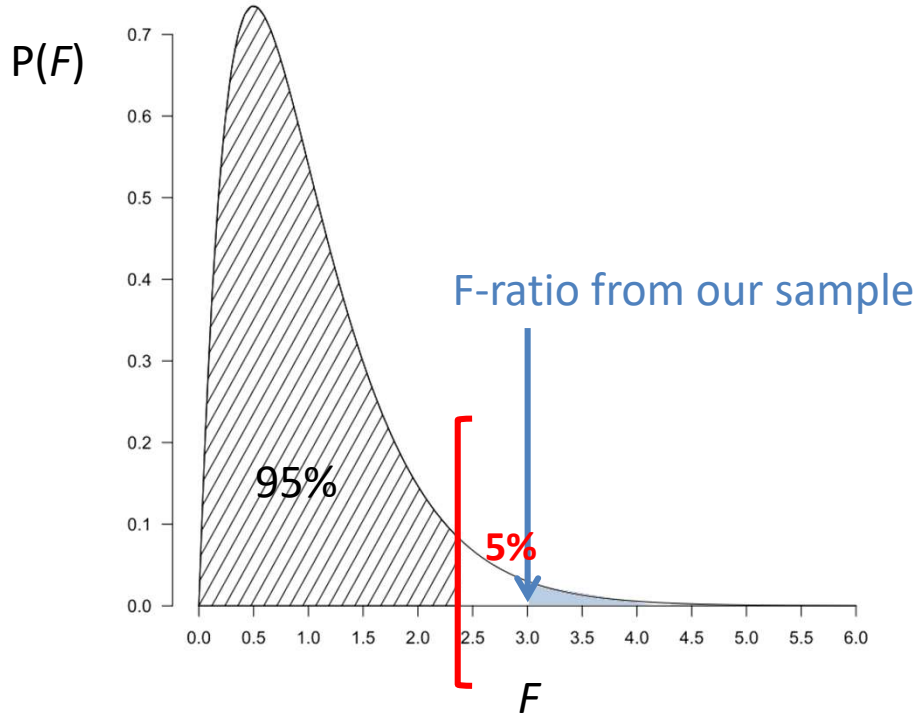
F-ratio=

$$\frac{\text{Explained variation}}{\text{Unexplained variation}}$$

Note that the F-ratio can only be positive!

Linear regression

- Again, we compare our sample F-ration to the F distribution, and we check if our p-value is larger than our significance level (α=0.05).

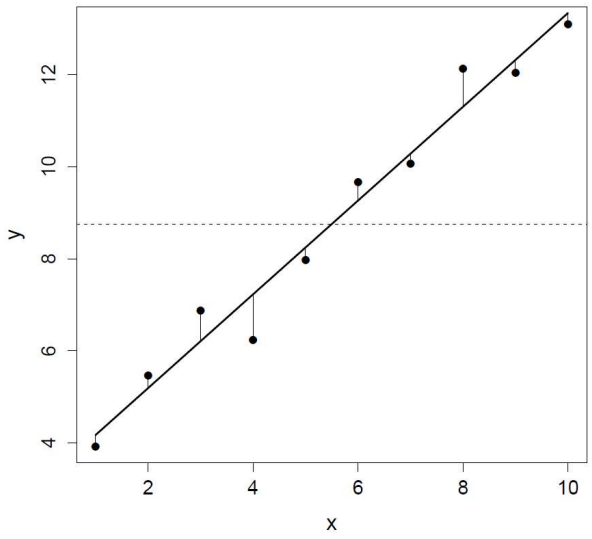


Coefficient of determination

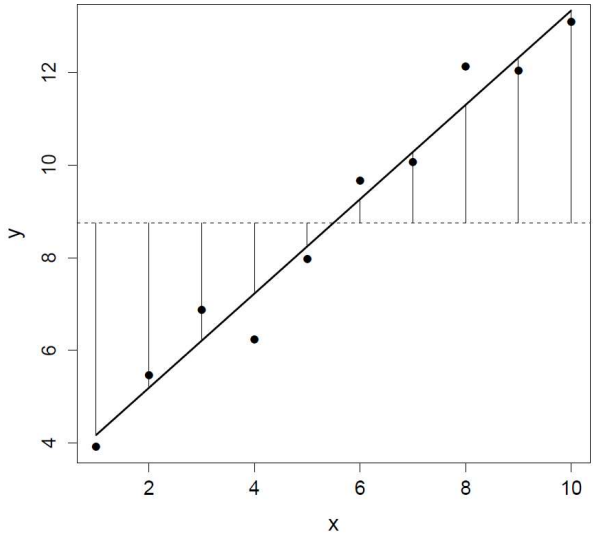
- The coefficient of determination **R^2** is the "**proportion of variance explained by the model**", i.e. the proportion of variance in the response variable that is predictable from the explanatory variable(s).

- Remember the sums of squares that we calculate in OLS regression:

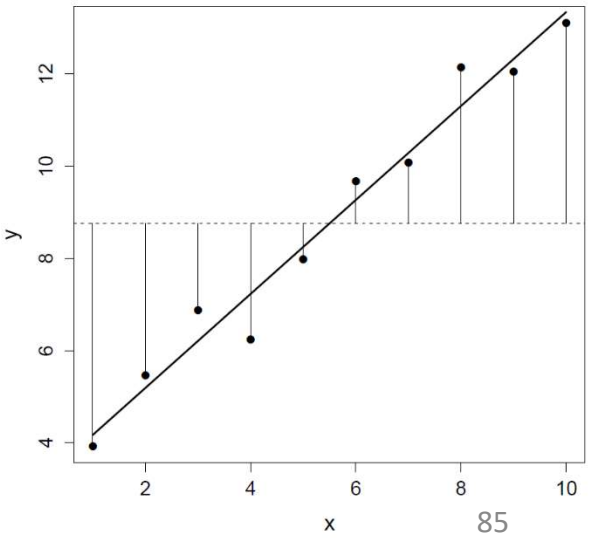
$$SS_{residual} = \sum_i {e_i}^2 = \sum_i (y_i - \hat{y}_i)^2$$



$$SS_{regression} = \sum_i (\hat{y}_i - \bar{y})^2$$



$$SS_{total} = \sum_i (y_i - \bar{y})^2$$



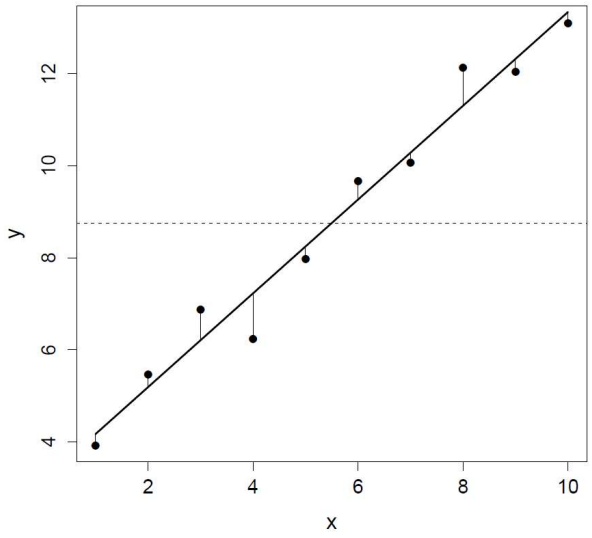
Coefficient of determination

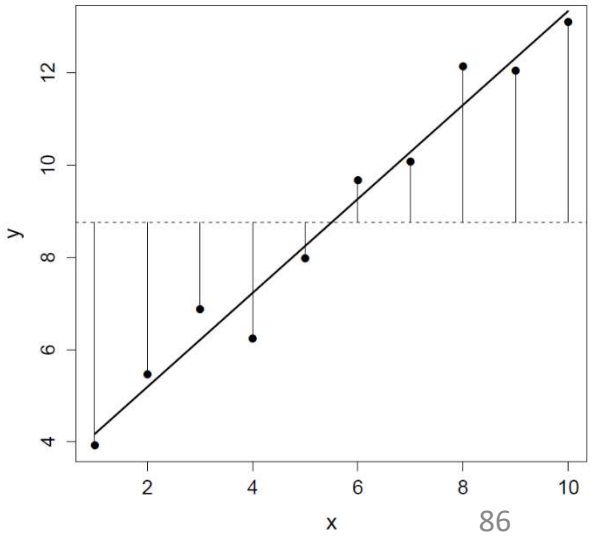
- The coefficient of determination R^2 is based on the ratio of $SS_{residual}$ (i.e. the residual variation left after calculating the variation explained by the explanatory variable(s)) and SS_{total} (i.e. the total variation in the response):

$$\mathbf{R^2 = 1 - \frac{SS_{residual}}{SS_{total}}}$$

$$SS_{residual} = \sum_i {e_i}^2 = \sum_i (y_i - \hat{y}_i)^2$$

$$SS_{total} = \sum_i (y_i - \bar{y})^2$$

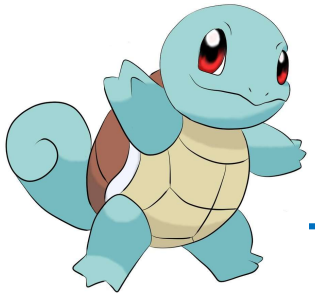




Coefficient of determination

- As we can always get a better fit (or at least not a worse fit) by adding more explanatory variables to the model, we often use an **'adjusted' R^2**; which is **penalized for including more variables** into the model:

$$\text{adj.}\, R^2 = 1 - \left(\frac{n-1}{n-p-1}\right)\left(\frac{SS_{residual}}{SS_{total}}\right)$$



Linear regression in R

Note that we use the 'lm' function, which means 'linear model'.

Linear regression

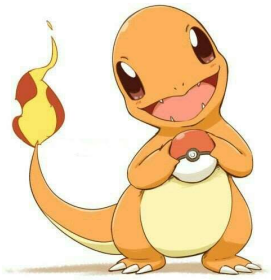
Exactly the same as for the ANOVA:

First, we create a model that you can examine further.

```
lg<-lm(Data$Response~Data$Explanatory)
```

Second, we get the summary table, which gives model parameter estimates and their significance, as well as the significance of variation explained by the model.

```
summary(lg)
```



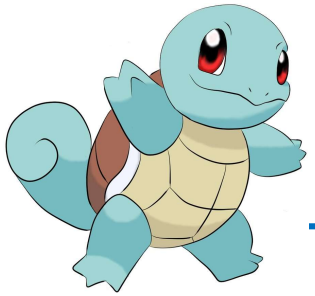
Time for an exercise

With our leaf traits dataset, we want to know the response of:

- Leaf N content to leaf LMA

Steps:

1) Load ‘leaftraits.txt’
2) Examine file structure
3) Run linear regression
4) Examine results
5) Check model assumptions

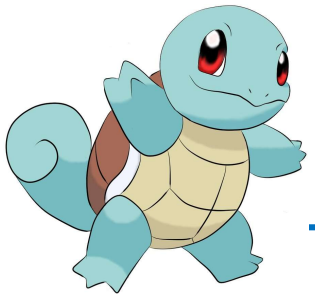


Model validation in R

Linear regression: checking model assumptions

Again, both assumptions (normality of residuals, variance homogeneity) can be checked using plots:

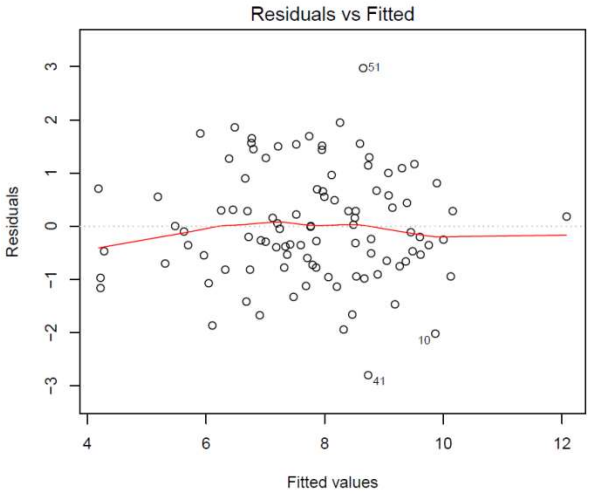
par(mfrow=c(2,2))
plot(lg)



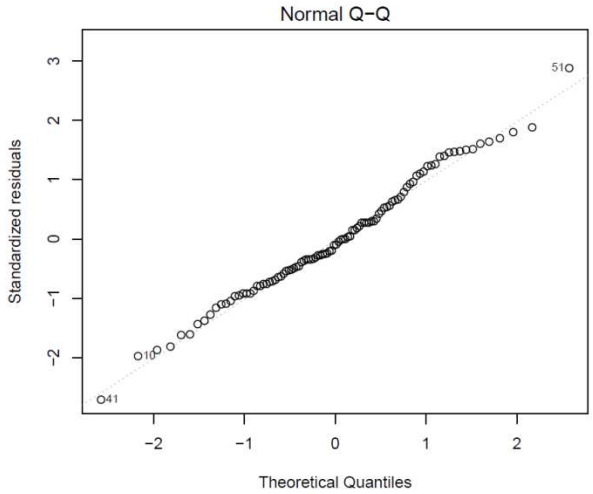
Model validation in R

Example of plots of residuals

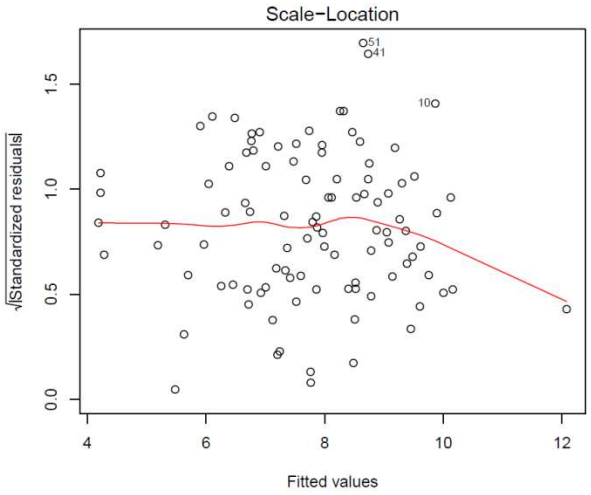
Variance heterogeneity/outliers



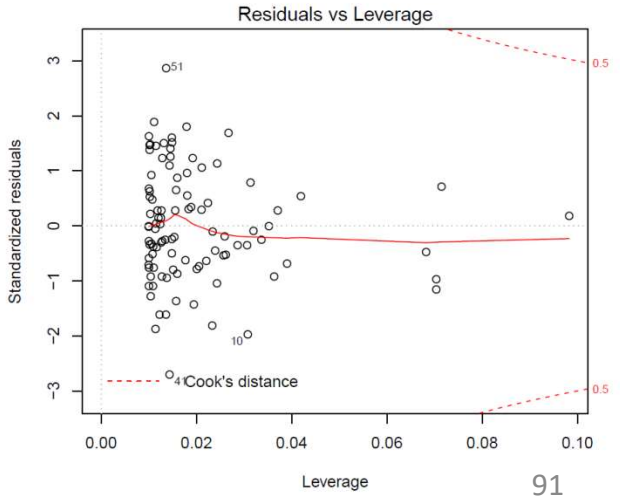
Normality/outliers



Variance heterogeneity



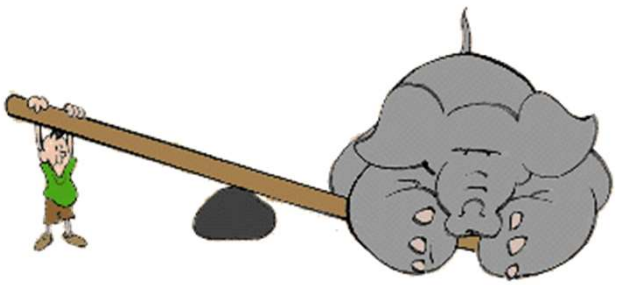
Leverage/Influential points (Cook's Distance)

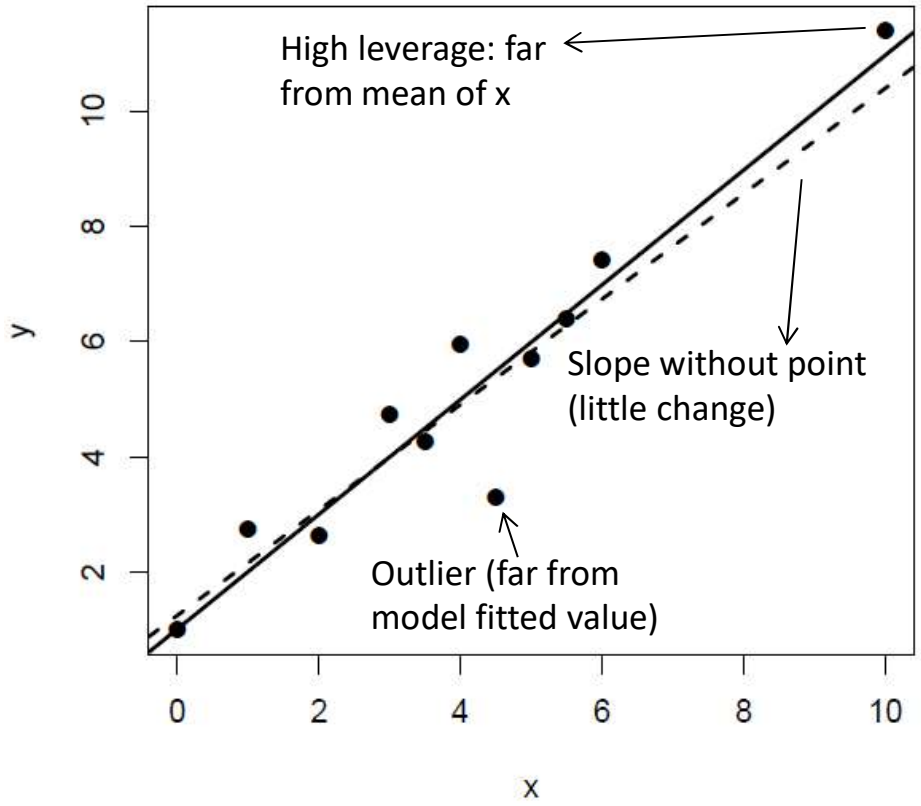


Leverage, outliers and influence

- **Outliers** are data points with very **large residuals** (far from the fitted line).

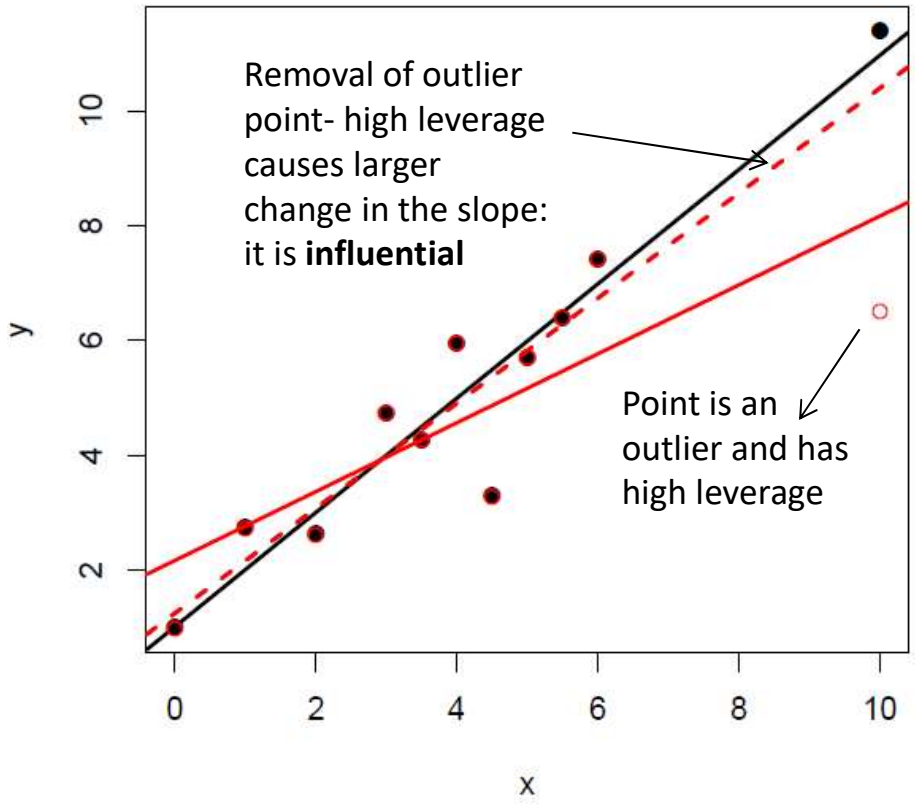
- Data points with high **leverage** are those with **extreme values** in the **explanatory variable (x)**.

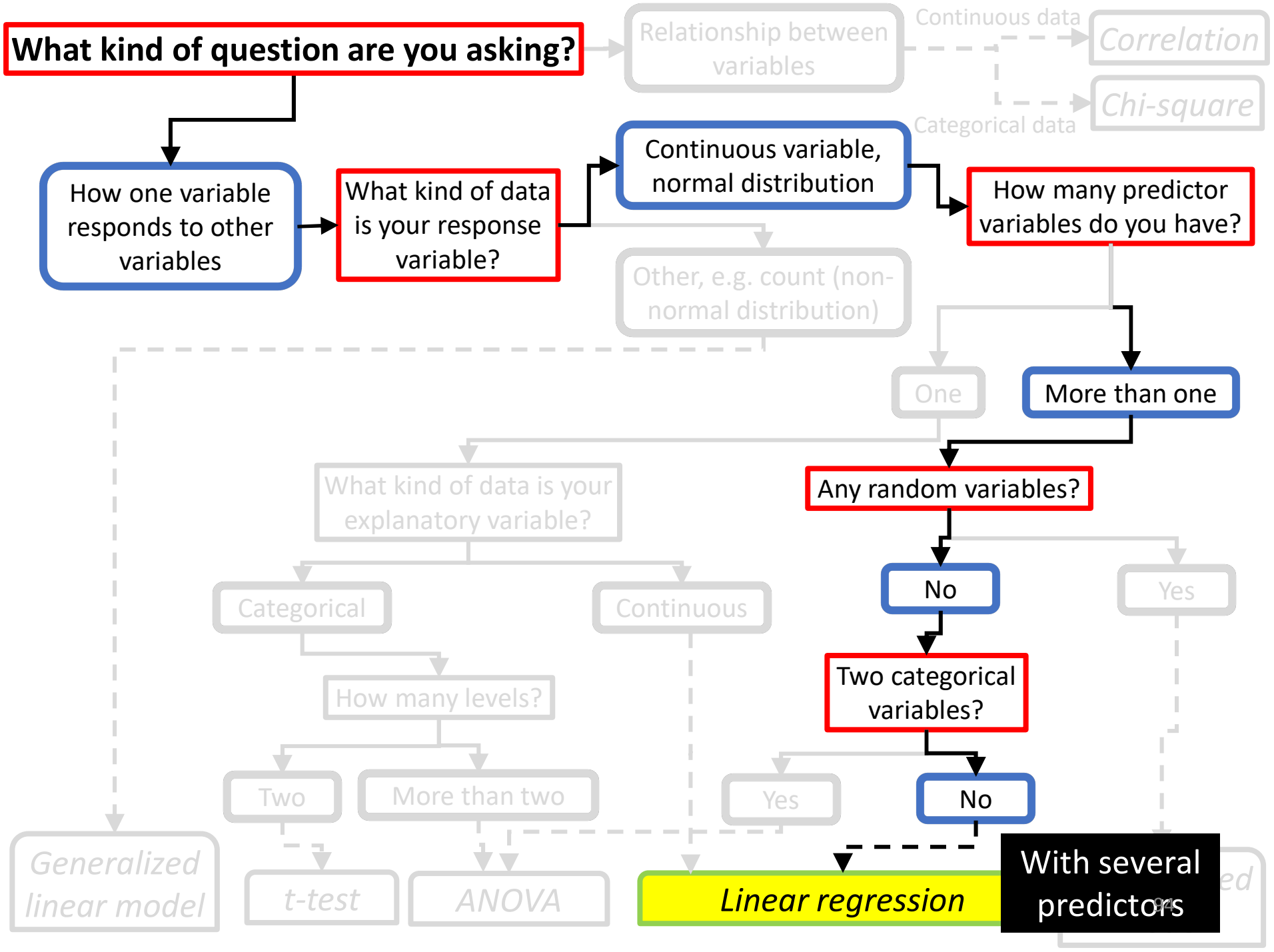




Leverage, outliers and influence

- When points have a high leverage AND are outliers for y, they are likely to be influential (having a larger effect on parameter estimate).




What kind of question are you asking?
Relationship between variables
Continuous data
Correlation
Chi-square
Categorical data
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
With several predictors

Linear model with two predictors

- The two explanatory variables may be both continuous, or one explanatory variable is categorical and the other is continuous (this is sometimes referred to as Analysis Of COVAriance, a.k.a. ANCOVA).
  - Example: does the relationship between leaf N content and leaf LMA differs between N-fixers and non-N-fixers?

- The two predictors may have **additive effects**:

$$Y_i = \alpha + \beta_1 * X_{1_i} + \beta_2 * X_{2_i} + \varepsilon_i$$

α : the **intercept** (value when X=0)

β_1: value by which leaf N content increases with LMA

β_2: the effect of being an N-fixer (X_{2_i}=1) compared to not being an N-fixer (X_{2_i}=0)

Linear model with two predictors

- Or, the two predictors may have **interacting effects**:

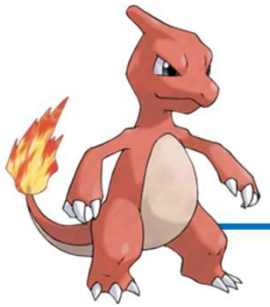
$$Y_i = \alpha + \left[\beta_1 + (\beta_3 * X_{2_i})\right] * X_{1_i} + \beta_2 * X_{2_i} + \varepsilon_i$$

α : the **intercept** (value when X=0)

β_1: value by which leaf N content increases with LMA

β_2: the effect of being an N-fixer (X_{2_i}=1) compared to not being an N-fixer (X_{2_i}=0)

β_3: the change in slope due to being an N-fixer (X_{2_i}=1) compared to not being an N-fixer (X_{2_i}=0)



Time for an exercise

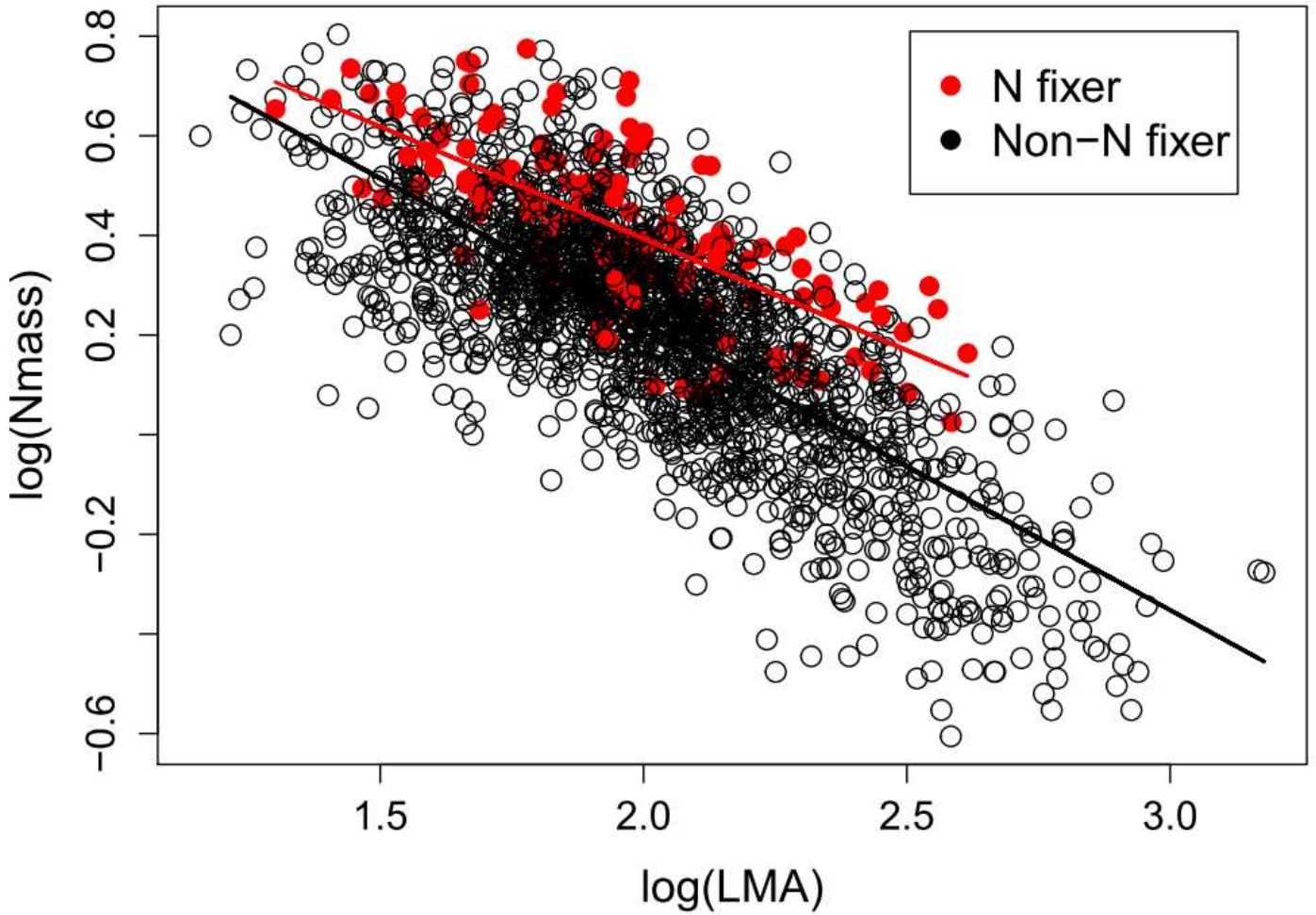
With our leaf traits dataset, we want to know the response of:

- Leaf N content to leaf LMA,
  and to N-fixing ability
- Leaf N content to leaf LMA,
  and if the response differs for N-fixers
  compared to non-N-fixers

Steps:

1) Load ‘leaftraits.txt’
2) Examine file structure
3) Run linear regression
4) Examine results
5) Check model assumptions

Plotting the interaction model



More practice with new data!

Load the ‘Forest_data.csv’ file. The data for each tree:

- Area: Sampling area
- Transect: Transect within area (~3-4 per area)
- Plot_Dist: Indiv. plot, noted as dist. along transect
- Soil: Plot soil type (fine, talus, both)
- Slope: Plot slope angle
- Profile: Plot profile (concave, convex, planar)
- TreeNum: Number of tree on the plot
- Species: Tree species
- Status: Live, dead
- DBH: Tree diameter at breast height
- Height: Tree height
- Age: Tree age (taken for largest tree on plot only)
- Basal.Area_km2: basal area of tree (as per DBH)
- cosAspect: Aspect of plot (transformed)

More practice with new data!

Answer the following questions:

1) Are tree height and DBH related?
2) Do the two tree species occur similarly on different plot profiles?
3) Does one species grow taller than another?
4) Does tree height vary with plot slope angle?
5) Does tree height vary with plot slope angle, and tree species?
6) Does tree height vary with plot slope angle, and does this vary with tree species?
7) Does tree DBH vary with plot profile?
8) Ask your own!

Tips:

1) Examine variables you will need
2) Identify appropriate analysis
3) Perform post-hoc analysis, if needed
4) Perform model validation, if needed

Acknowledgements

People:

Emily Haeuser
Wayne Dawson
Fränzi Körner
... and all the others!

International Max Planck Research School for Organismal Biology

Ecology Konstanz

Universität Konstanz

LGFG GEFÖRDERT VOM Bundesministerium für Bildung und Forschung

The Company of Biologists

TReND

Ikala STEM

Supported by

Development

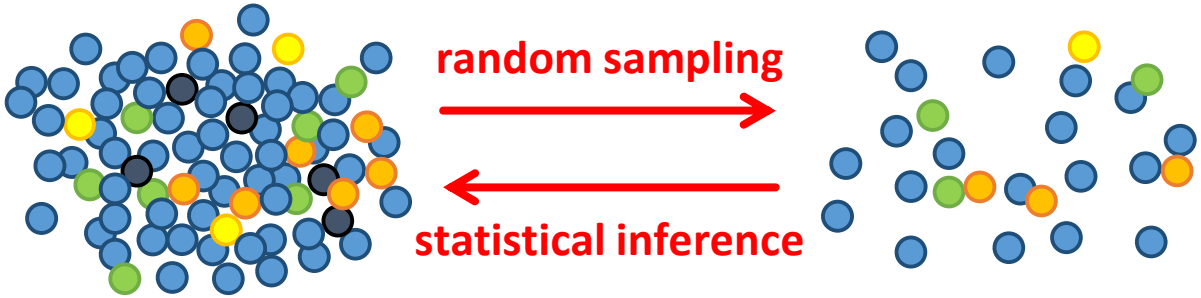
Journal of Cell Science

Journal of Experimental Biology

Disease Models & Mechanisms

Biology Open

Reminder: summary statistics



Population mean μ ← is an estimate of — Sample mean $\bar{x} = \frac{\sum x_i}{N}$

Population variance σ^2 ← is an estimate of — Sample variance $s^2 = \frac{1}{N-1}\sum_{i=0}^{N}(x_i - \bar{x})^2$

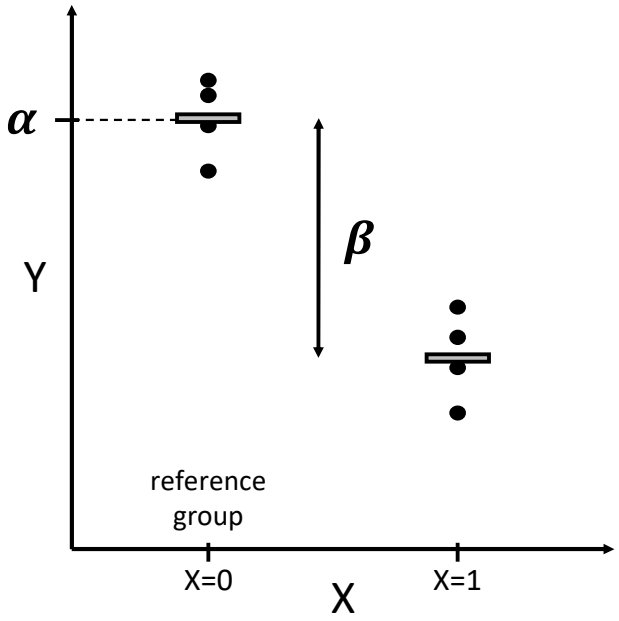
Population standard deviation σ ← is an estimate of — Sample standard deviation $s = \sqrt{s^2}$

Sample standard error of the mean $SE = \frac{s}{\sqrt{N}}$

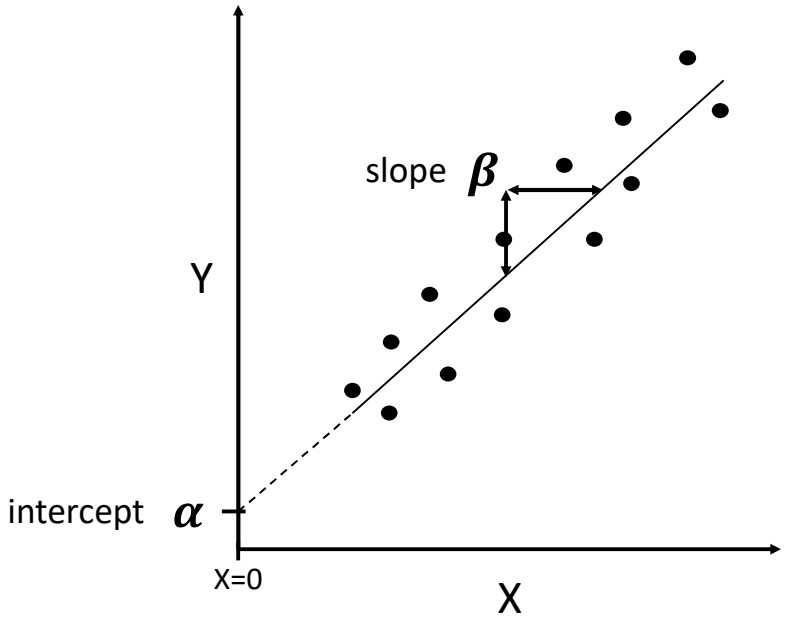
ANOVA and linear regression

$$Y_i = \alpha + \beta * X_i + \varepsilon_i$$

X_i categorical: ANOVA



X_i continuous: linear regression



*When X=0, Y=α+β*0= α (mean of the reference group).*
*When X=1, Y= α+β*1=α+β (mean of the second group, i.e. mean of the reference group + difference between group means).*