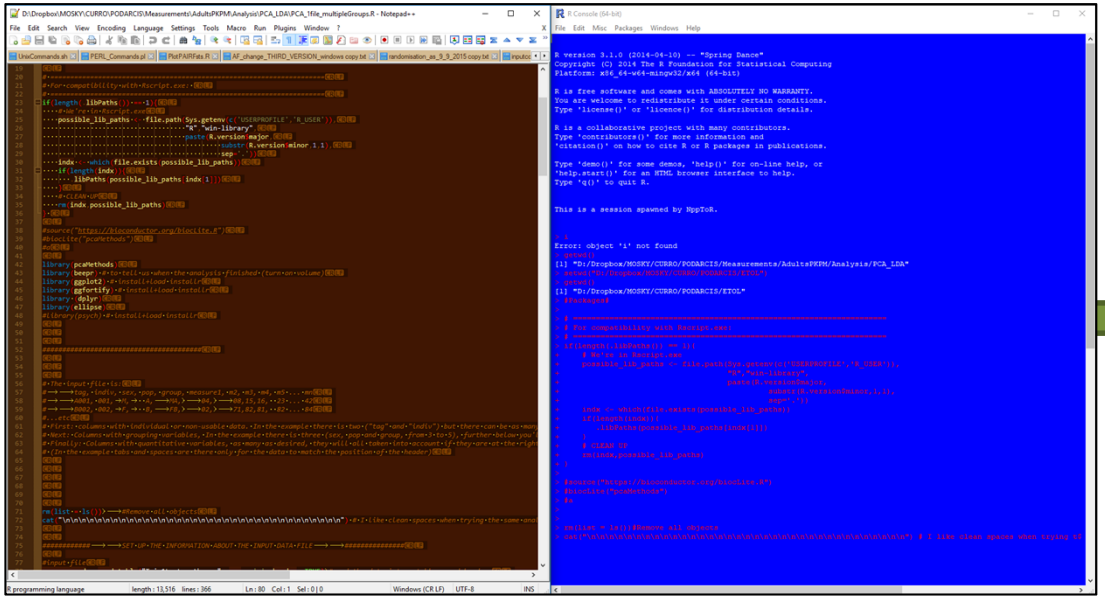
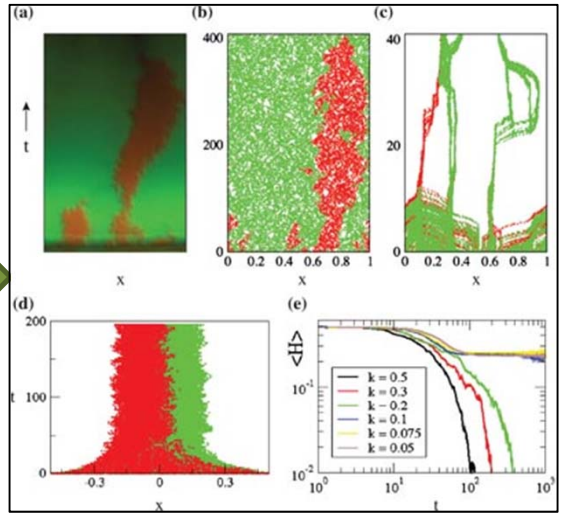
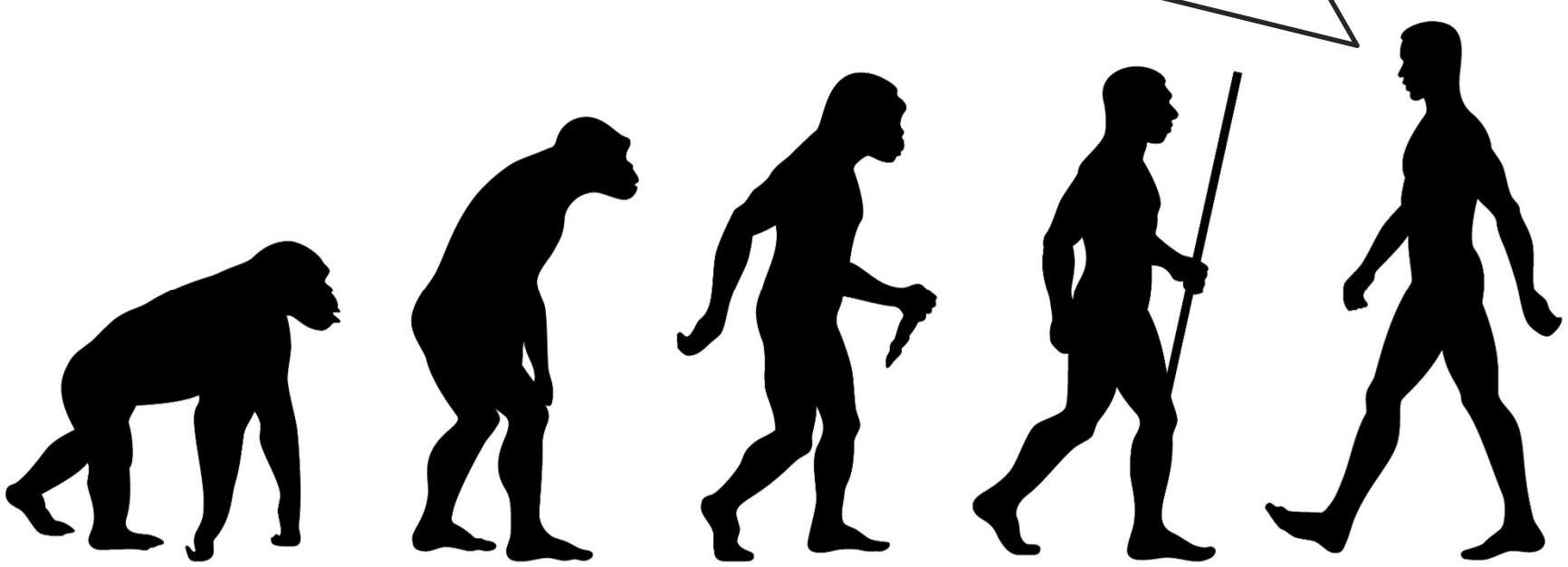
Use of R environment in **Evolutionary Ecology**





OK guys, let's go

Back to the basics



From Basic R to Basic statistics

Is pointless to learn new ways to analyse your data if you don't understand what you are doing

0) What do we want to know?
1) Which type of data do we have?
2) Descriptive analysis/plot
3) Limitations of the analysis -> Choose

***DISCLAIMER**

Basic statistics

Is pointless to learn new ways to analyse your data if you don't understand what you are doing

0) What do we want to know?
1) Which type of data do we have?
2) Descriptive analysis/plot
3) Limitations of the analysis -> Choose

Don't just perform analysis frenetically

Basic statistics

0) What do we want to know?

In most of the cases:

Are there significant differences between groups?

We may also be interested in:

- Which are those groups?
- Which variables define groups?
- Is there any variable bigger/smaller in any group?
- How the variables affect to each other?

Basic statistics

Is pointless to learn new ways to analyse your data if you don't understand what you are doing

0) What do we want to know?

1) Which type of data do we have?

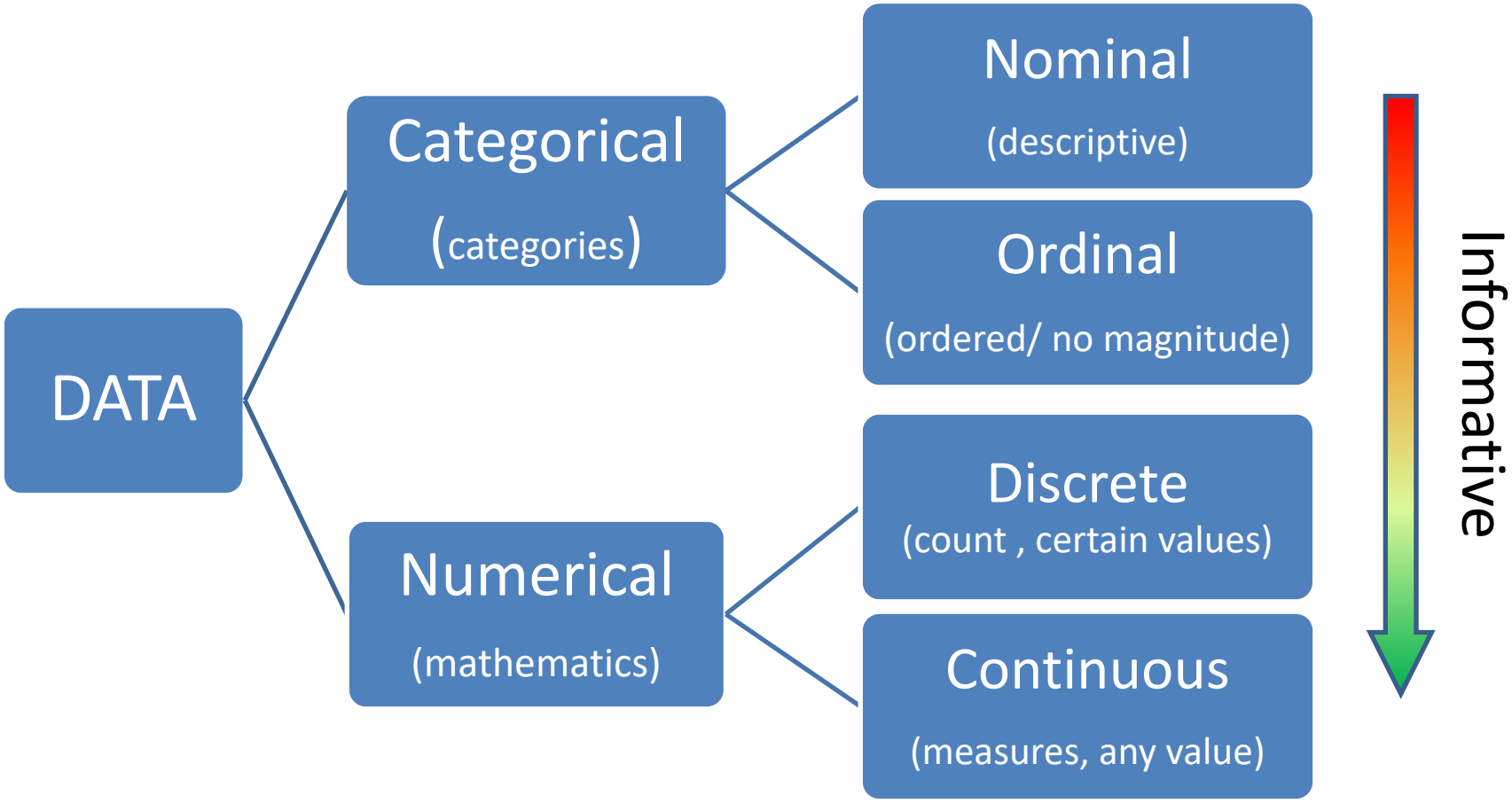
2) Descriptive analysis/plot

3) Limitations of the analysis -> Choose

We should have thought about point **0** (and 1) while designing the experiment

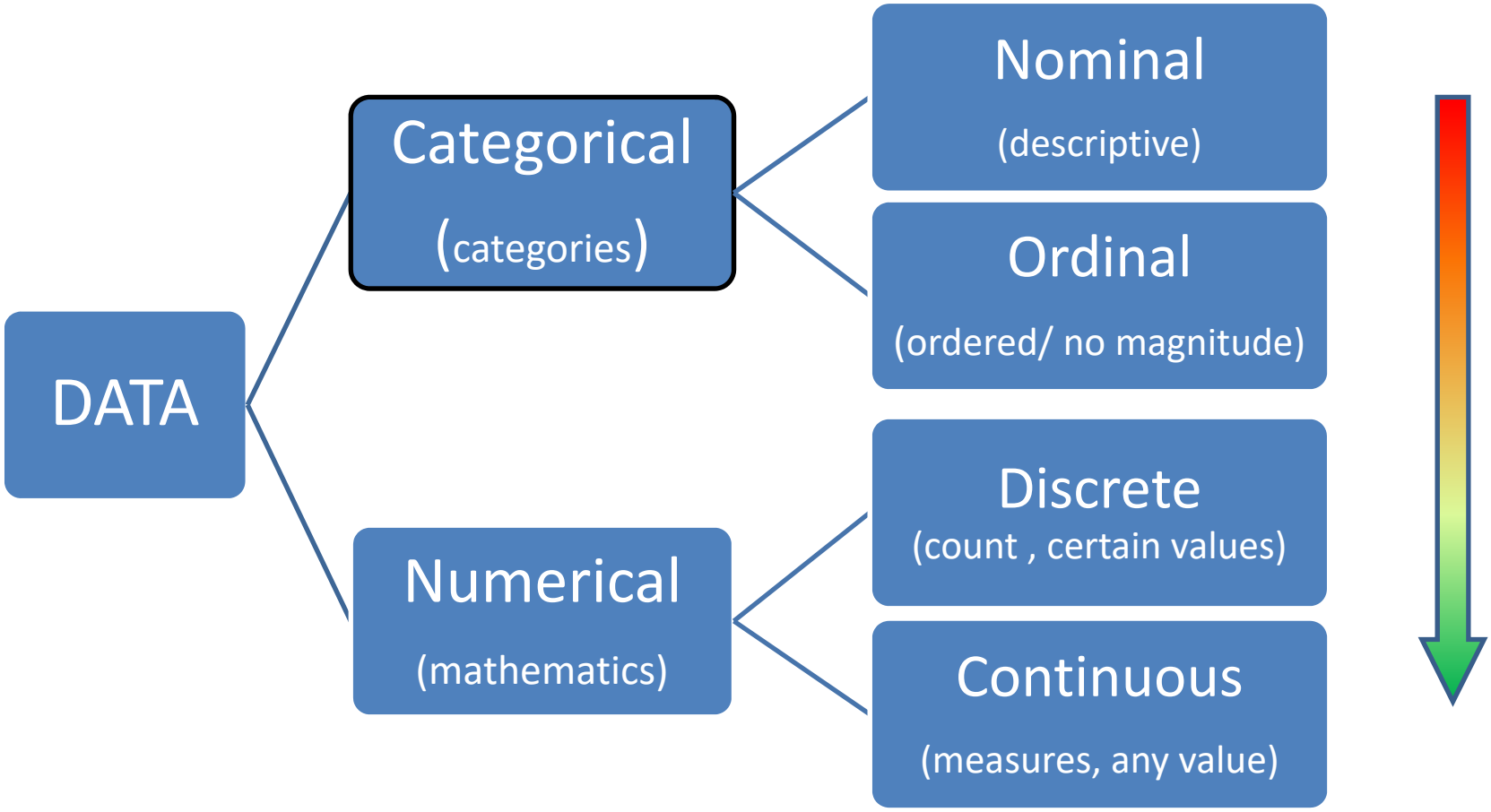
Basic statistics

1) Which type of data do we have?

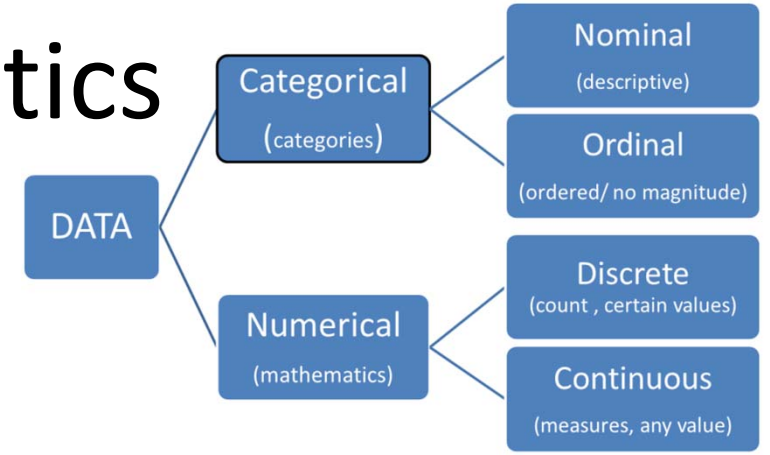


Basic statistics

1) Which type of data do we have?



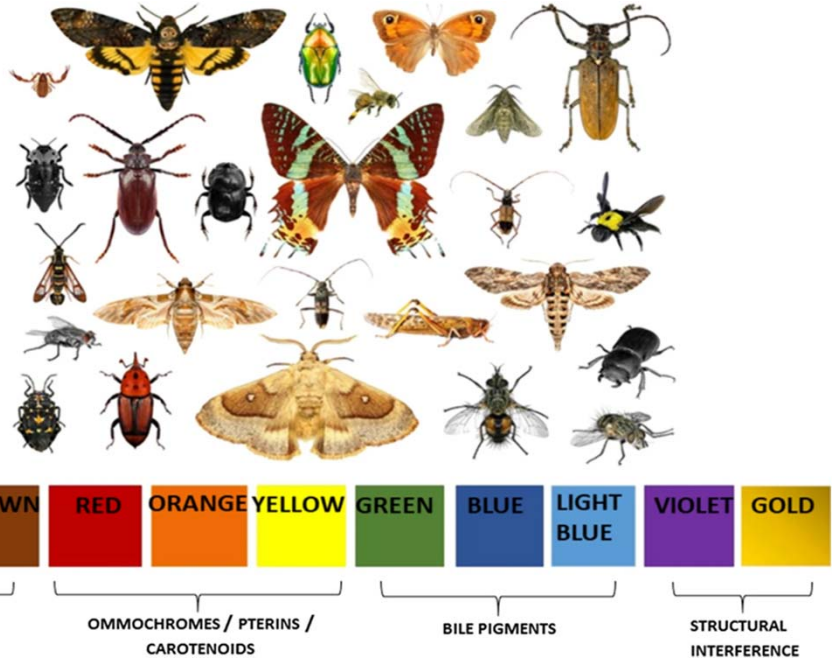
Basic statistics



1) Type of data

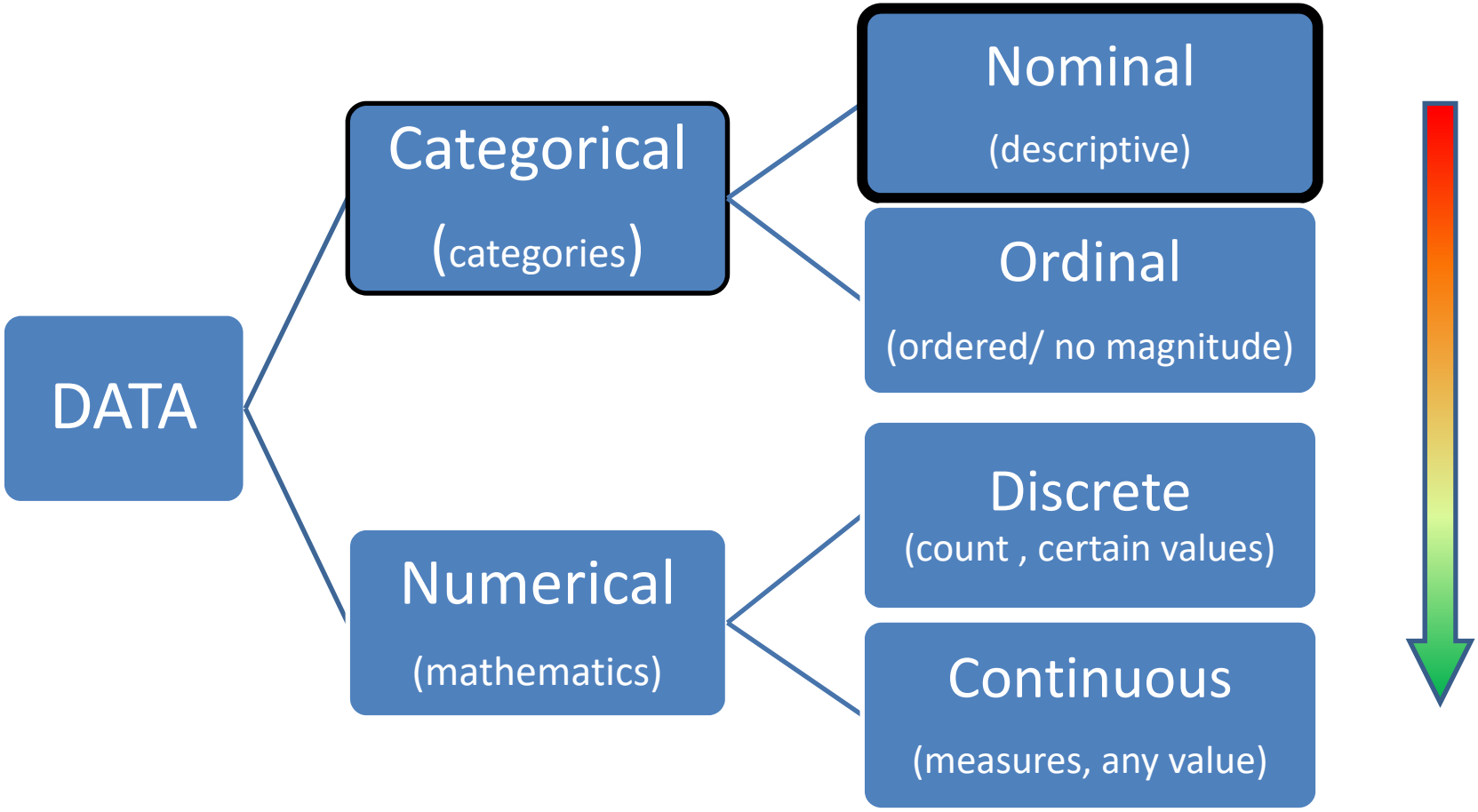
CATEGORICAL (chr/Factor)

- Anything that is not a numerical measured value
- Characters/categories
- Finite possible values
- Usually qualitative

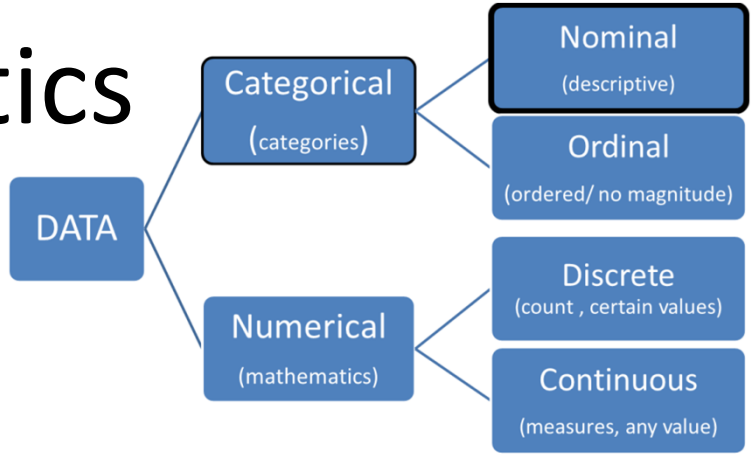


Basic statistics

1) Which type of data do we have?



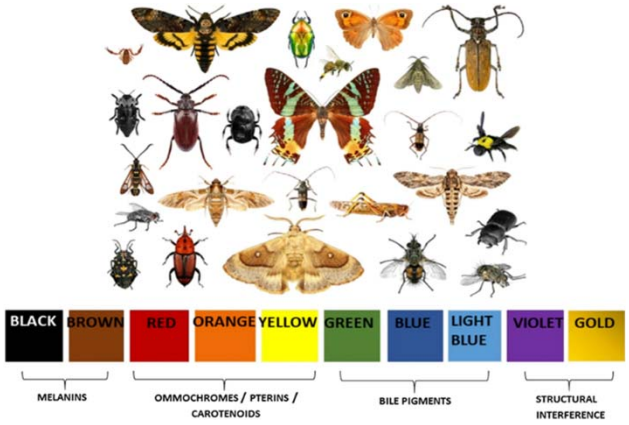
Basic statistics



1) Type of data

NOMINAL (Categorical)

- Descriptive, qualitative (no value)
- Often used as "grouping" variables



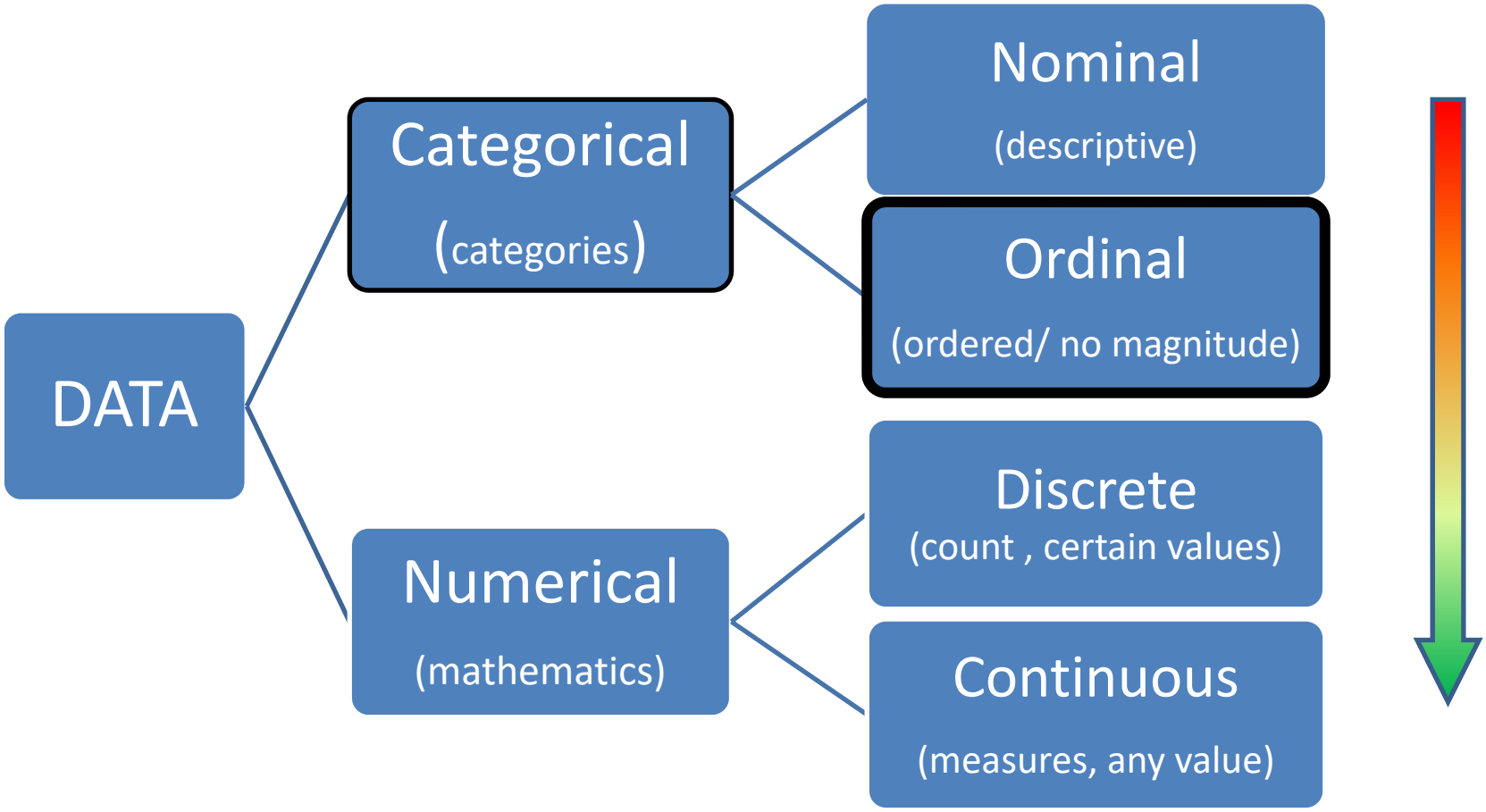
Colour, sex, population,
species, ecotype, host-plant...

- Can be dichotomous (binary)

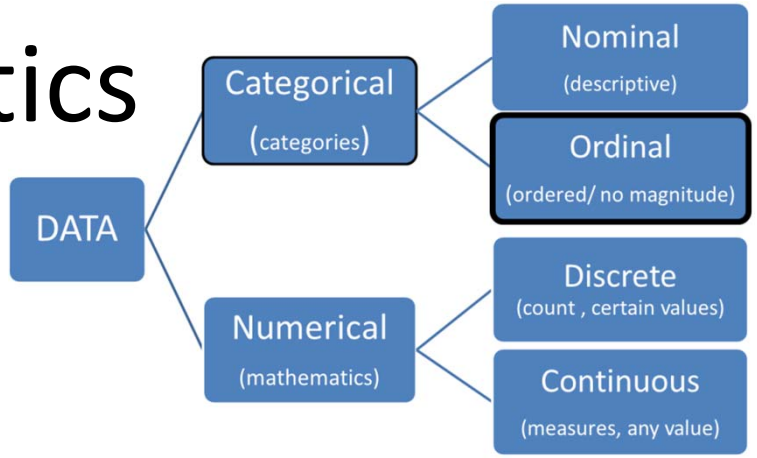
yes/no; TRUE/FALSE; 0/1; dead/alive;

Basic statistics

1) Which type of data do we have?



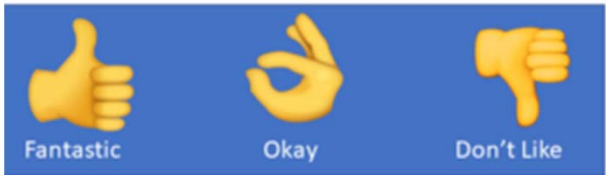
Basic statistics



1) Type of data

ORDINAL (Categorical)

- They are sortable (~value)
- Not exact numeric equivalence



juvenile, subadult, adult, old

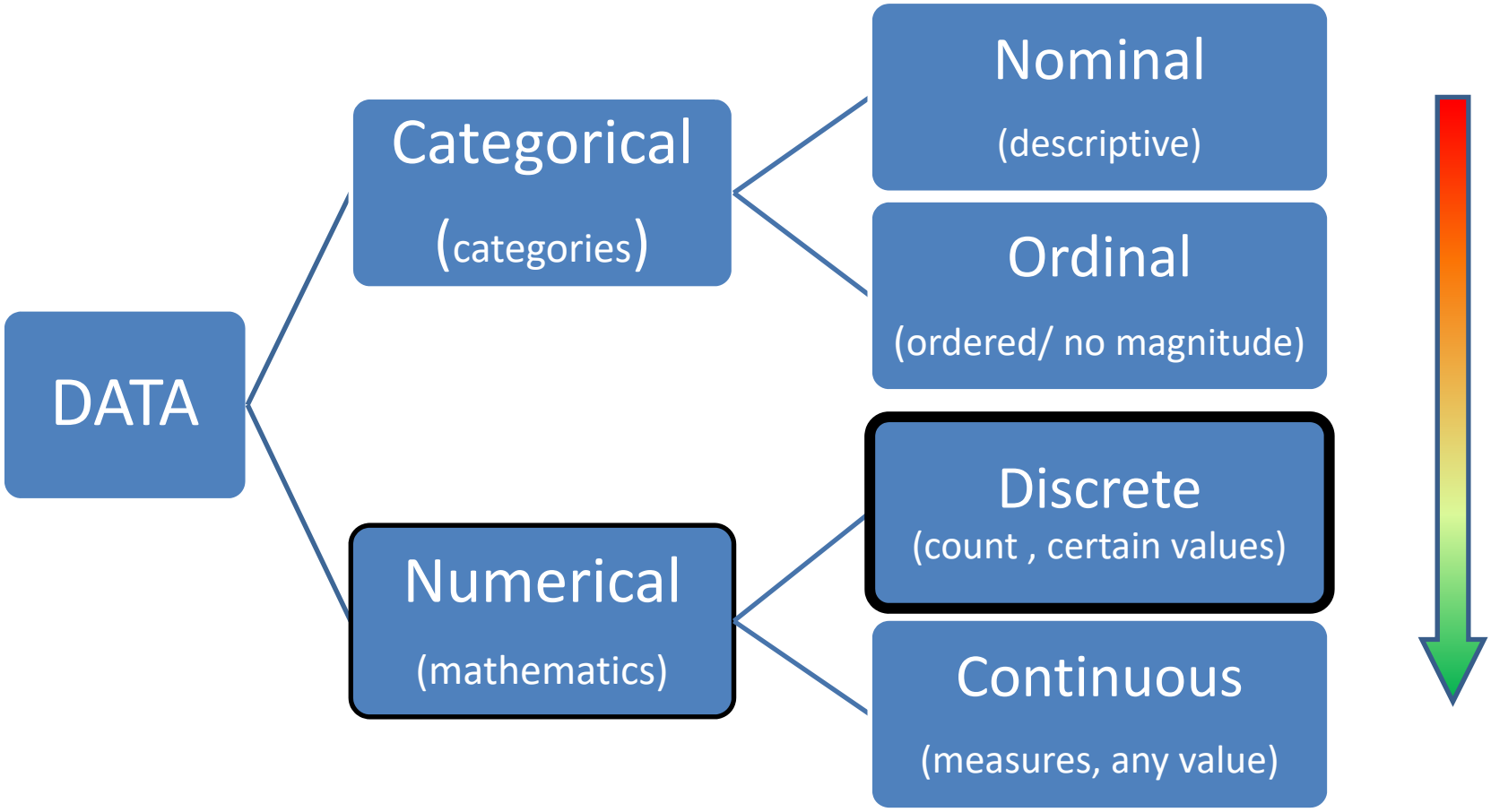
slow, medium, fast

bad, good, excelent

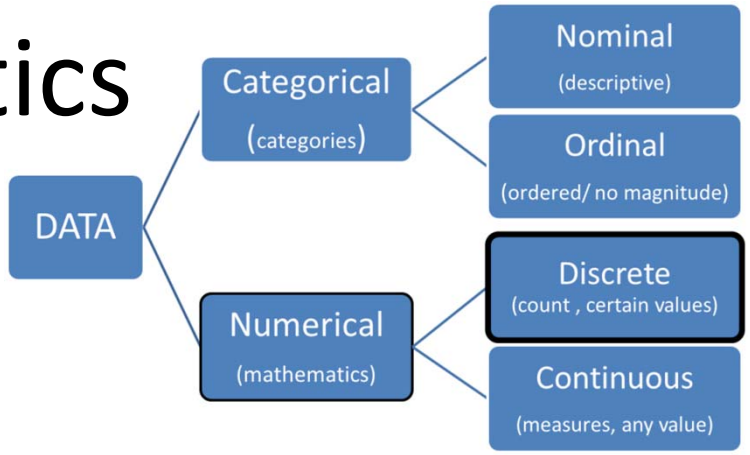
light, medium, dark

Basic statistics

1) Which type of data do we have?



Basic statistics



1) Type of data

DISCRETE (Numerical)

- Integers (int)
- Not all values *in-between* exist (can't be subdivided)

Usually:

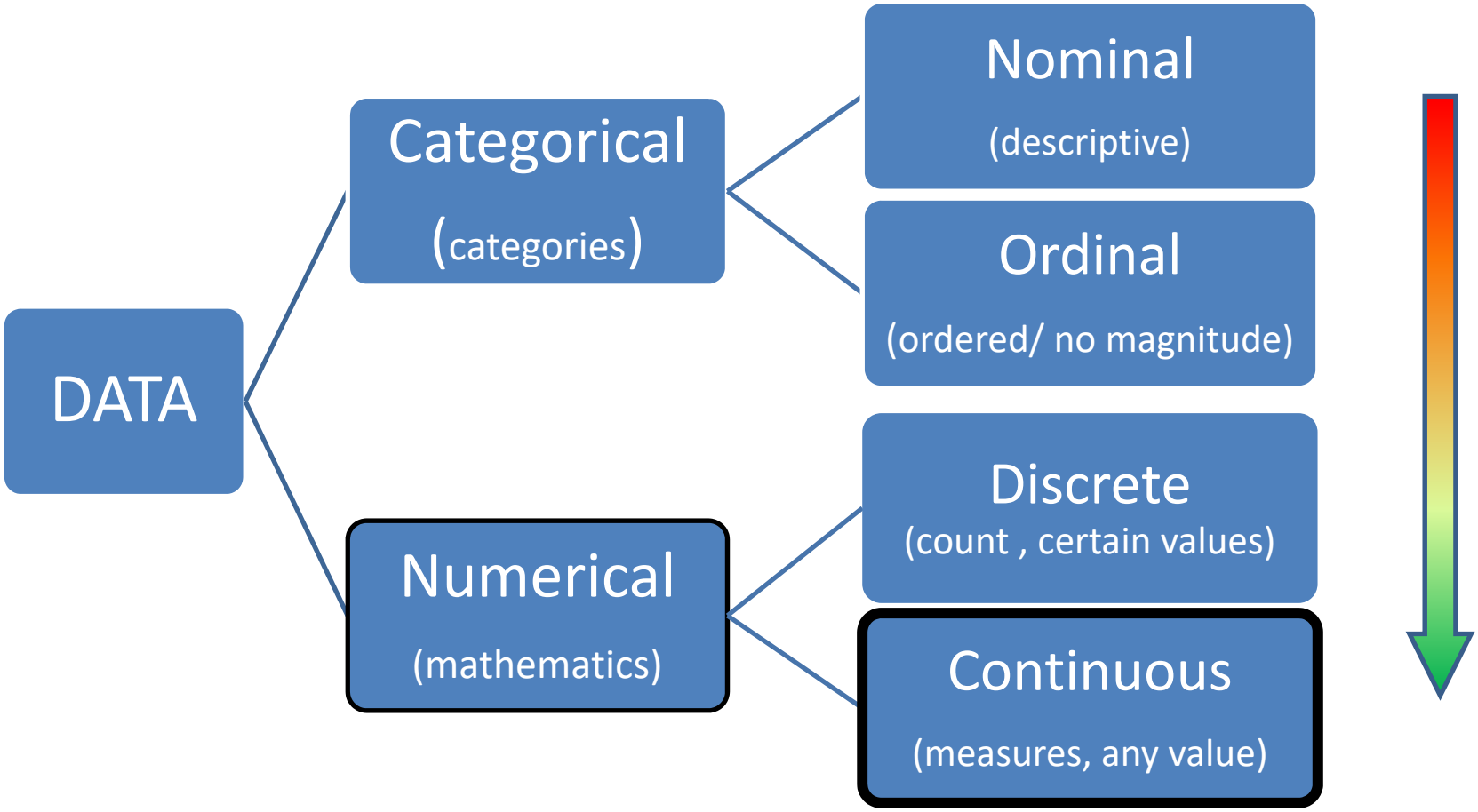
Number of something/counts

Number of beans, headbobs,

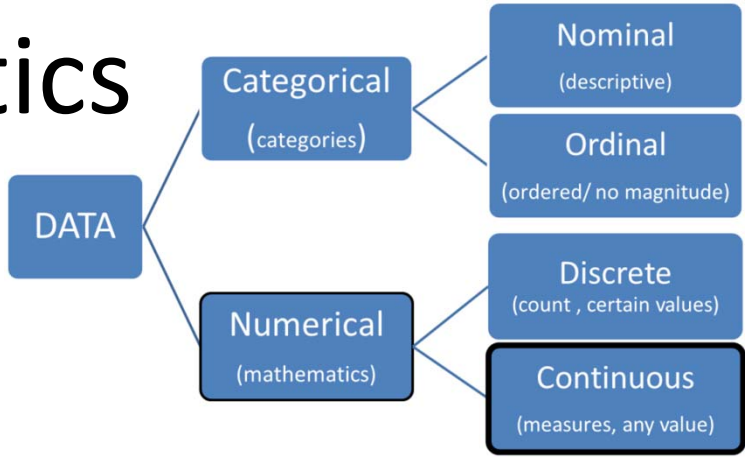
students per class, eggs laid, petals

Basic statistics

1) Which type of data do we have?



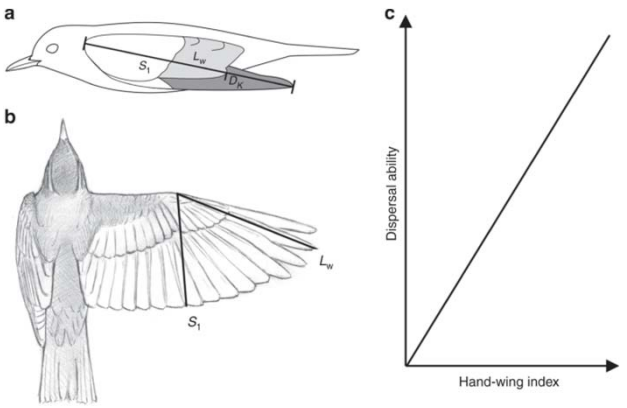
Basic statistics

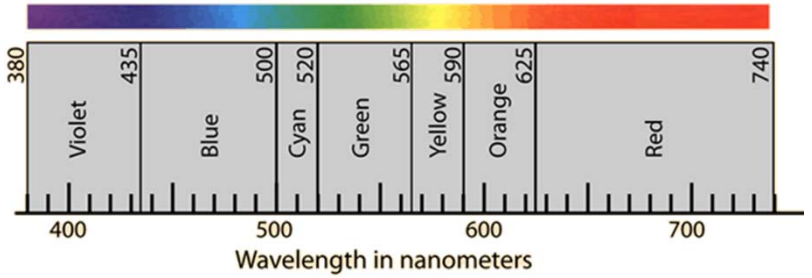


1) Type of data

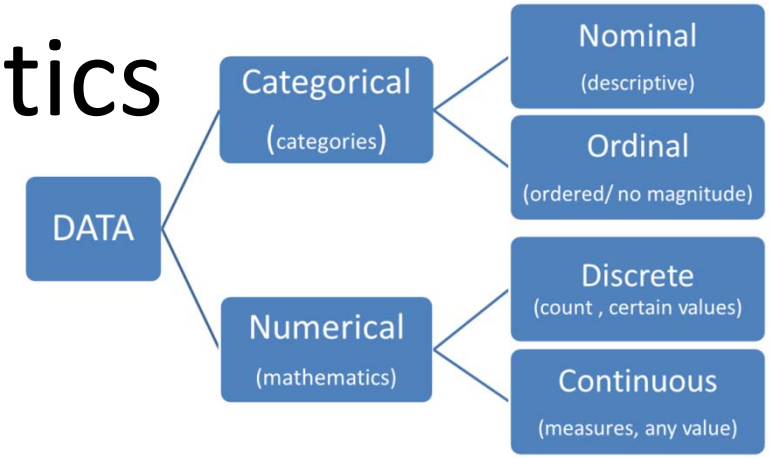
CONTINUOUS (Numerical)

- Real numerical values (num) ($\in \mathbb{R}$)
- Usually direct measurements
- More information!
  More powerful!
- More options!
  More complications!!
  (distribution)





Basic statistics



1) Types of data

- Number of spots on ladybugs
- Radiation levels on Gallifrey's atmosphere
- Sandworms sightings per year on Arrakis deserts
- Age group (babies, teenagers, adults, old)
- Education level
- Nationality
- Time spent exploring
- Passed/Did not pass the final test

Basic statistics

Always check your data before analysing it

0) What do we want to know?
1) Which type of data do we have?
2) Descriptive analysis/plot
3) Limitations of the analysis -> Choose

Basic statistics

2) Descriptive analysis / plots

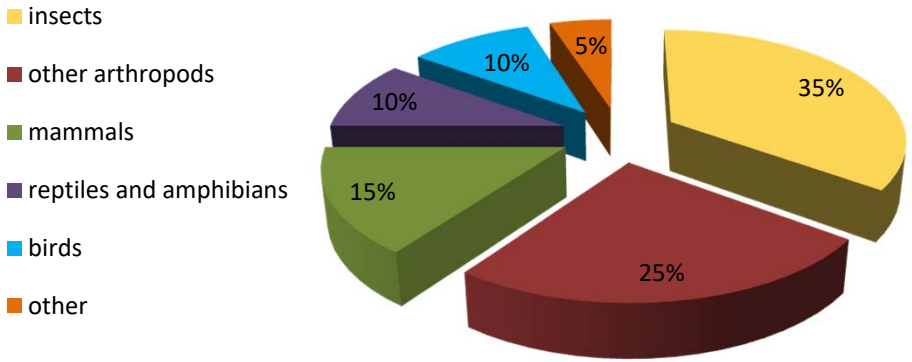
- Over-view of the data
- Descriptive statistics
  Average, SD, Median and quartiles
- Skewness and Outliers
- A priory differences and groups

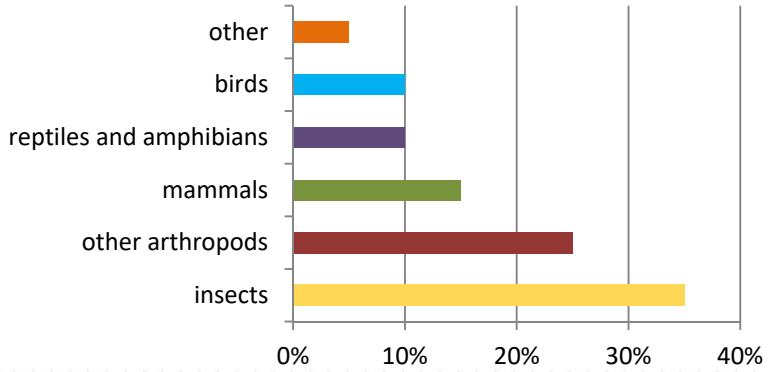
Basic statistics

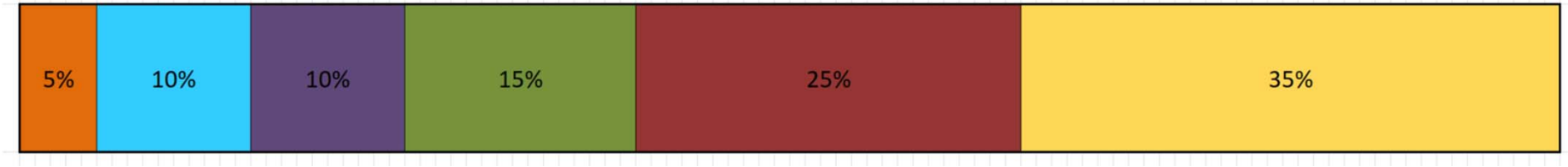
2) Descriptive plots

For categories

insects	35%
other arthropods	25%
mammals	15%
reptiles and amphibians	10%
birds	10%
other	5%





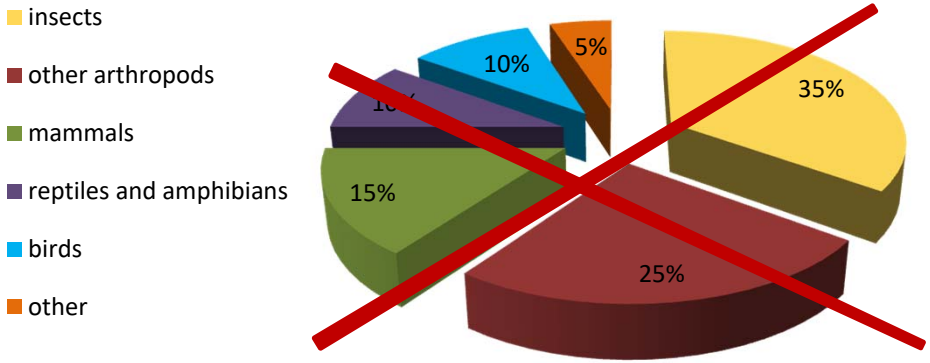


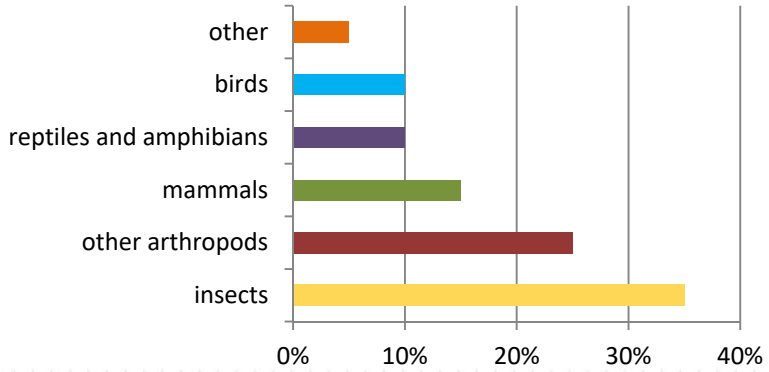
Basic statistics

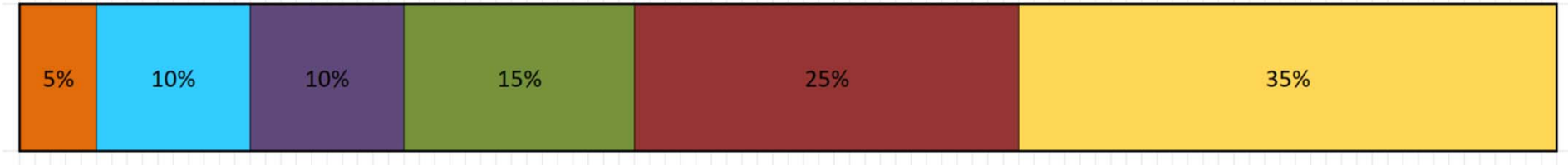
2) Descriptive plots

For categories

insects	35%
other arthropods	25%
mammals	15%
reptiles and amphibians	10%
birds	10%
other	5%



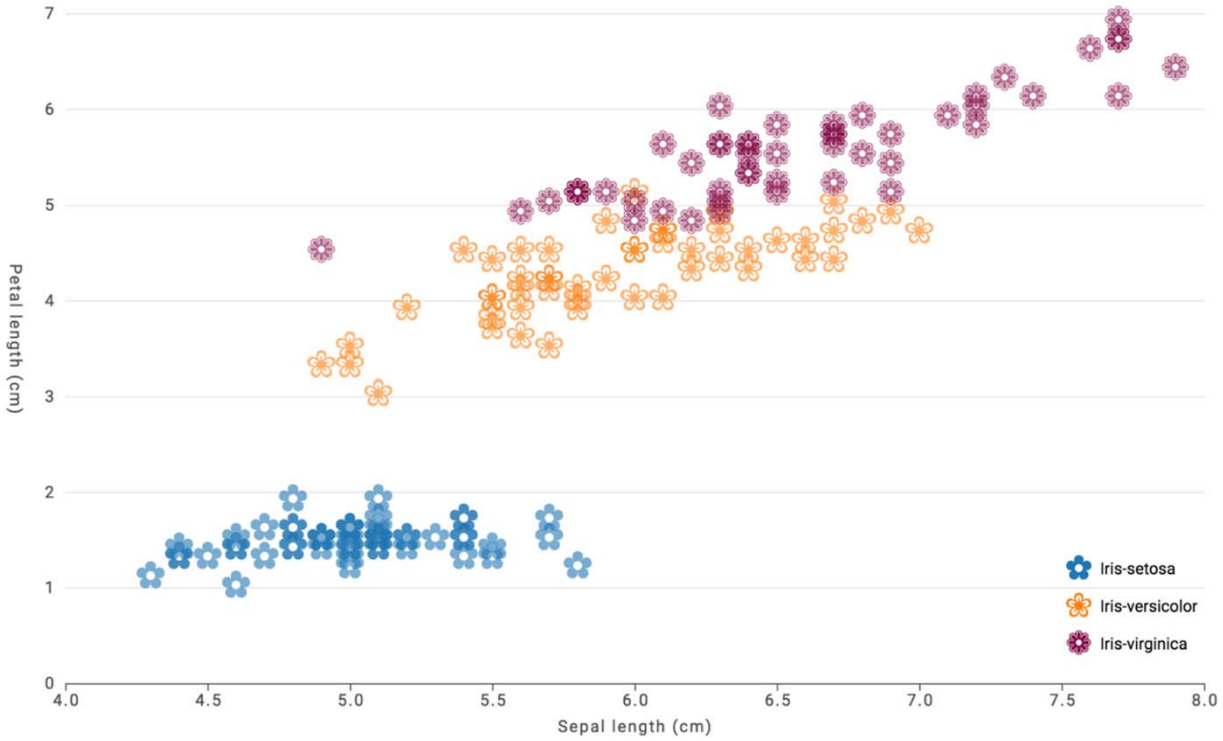




Basic statistics

2) Descriptive plots

Scatter plot



Iris Versicolor

Iris Setosa

Iris Virginica

Basic statistics

2) Descriptive plots

Box plot

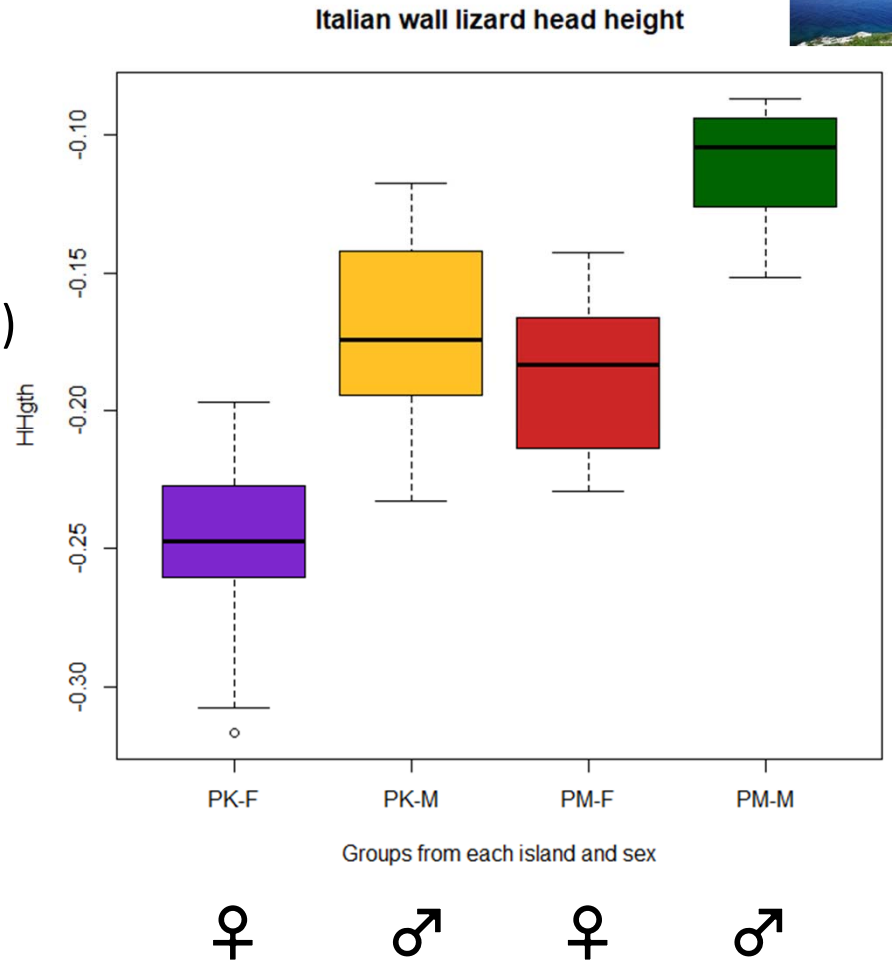
Median (50%)

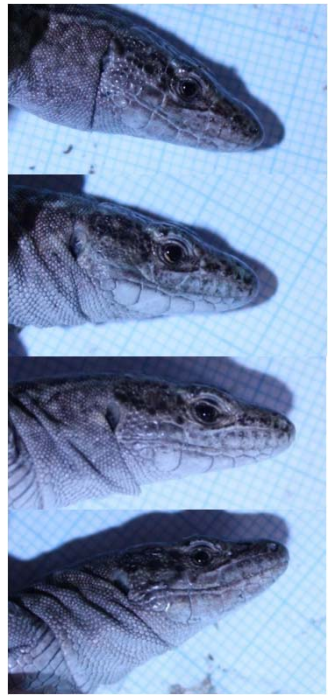
Quartiles (25-75%)

Minimum

Maximum

Outliers

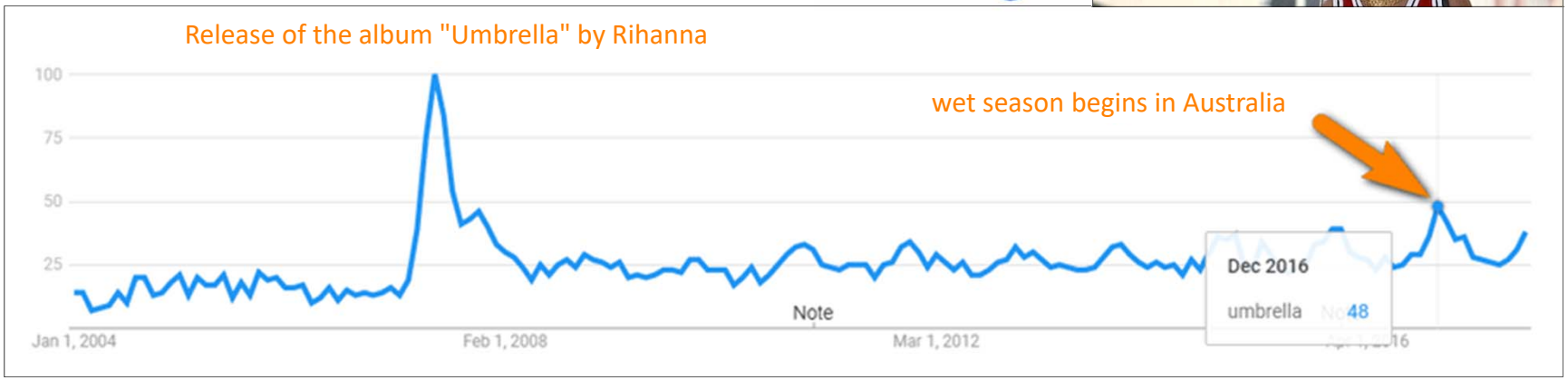


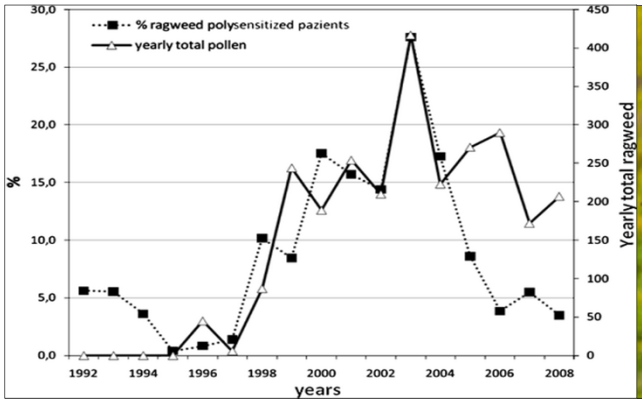


Basic statistics

2) Descriptive plots

Trend





Basic statistics

Is pointless to learn new ways to analyse your data if you don't understand what you are doing

0) What do we want to know?
1) Which type of data do we have?
2) Descriptive analysis/plot
3) Limitations of the analysis -> Choose

Basic statistics

3) Limitations and pre-requisites of Analysis

MOST ANALYSIS NEED:

- Sample size / independence of datasets
- Equality of variances
- Degrees of freedom

Basic statistics

3) Limitations and pre-requisites of Analysis

- **Sample size / independence of datasets**

  So you have enough samples?

  Are your experiments replicated?

  Are the replicates independent?

  pseudo replication

  multiple samples of the same subject/area

  Too late to change the experiment!

Basic statistics

3) Limitations and pre-requisites of Analysis

MOST ANALYSIS NEED:

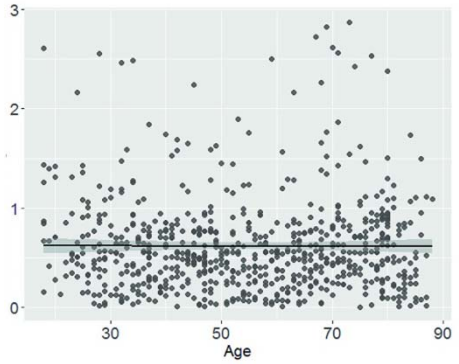
- Sample size / independence of datasets
- Equality of variances
- Degrees of freedom

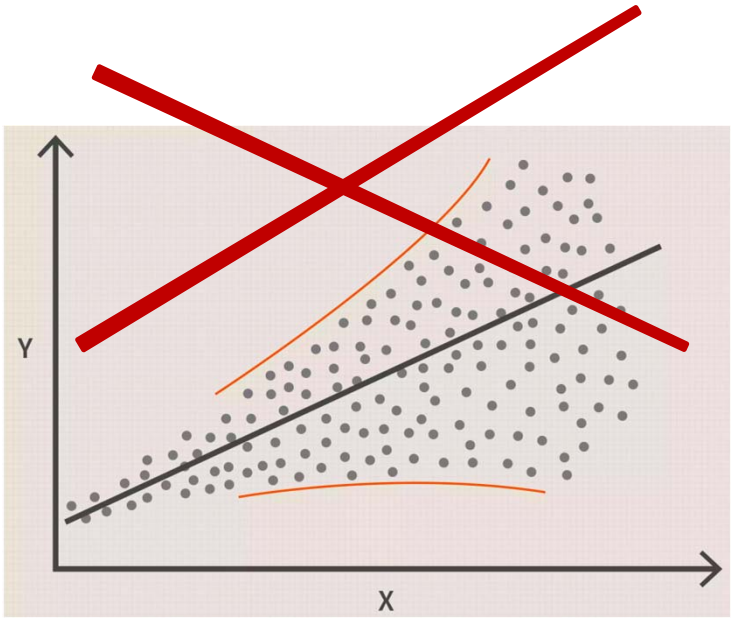
Basic statistics

3) Limitations and pre-requisites of Analysis

- **Equal variances**

  Assumption of homoscedasticity:
  Different samples have similar variance (error)
  The error "scatter" is independent from the population or have value.





Basic statistics

3) Limitations and pre-requisites of Analysis

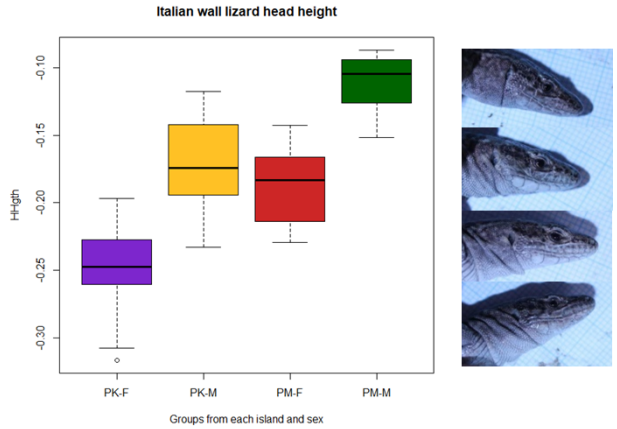
MOST ANALYSIS NEED:

- Sample size / independence of datasets
- Equality of variances
- Degrees of freedom

Basic statistics

3) Limitations and pre-requisites of Analysis

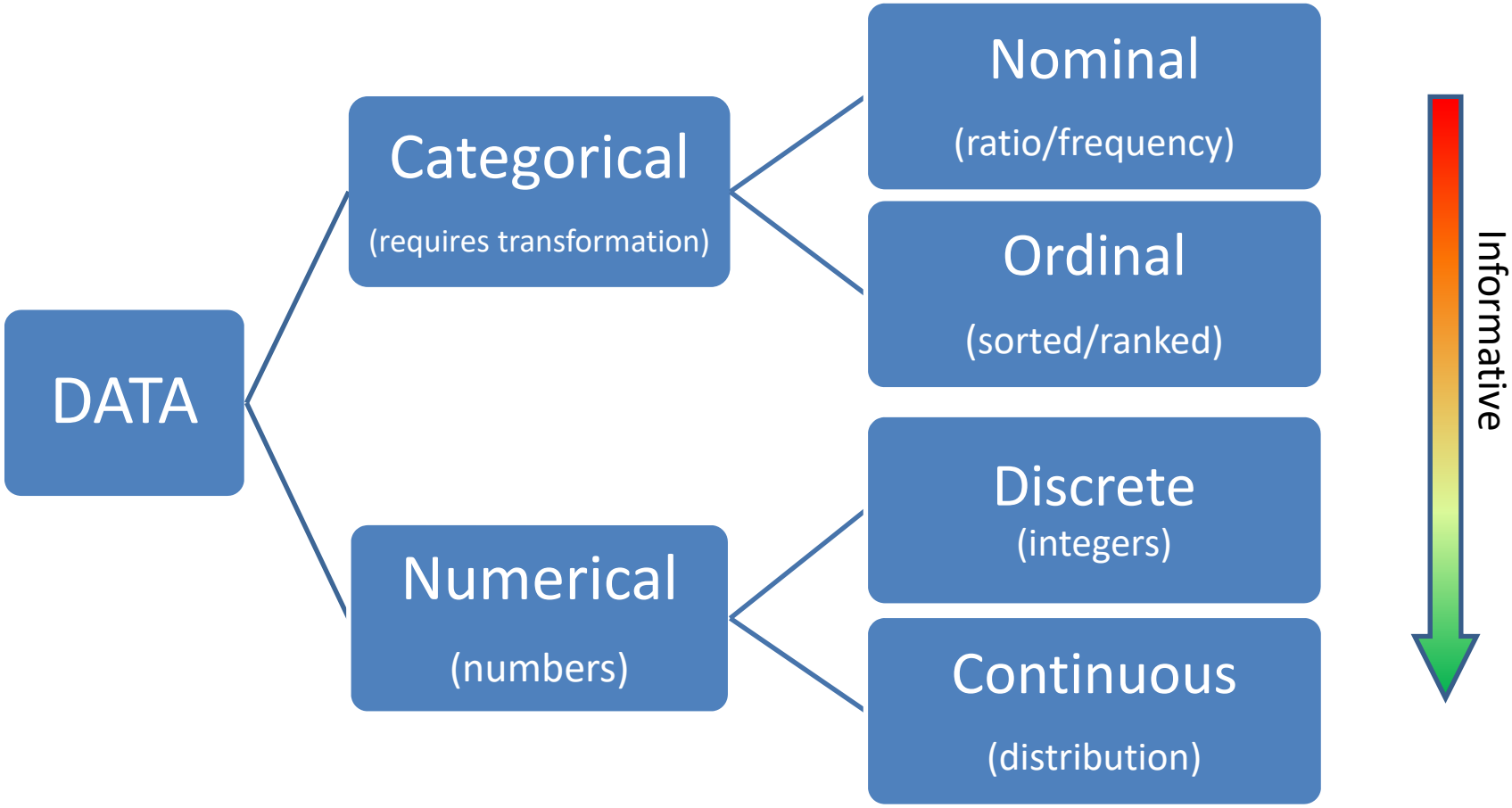
- **Degrees of freedom?**
  Number of clusters of information in which the estimated values are free to vary
  - Hard to understand
  - Easy to calculate -> levels of grouping - 1



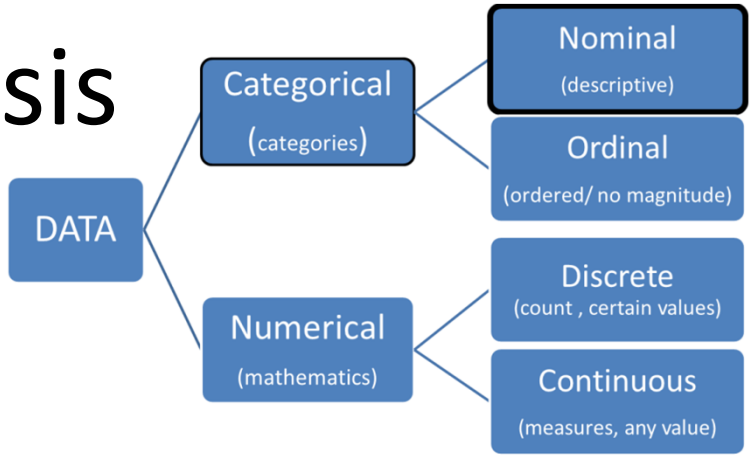
4 – 1 = 3 df

Basic statistics

3) Choose analysis



Basic statistics: Analysis

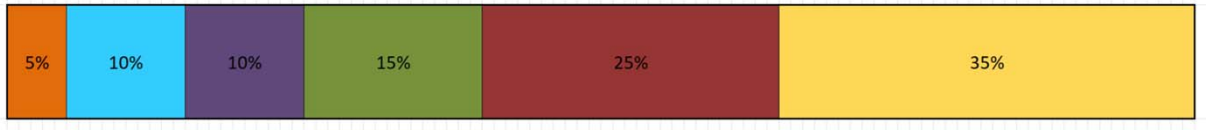


NOMINAL (Categorical)

species, colour, hostplant, population...

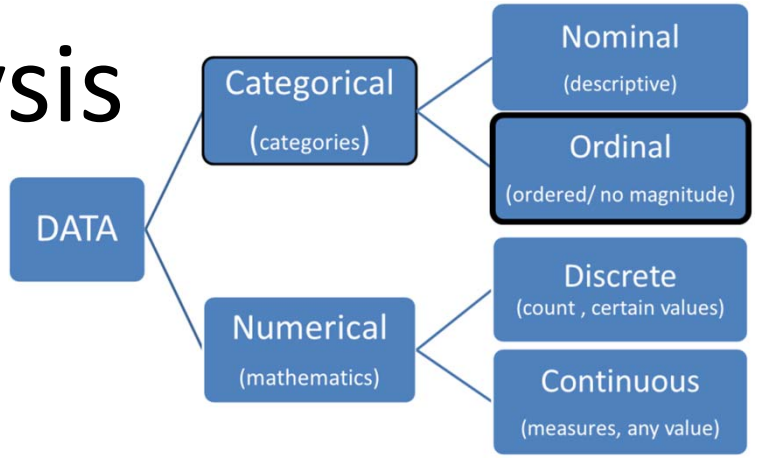
We need to calculate the ratio / frequency

- Chi-square χ^2

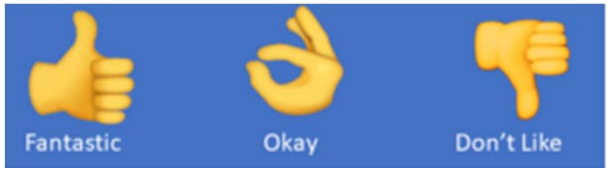


- binomial test, fisher exact test (binary data)

Basic statistics: Analysis



ORDINAL (Categorical)

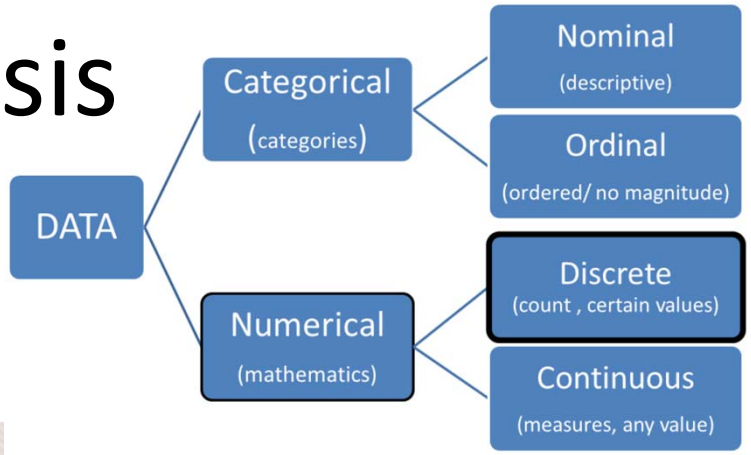


fast, medium, slow // light, medium, dark

Need to sort the observations and/or rank them

- Wilcoxon test
- Median test
- Kruskal-Wallis test

Basic statistics: Analysis



DISCRETE (Numerical)

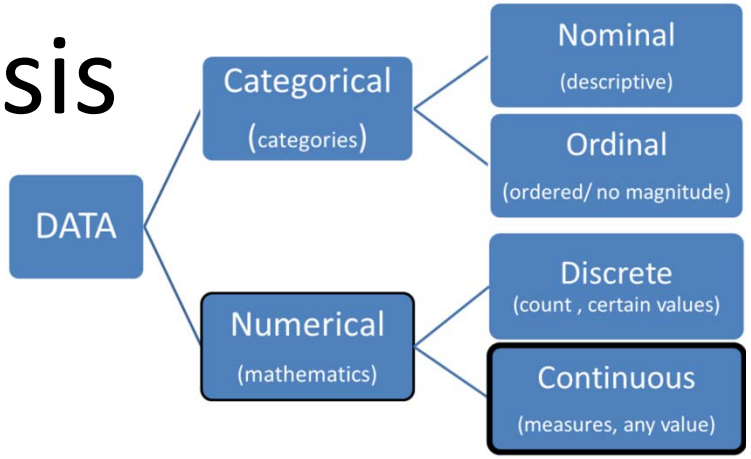
number of spots // eggs laid

Same as Ordinal (better resolution)

- Wilcoxon test
- Median test
- Kruskal-Wallis test

*** some test for continuous numerical data can analyse discrete data**

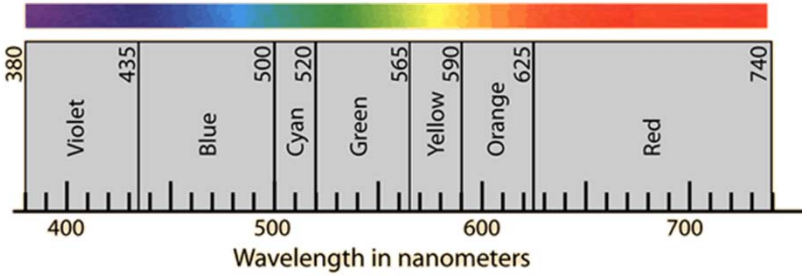
Basic statistics: Analysis



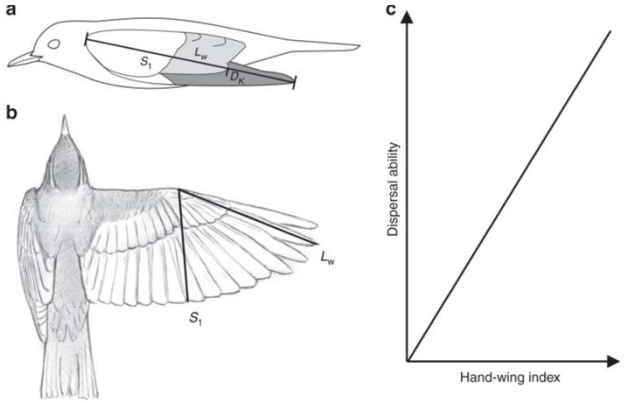
CONTINUOUS (Numerical)

This is most informative data

Predictions vs ideal distribution



Most tests establish if there are differences between groups by comparing the variance/average values of your data with the predicted values from an ideal distribution



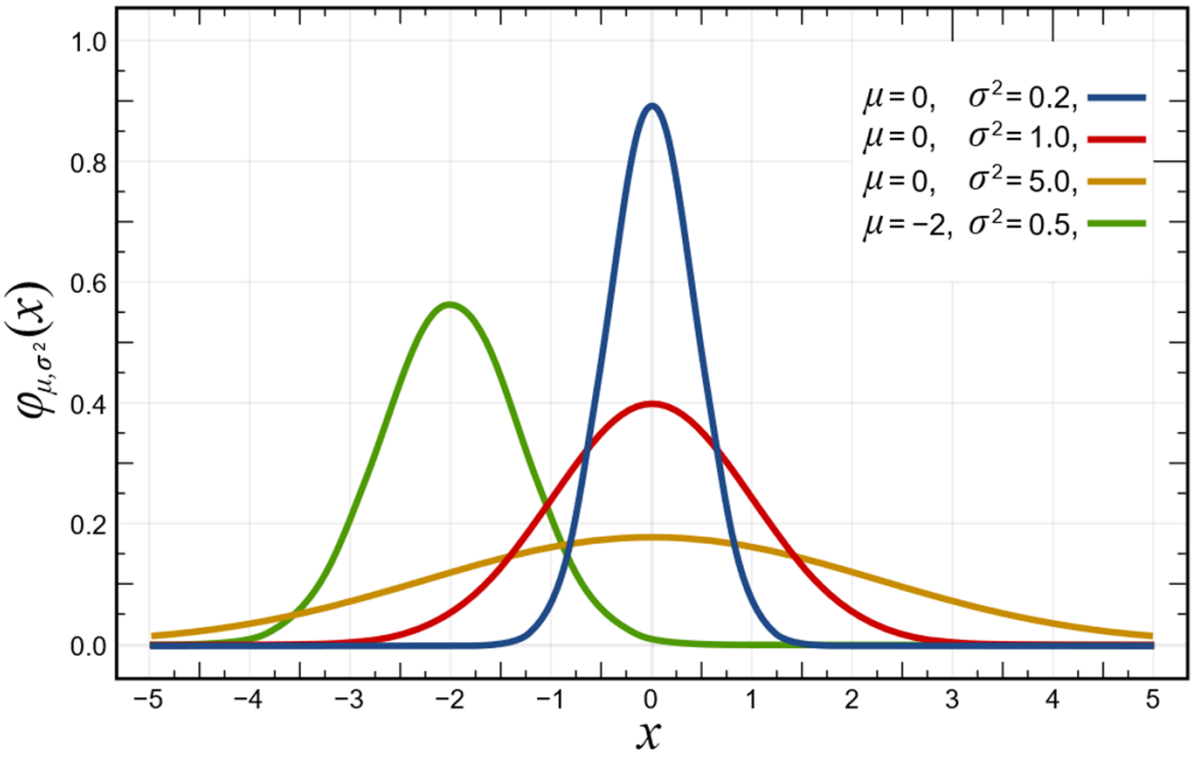
Your data needs to fit the distribution

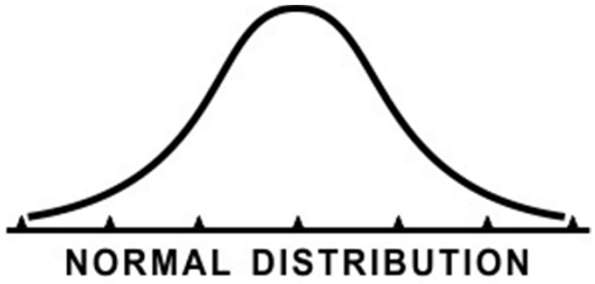
Basic statistics

3) Limitations and pre-requisites of Analysis

<u>Distribution</u>

Normality?





Basic statistics

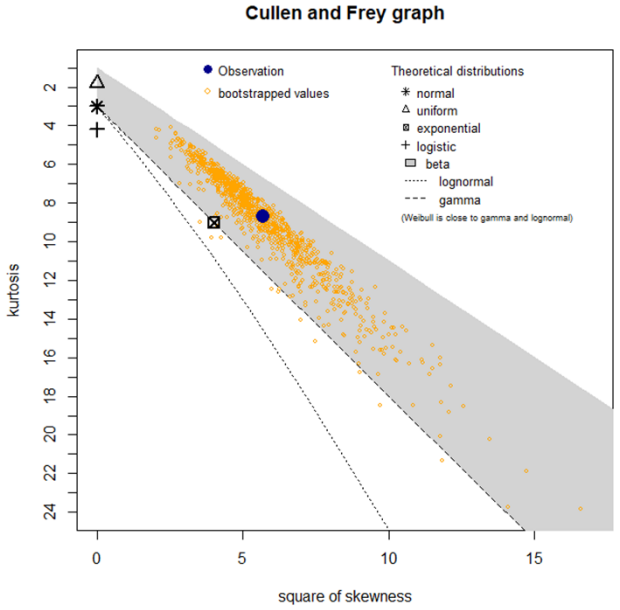
3) Limitations and pre-requisites of Analysis

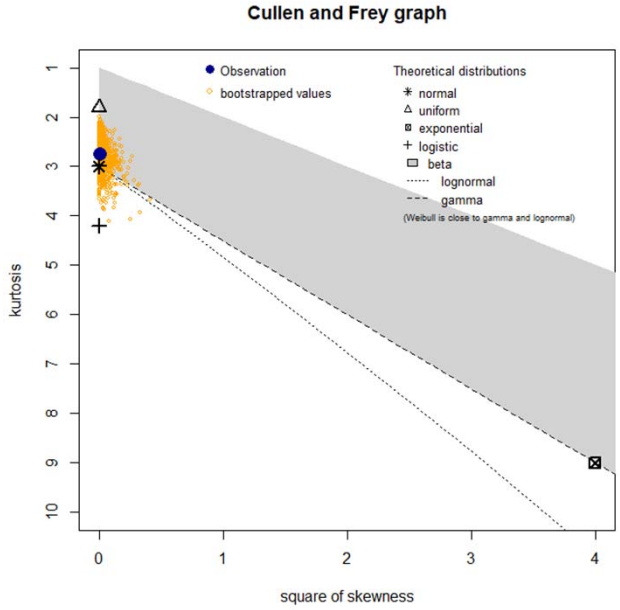
Distribution

Shapiro-Wilk test of normality

Transformation? (we will see this in R)

Robustness of the test?





Basic statistics

ANOVA

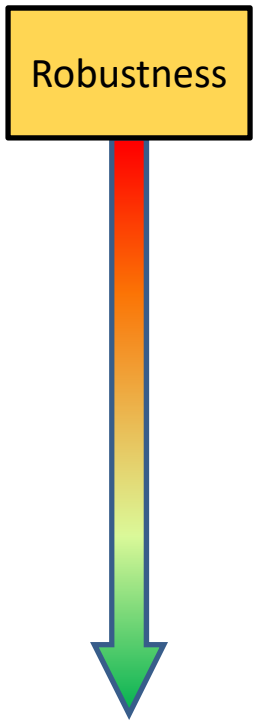
Continuous (numerical)

Samples are chosen randomly

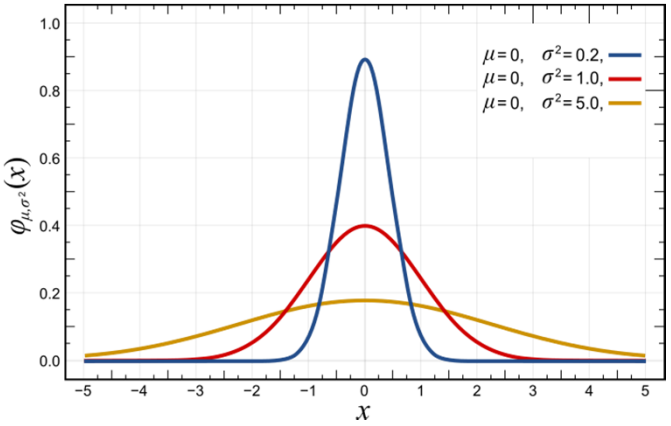
Normal distribution for each group

Homoscedasticity

Degrees of freedom = n-1 ≥ 2



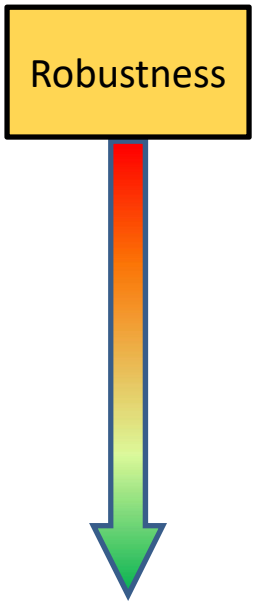
Standard Deviation:
determines shape of the distribution
measures amount of dispersion
correlated with variance



Basic statistics

KRUSKAL-WALLIS

Less powerful detecting variance



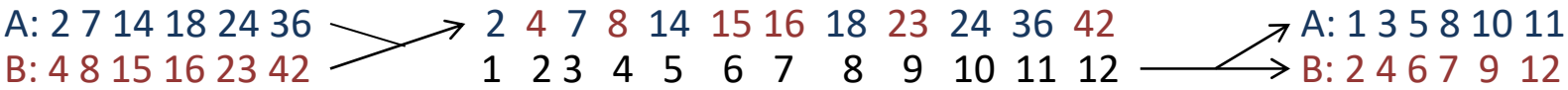
Sortable data

Samples are chosen randomly

Degrees of freedom = n-1 ≥ 2

Homoscedasticity

Unique values:
It works by sorting the data according to the values, ranking them and then comparing the ranks assigned to each group. The p-value will not be accurate if there is more than one sample with the same rank



Basic statistics

We know the characteristics of our data
We know which test to use

0) What do we want to know?
1) Which type of data do we have?
2) Descriptive analysis/plot
3) Limitations of the analysis -> Choose

DON'T FORGET "0"

Basic statistics

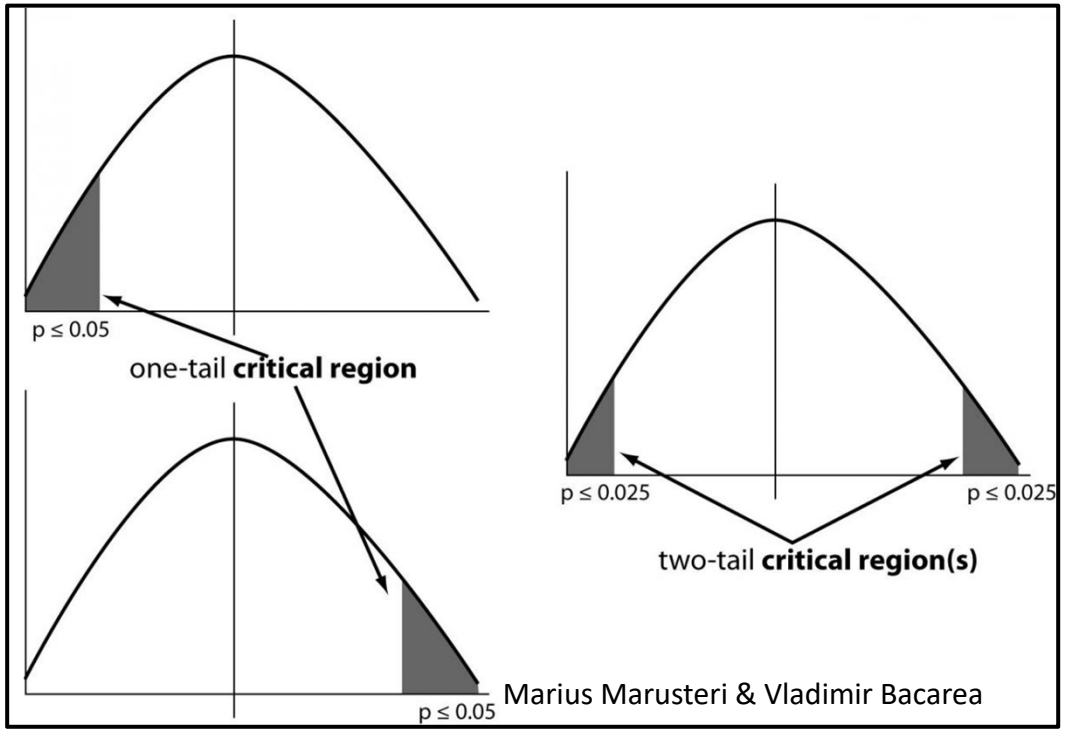
What does the p-value tells us?

Ho = all the same

Ha = there are differences

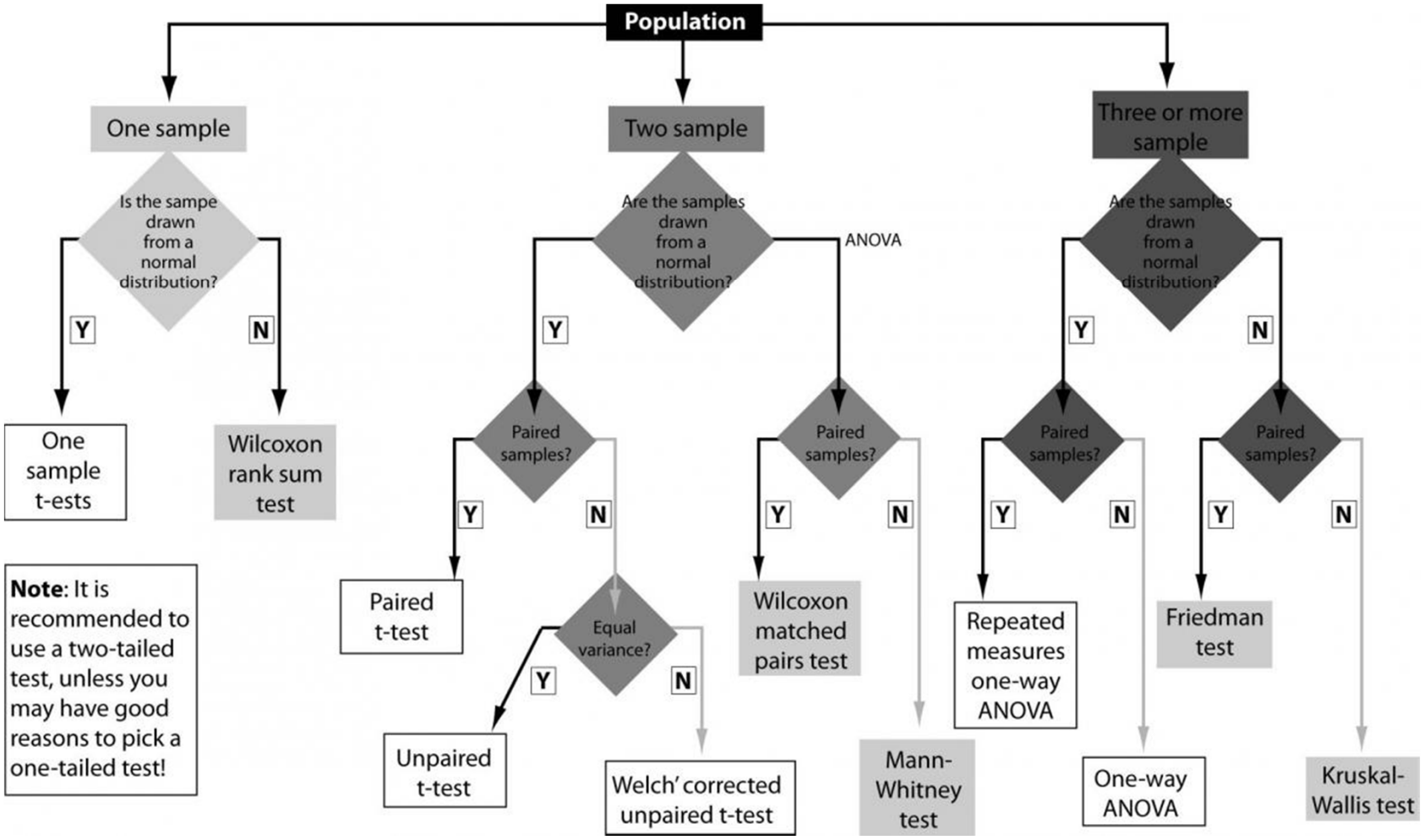
p-values

Probability by chance, that the value found belongs to the distribution but is rejected as independent



Basic statistics

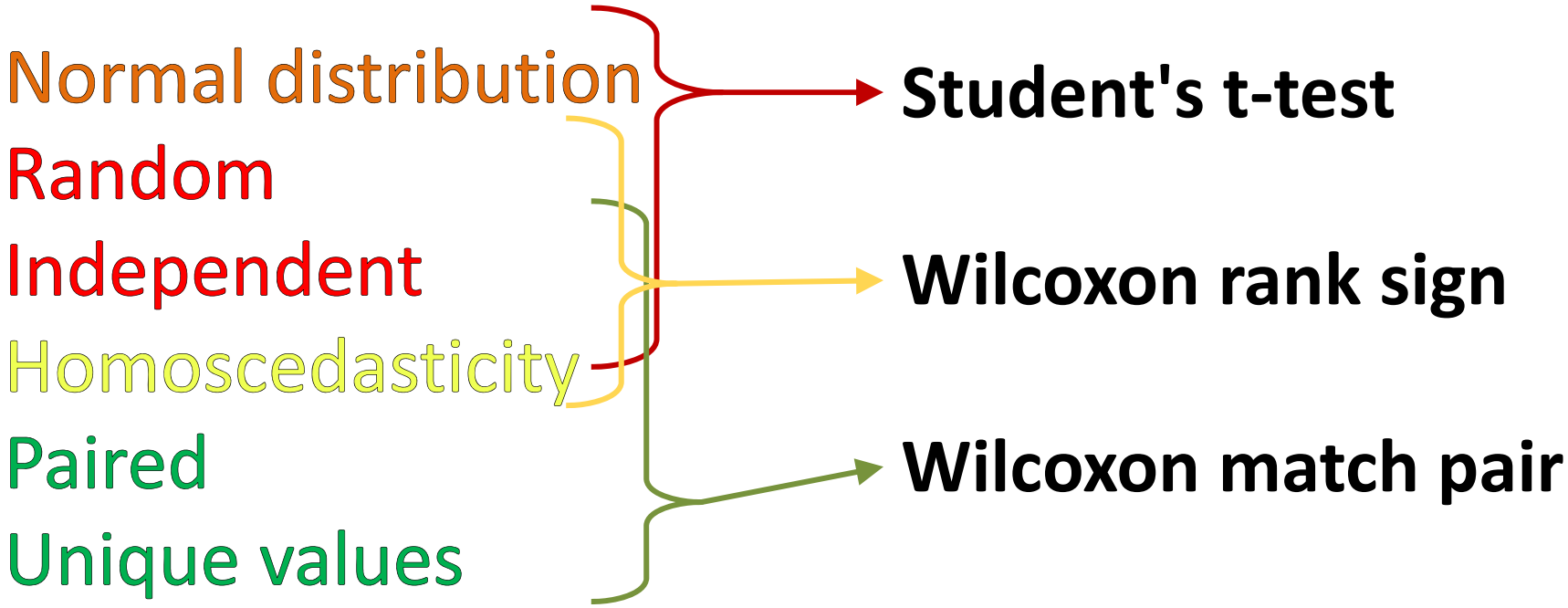
Differences of variance between groups (continuous data)



Marius Marusteri & Vladimir Bacarea

Basic statistics

Comparing two groups (df = 1)



Error Type I: **Rejecting a true null hypothesis <- correction for multiple tests**
Error Type II: Not rejecting a false null hypothesis