BIOSTATISTICS COURSE #2

Statistical Modeling #1

SEPTEMBER 2025

SUMMARY OF THE COURSE #2

INTRODUCTION 01

INTRODUCTION
STATISTICAL MODELING VS MACHINE LEARNING

Statistics

High certainty that **most assumptions will be satisfied,** prior to constructing your model

Small-to-mid sized data sets

Expectations that there will be **some uncertainty in predictions**

High interpretability

A need for a **simple structure/ model**

ML

There are **several or even countless ways to train your algorithm**

You have a **large data set**

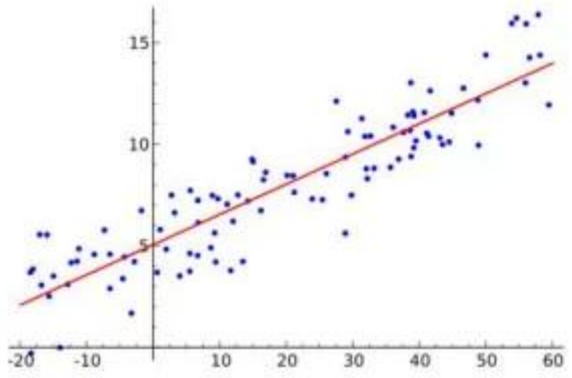
You're looking to make a prediction that **is not based on other independent variables or their relationships with each other**

There are **low interpretability options.**

INTRODUCTION
WHAT IS A STATISTICAL MODEL ?

- **More or less complex equation which aims to link a variable** (called "dependant *variable*") with to explanatory variables (also called "*factors*")

- Example with the classical linear model :



$$Y = a \times X + b + \varepsilon$$

with Y : variable to explain (continuous)
X : explanatory variable (continuous)
a : coefficient of explanatory variable X (slope of the regression)
b : intercept (value of Y when X = 0)
ε : model residuals (proportion of the variability of Y not explained)

INTRODUCTION
WHAT IS A STATISTICAL MODEL ?

- Two possible goals for statistical modeling :
  - ➢ Explain and quantify the relationship (linear or not) between a variable to explain $\boldsymbol{Y}$ (often continuous) and explanatory variables $\boldsymbol{X}$ (***inference***)
  - ➢ Predict a value $\hat{Y}$ from explanatory variables $\boldsymbol{X}$ (***prediction***)
- The best explanatory model is not always the best predictive model !
- **Various metrics** are available for model quality assessment : R^2, Q^2 (also called predictive R^2), AIC, BIC…

INTRODUCTION
EXHAUSTIVE LISTING OF STATISTICAL MODELS

- Modeling of a **continuous variable** :
  - Linear model (univariate or multivariate)
  - ANOVA (**AN**alysis **O**f **VA**riance), MANOVA, ANCOVA, MANCOVA
  - Non-linear model (quadratic, cubic, exponential, logarithmic…)
  - Mixed models
  - Factorial analysis : PCA

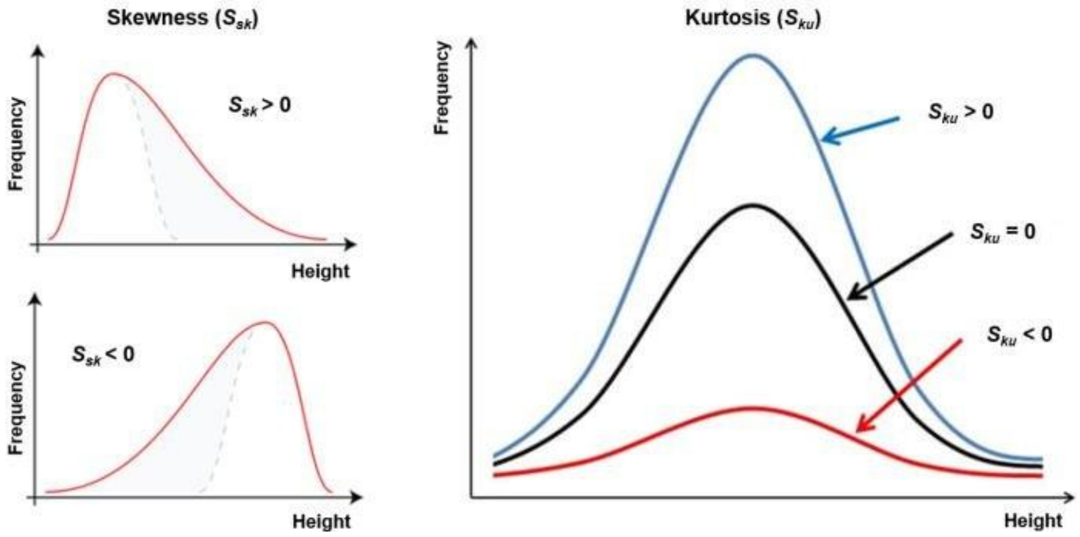
- Modeling of a **categorical variable** :
  - Two modalities (Yes vs No) : binary logistic regression
  - More than two modalities : ordinal or nominal logistic regression
  - Factorial analysis : CFA, MCA

DENSITY DISTRIBUTIONS

02

DENSITY DISTRIBUTIONS
INTRODUCTION

- Hundreds of different distributions are available
- The most useful one : the gaussian distribution (also called normal law)
- Two metrics which resume the shape of a distribution :
  - ➢ Skewness (asymmetry of distribution)
  - ➢ Kurtosis (tailedness of distribution)
- Normal distribution : skewness of 0



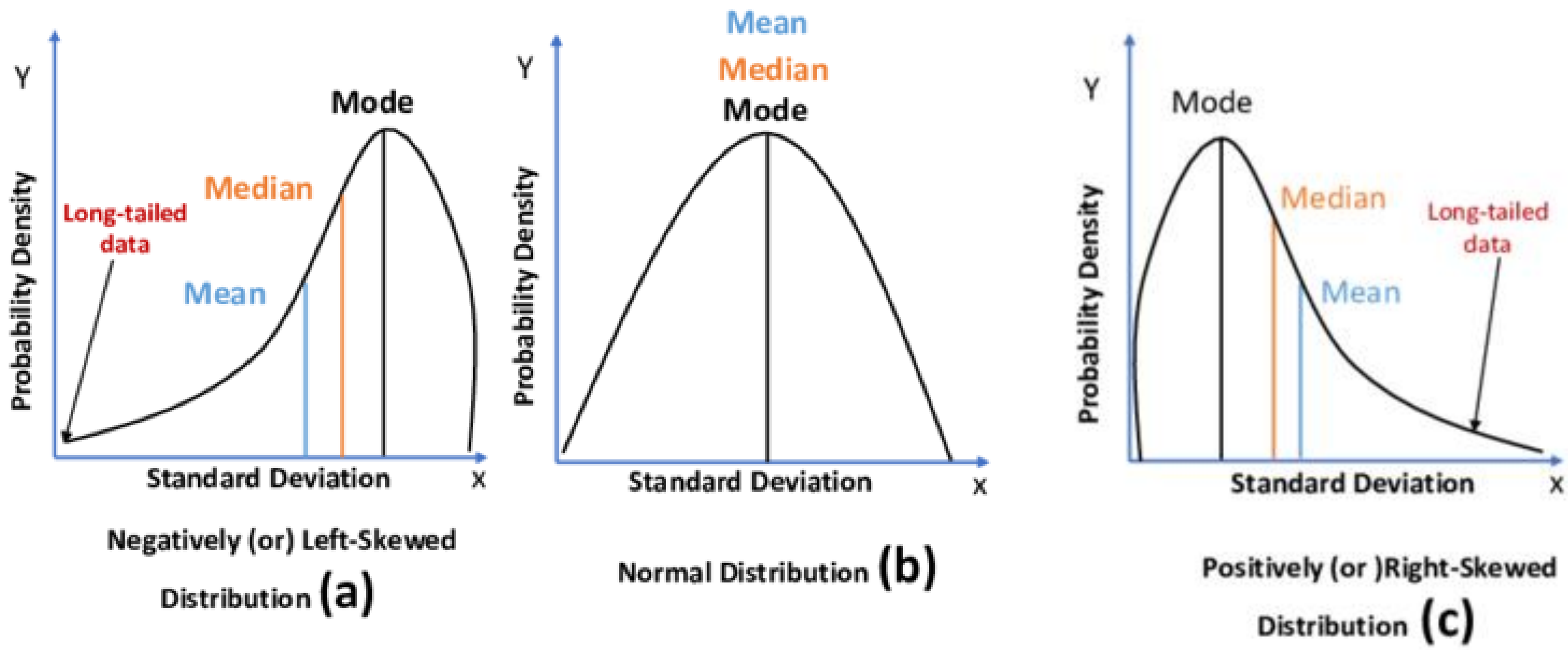
DENSITY DISTRIBUTIONS

FUNCTIONS FOR SKEWNESS AND KURTOSIS

- Skewness of a distribution : *skewness* function (package ***parameters***)
  Parameters : *x* = list of continuous or categorical values
  na.rm (boolean : true or false) = remove NA values ?

- Kurtosis of a distribution : *kurtosis* function (package ***parameters***)
  Parameters : *x* = list of continuous or categorical values
  na.rm (boolean : true or false) = remove NA values ?

DENSITY DISTRIBUTIONS
IMPACT OF DISTRIBUTION ON CENTRAL TENDENCY

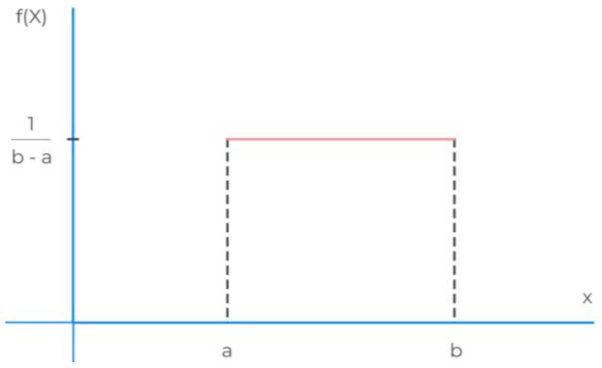


DENSITY DISTRIBUTIONS

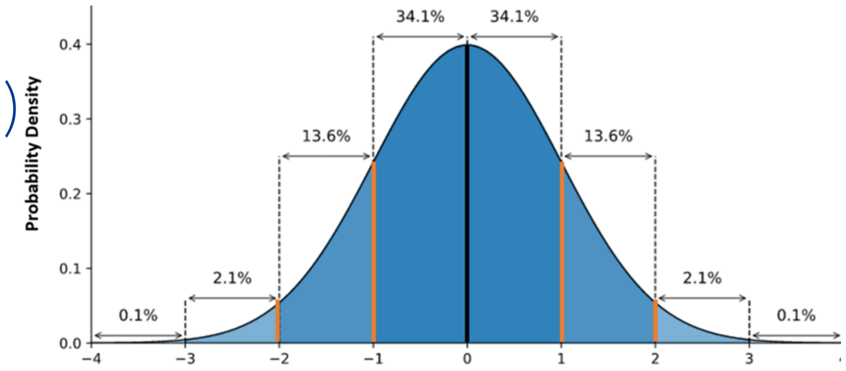
FUNCTION FOR DENSITY DISTRIBUTION SIMULATIONS

Generation of random variables is easy with R

- Uniform distribution : *runif* function (*stats* package)
  Parameters : *n* = number of random values
  min = minimum threshold of values
  max = maximum threshold of values



- Normal distribution : *rnorm* function (*stats* package)
  Parameters : *n* = number of random values
  mean = mean of values
  sd = standard-deviation of values

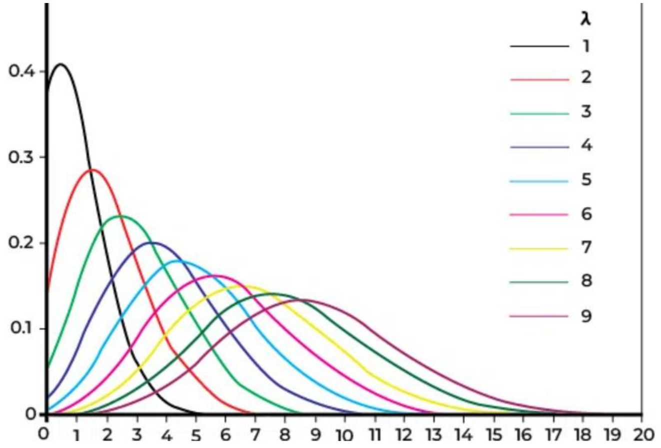


DENSITY DISTRIBUTIONS

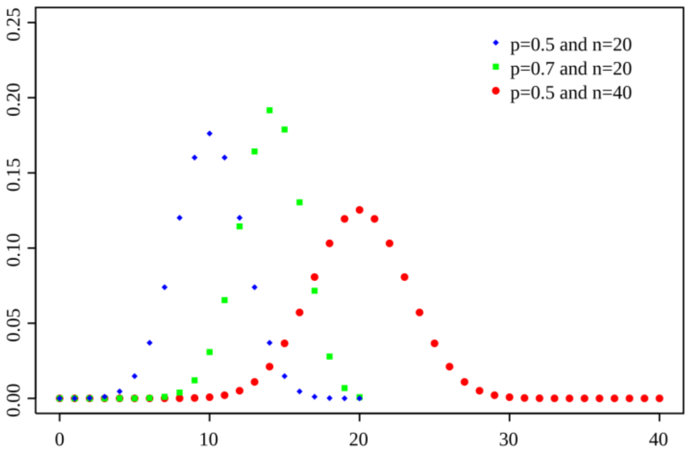
FUNCTION FOR DENSITY DISTRIBUTION SIMULATIONS

Generation of random variables is easy with R

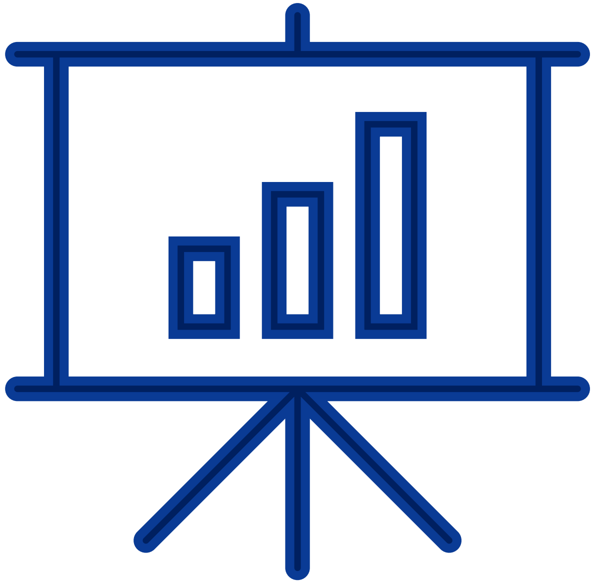
- Poisson distribution : *rpois* function (*stats* package)
  Parameters : *n* = number of random values
  lambda = shape of the distribution



- Binomial distribution : *rbinom* function (*base* package)
  Parameters : *n* = number of random values
  size = number of trials
  prob = probability of success

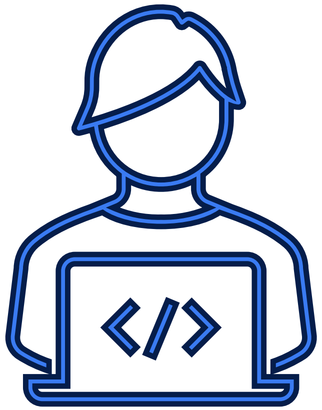


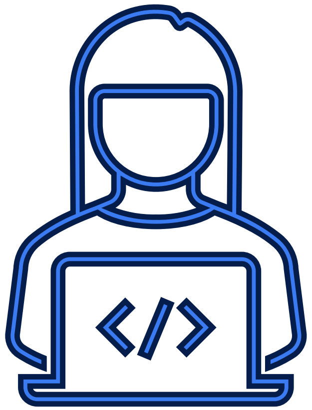
DENSITY DISTRIBUTIONS



Live demo

DENSITY DISTRIBUTIONS





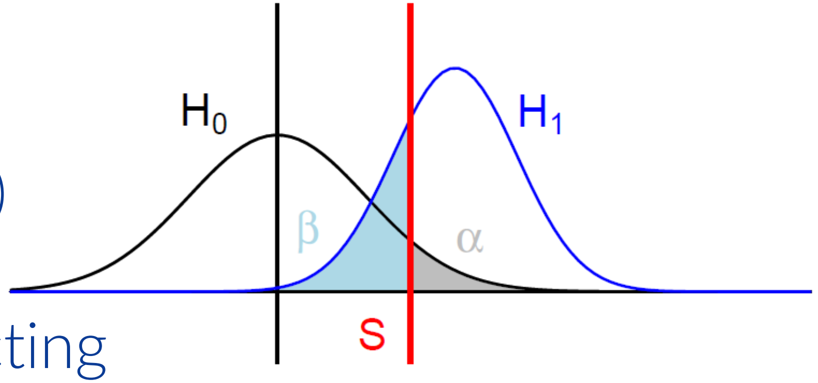
Time to play !
(20 minutes)

STATISTICAL TESTS

03

STATISTICAL TESTS
DEFINITION

- Statistical **inferential** procedure which allows to test **hypotheses**



- Two tested hypotheses : H_0 (null hypothesis) and
  H_1 (alternative hypothesis)

- Result : **p-value**, probability to be wrong when rejecting the null hypothesis.

- Two levels of risk : **alpha** (σ) and **beta** (β)

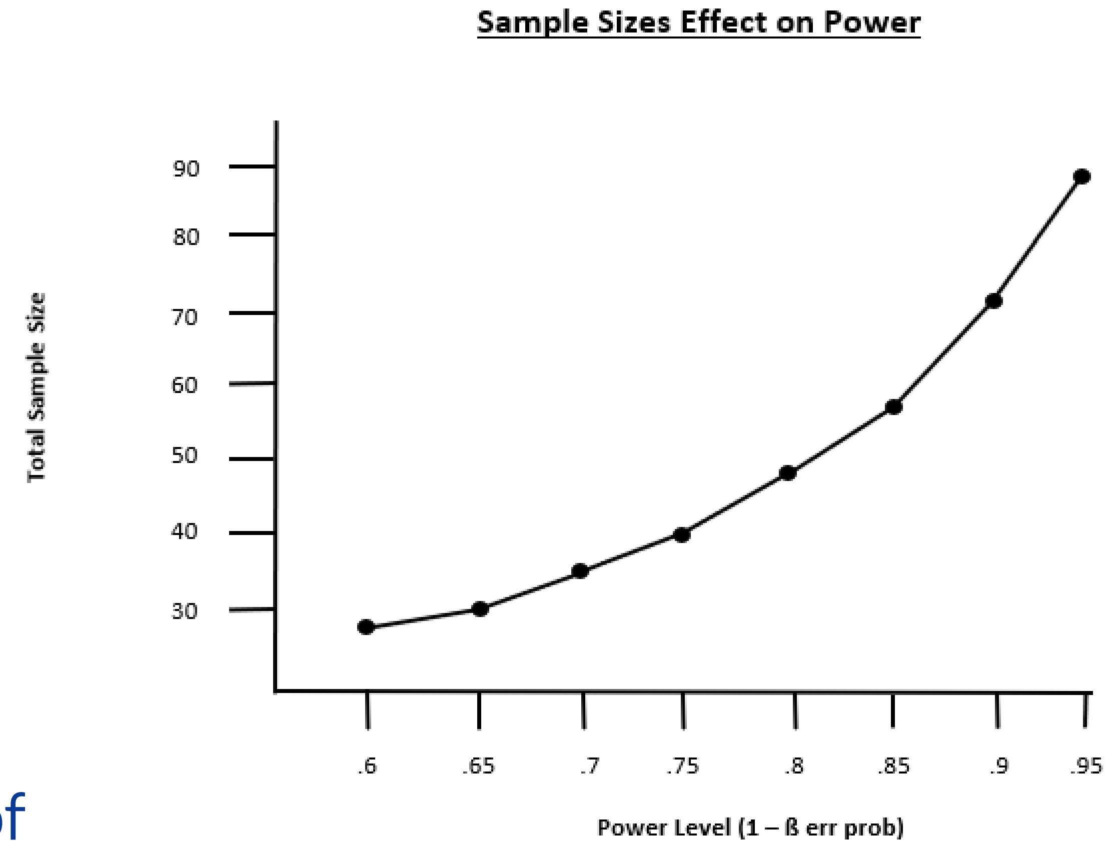
- H_0 rejected when pvalue is < α (0.05 or 5%)

- **Two-sided test** (difference) vs **one-sided** (inferiority or superiority)

		The Truth	
		H_0 is True	H_0 is False
The Outcome of the Hypothesis Test	Fail to Reject H_0	Correct Decision	INCORRECT DECISION (Type II Error) Beta (β) Risk
	Reject H_0	INCORRECT DECISION (Type I Error) Alpha (α) risk	Correct Decision *Power (1 - β)*

STATISTICAL TESTS
STATISTICAL POWER

- Statistical power = 1 - β (beta risk)
- Represents the **ability of a test to detect a significant effect size**
- Strongly correlated with **the number of observations** and **effect size**



Note: As the sample size increases in the model, so does power.

STATISTICAL TESTS
FAMILIES OF TESTS

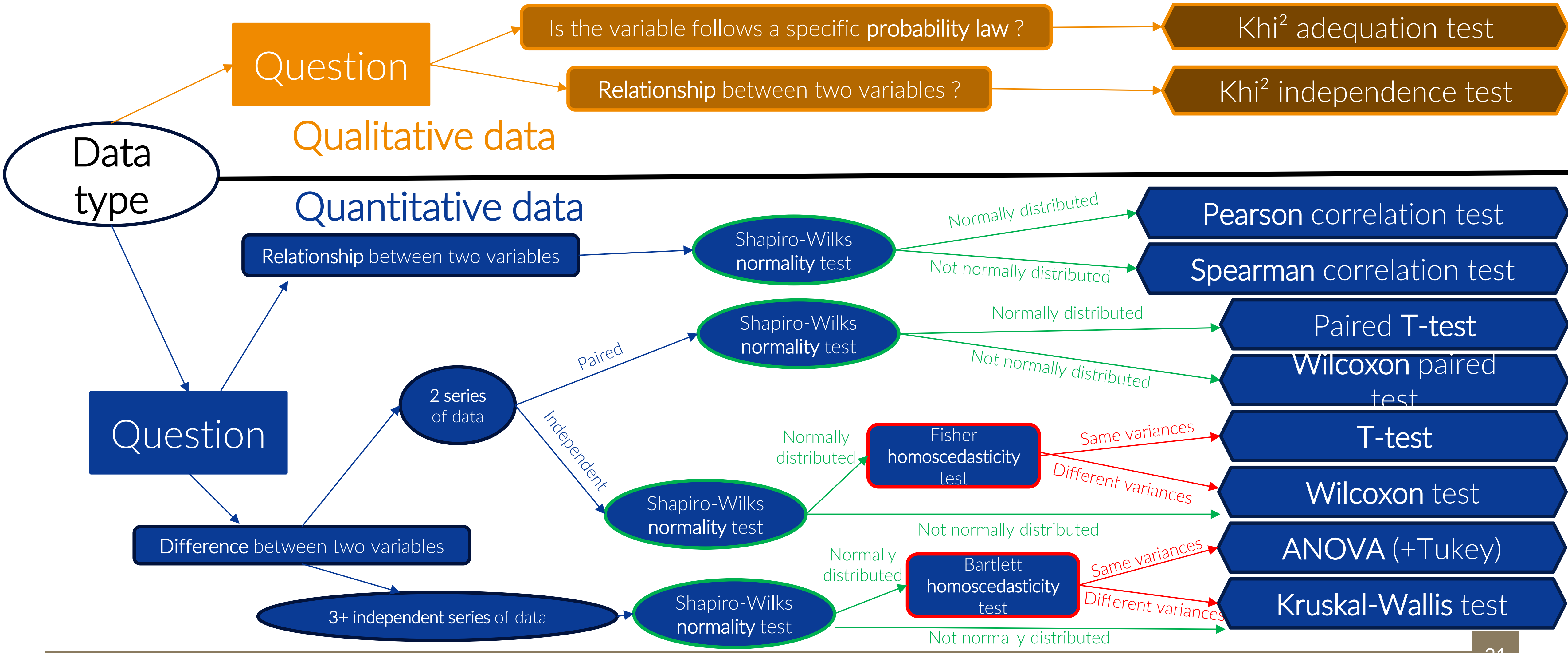
- Two families of tests :
  - **Parametric tests** : require to check assumptions on data (normality of distribution at least)
  - **Non-parametric tests** : alternative tests when parametric tests are not valid
- Two dimensions of tests :
  - **Univariate tests** : allows to test hypotheses at the overall level (no groups)
  - **Multivariate tests** : allows to test hypotheses at group level (2 or more)

STATISTICAL TESTS
EXHAUSTIVE LIST OF TESTS

- Quantitative data :
  - Test the **distribution** of a variable (normality…) : Kolmogorov-Smirnov, Shapiro-Wilks…
  - Compare the **level** of a variable (mean, median…) **to a reference level**.
  - Highlight **outliers** (test de Grubbs)
  - Compare the **level of a variable between two or more groups** (mean, median…) : T-test, univariate ANOVA, Wilcoxon test, Kruskal-Wallis test…
  - Test the **strength of the relationship between two variables** (correlations)
  - Compare the **variability of a variable between groups** (variance)

- Qualitative data : Compare occurrence of factors (counts and percentages) between groups (Chi^2 test)

STATISTICAL TESTS
HOW TO CHOOSE THE RELEVANT TEST ?

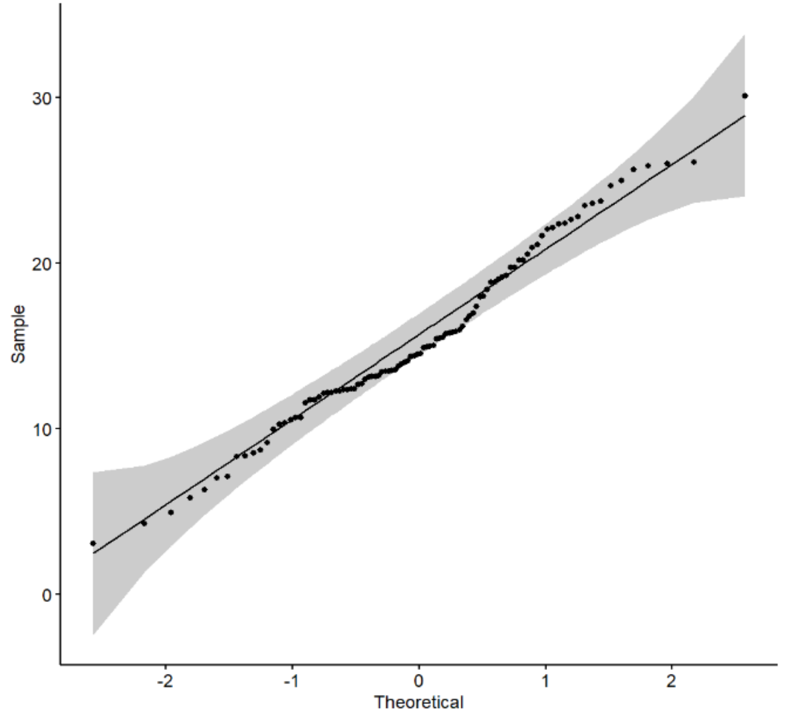


UNIVARIATE TESTS

STATISTICAL TESTS
UNIVARIATE TEST : NORMALITY TEST

- **Goal** : test if the **distribution of the values** of a quantitative variable **is gaussian** (or normal).

- **Hypotheses** :
  - H_0 : the values follow the **normal law**
  - H_1 : the values do not follow the **normal law**

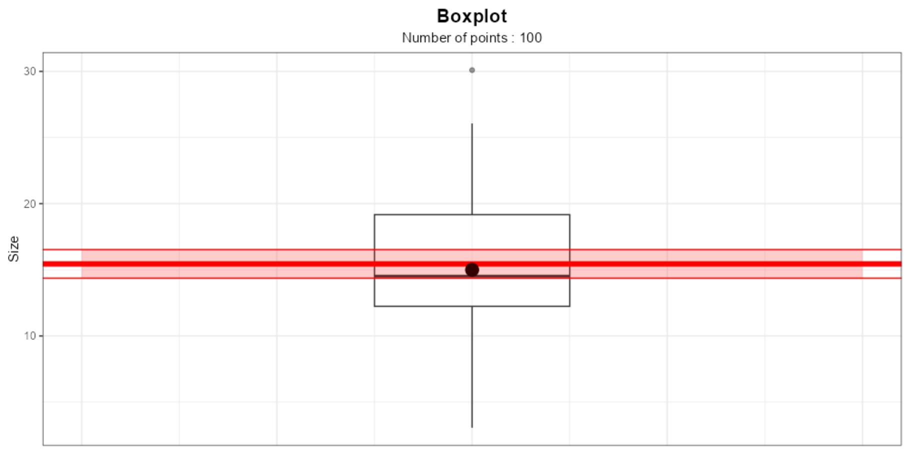


- In R : function ***shapiro.test***(data) (package ***stats***)
  if pvalue ≥ 0.05 = data is normally distributed

- Plot for normality : function ***ggqqplot***(data) (package ***ggpubr***) : if all points are approximately aligned on the diagonal : data is normality distributed

STATISTICAL TESTS
UNIVARIATE TEST : TEST MEAN TO A REFERENCE

- **Goal** : test if a variable has the same **mean** compared to a **reference value** (normally distributed variable)



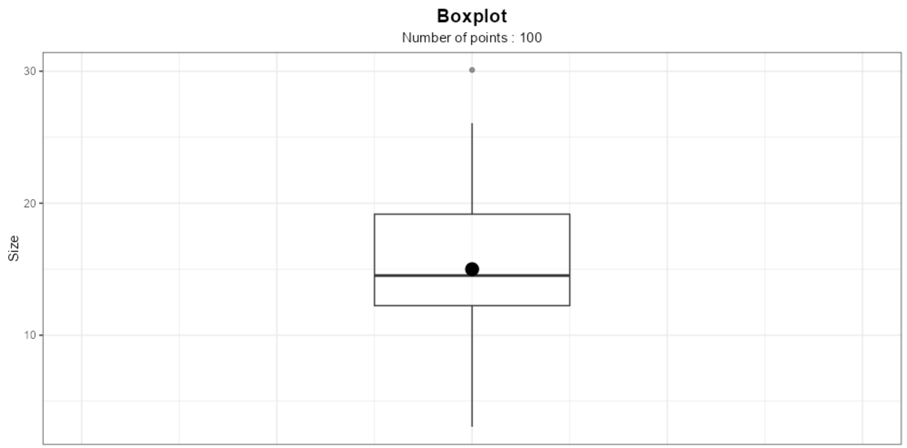
- **Hypotheses** :
  - H_0 : the variable has the same **mean**
  - H_1 : the variable has a different **mean**

- In R : function ***t.test***(*data*, *mu = reference level*) (package ***stats***)
  if pvalue < 0.05 = mean is significantly different from the reference level

- **Outputs** : test statistic, degrees of freedom, pvalue, 95% confidence interval (CI) of the mean.

STATISTICAL TESTS
UNIVARIATE TEST : TEST MEDIAN TO A REFERENCE

- **Goal** : test if a variable has the same **median** compared to a reference value (not-normally distributed variable)



- **Hypotheses** :
  - H_0 : the variable has the same **median**
  - H_1 : the variable has a different **median**

- In R : function ***wilcox.test***(*data*, *mu = reference level*) (package ***stats***)
  if pvalue < 0.05 = median is significantly different from the reference level

- **Outputs** : test statistic, pvalue

STATISTICAL TESTS
UNIVARIATE TEST : TEST PROPORTION TO A REFERENCE

- **Goal** : test if the **proportion** of a modality of a binary variable is equal to a target percentage.

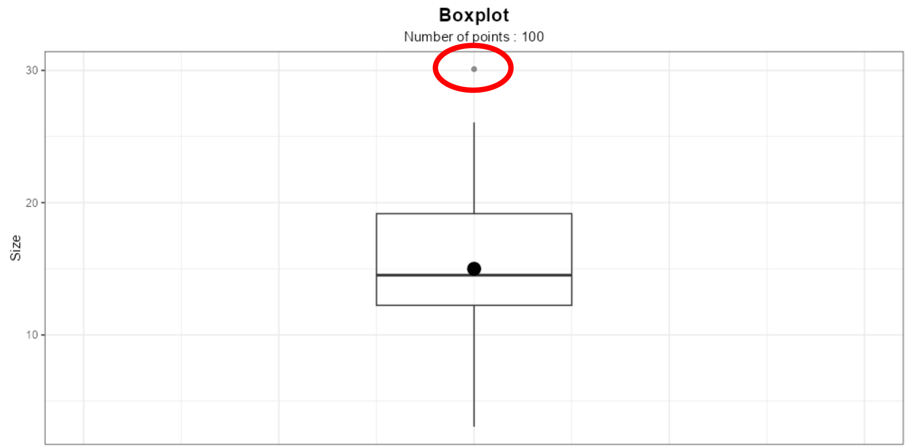
- **Hypotheses** :
  - H_0 : modality X has the same **proportion**
  - H_1 : modality X has a different **proportion**

- In R two options :
  - If small sample (n < 30) : function ***binom.test***(*x = number of success, n = number of trials, p = target proportion*) (package ***stats***)
  - If large sample (n ≥ 30) : function ***prop.test***(*same parameters*) (package ***stats***)

- **Outputs** : test statistic, proportion, pvalue, 95% CI of the proportion.

STATISTICAL TESTS
UNIVARIATE TEST : OUTLIER DETECTION (GRUBBS TEST)

- **Goal** : test if some **outliers exist** in a variable

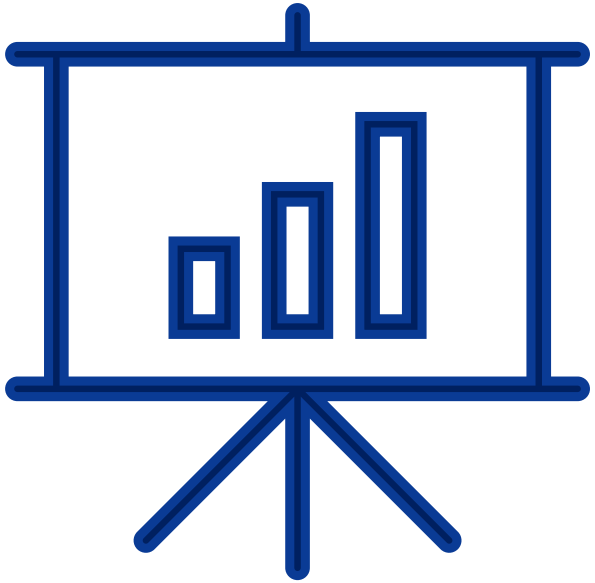


- **Hypotheses** :
  - H_0 : the highest (or lowest) value **is not an outlier**
  - H_1 : the highest (or lowest) value **is an outlier**

- In R : function ***grubbs.test***(*data*, *type*, *opposite*) (package ***outliers***)
  if pvalue < 0.05 = an outlier is detected

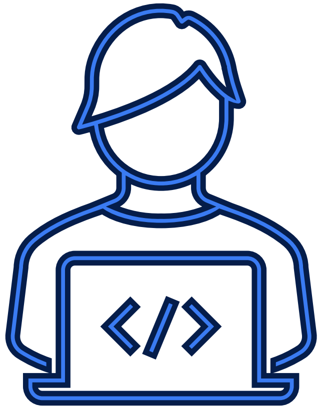
- **Outputs** : test statistic, pvalue

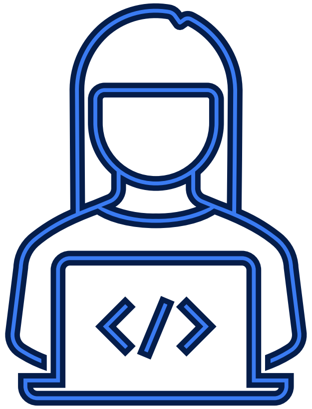
STATISTICAL TESTS
UNIVARIATE TESTS



Live demo

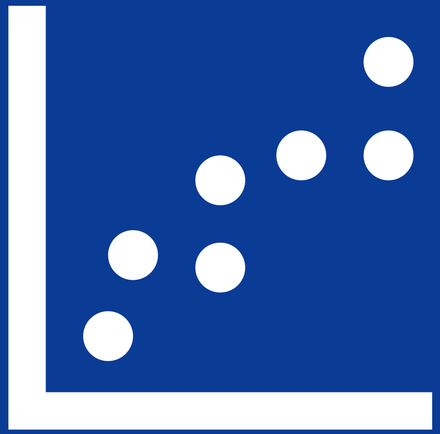
STATISTICAL TESTS
UNIVARIATE TESTS





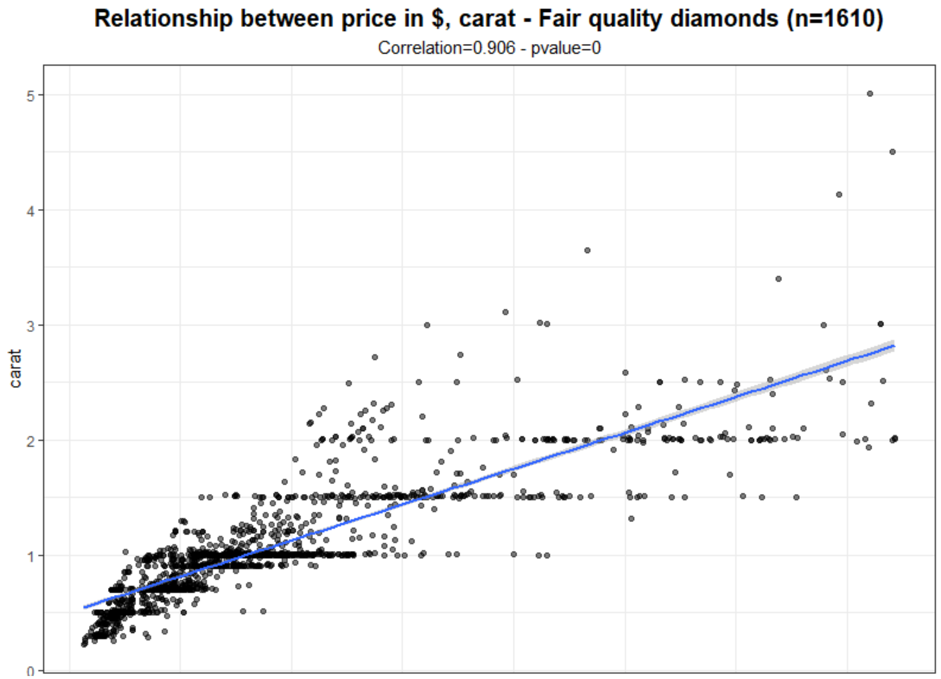
Time to play !
(15 minutes)

MULTIVARIATE TESTS



STATISTICAL TESTS
MULTIVARIATE TEST : CORRELATION

- **Goal** : test if the **coefficient of correlation** between two variables is significantly different from 0.

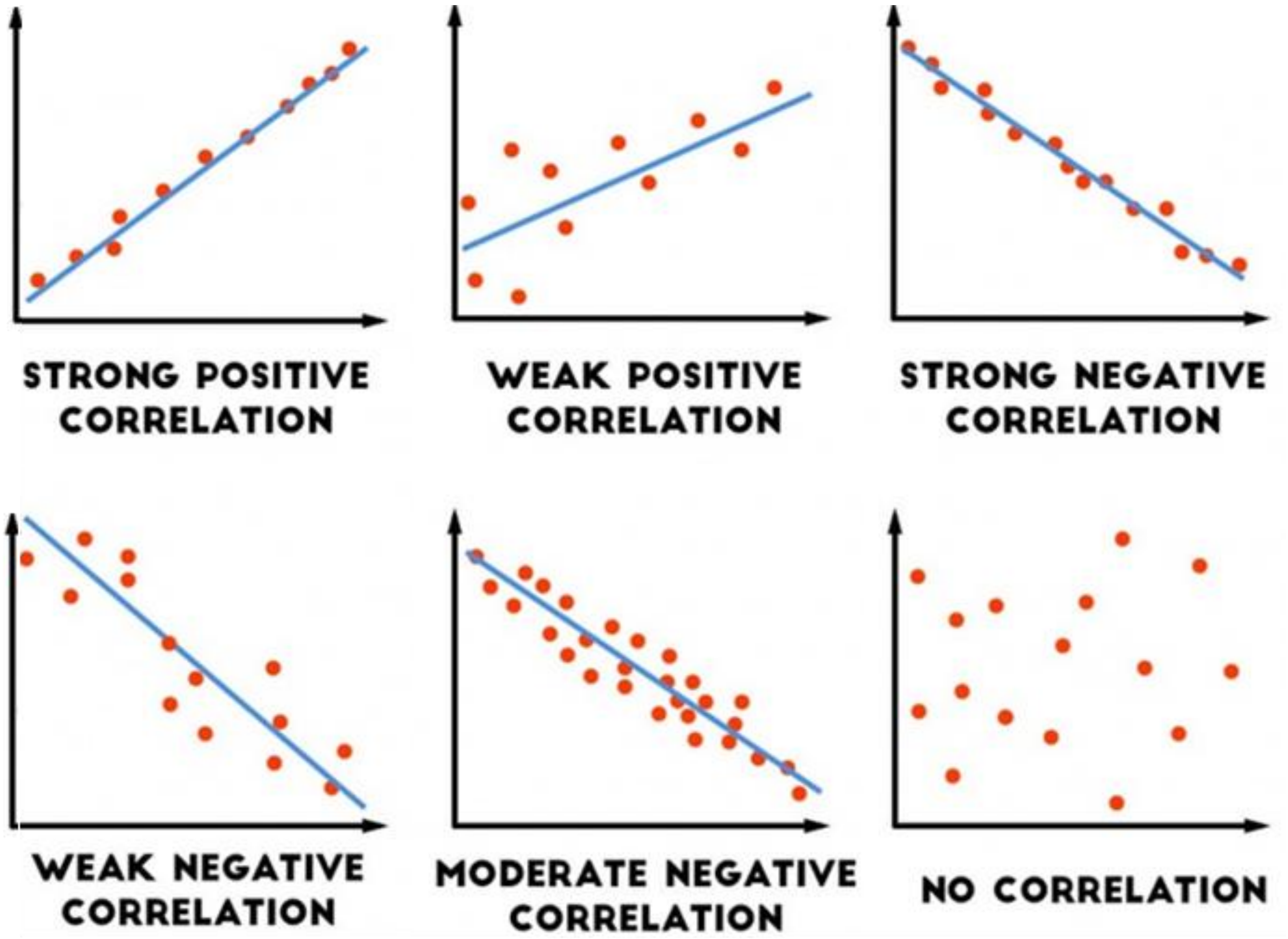


- **Hypotheses** :
  - ➢ H_0 : the coefficient is 0
  - ➢ H_1 : the coefficient is different from 0

- **Warning : a coefficient of correlation with a significant pvalue is not necessarily strong !**
  Correlation > 0.75 in absolute : strong.

- In R : function ***cor.test***(*x*, *y*, *method*) (package ***stats***)
  Method to use : **pearson** (normality of distributions) vs **spearman**

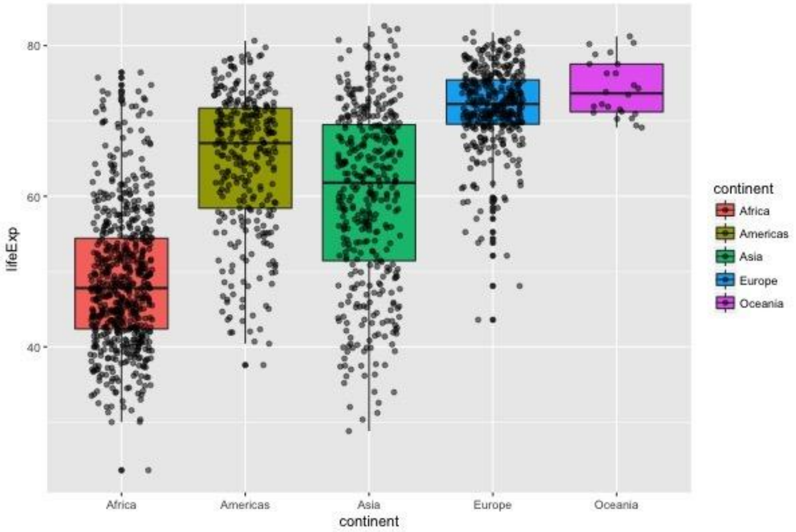
STATISTICAL TESTS
MULTIVARIATE TEST : CORRELATION



STATISTICAL TESTS
MULTIVARIATE TEST : COMPARISON OF VARIANCES

- **Goal** : compare the **variance** of a variable between **independents groups** (homoscedasticity)

- **Hypotheses** :
  - ➢ H_0 : the **variances** of each group are equivalent
  - ➢ H_1 : At least one group has a different **variance**

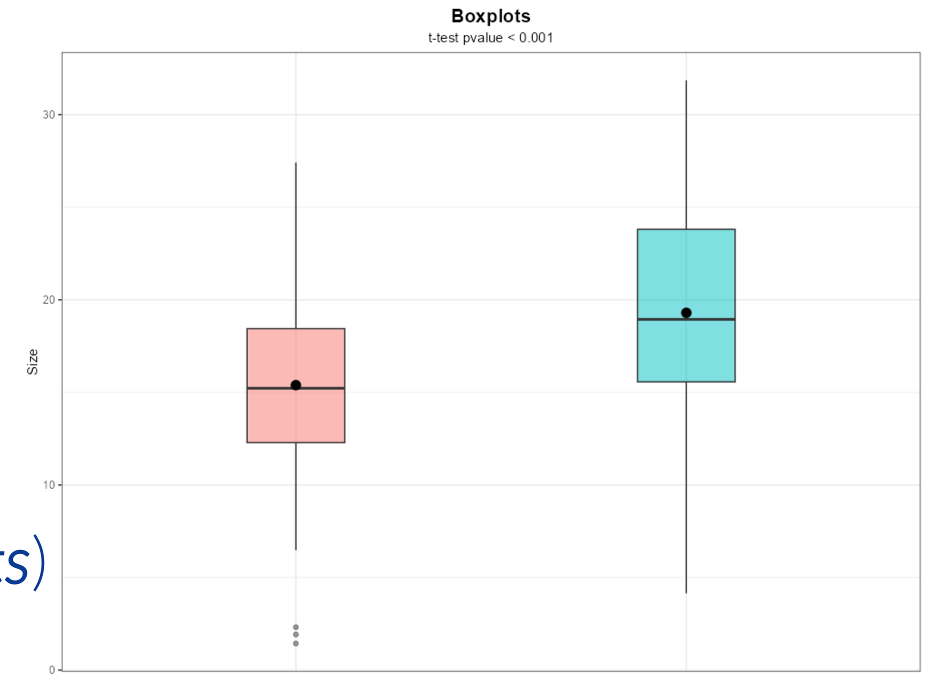


- In R two options (package ***stats***) :
  - ➢ If two groups : function ***var.test***(*x*, *y*, *ratio*)
  - ➢ If more than two groups : function ***bartlett.test***(*values ~ group*)

- **Outputs** : test statistic, pvalue

STATISTICAL TESTS
MULTIVARIATE TEST : COMPARISON OF MEANS

- **Goal** : compare **means** between two **independent groups** (normally distributed variable and **homoscedasticity**)



- **Hypotheses** :
  - H_0 : the two **means** are equivalent
  - H_1 : the two **means** are different

- In R : function ***t.test***(*x*, *y*, *mu*) (package ***stats***)

- **Outputs** : test statistic, difference of means, pvalue, 95% CI of the difference.

STATISTICAL TESTS
MULTIVARIATE TEST : COMPARISON OF MEANS

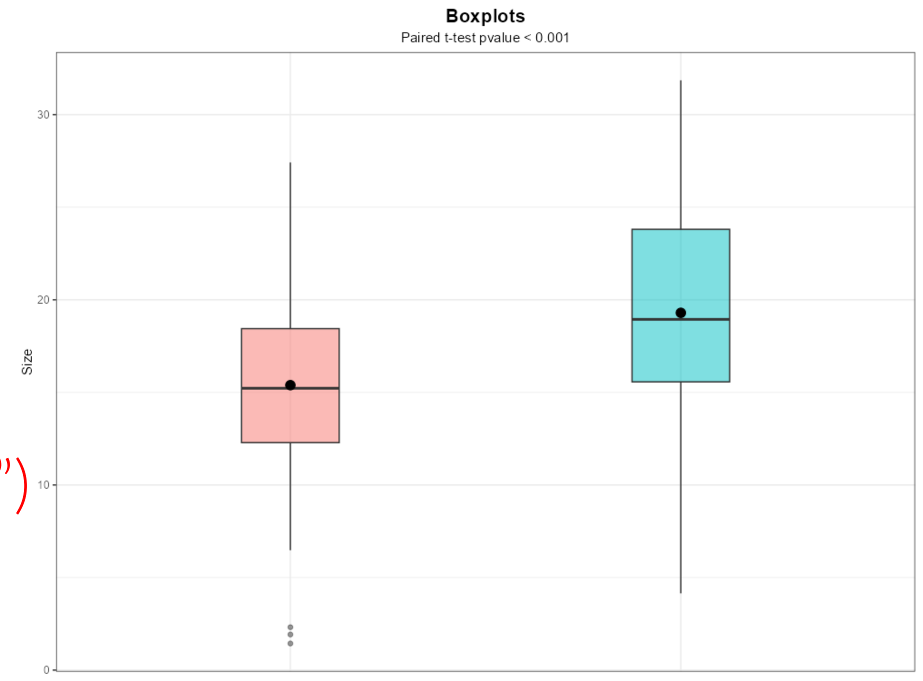
- **Goal** : compare **means** between two **paired groups** (normally distributed variable and **homoscedasticity**)

- Hypotheses :
  - H_0 : the two **means** are equivalent
  - H_1 : the two **means** are different

- Only applicable on **paired** data (“*before / after*”)

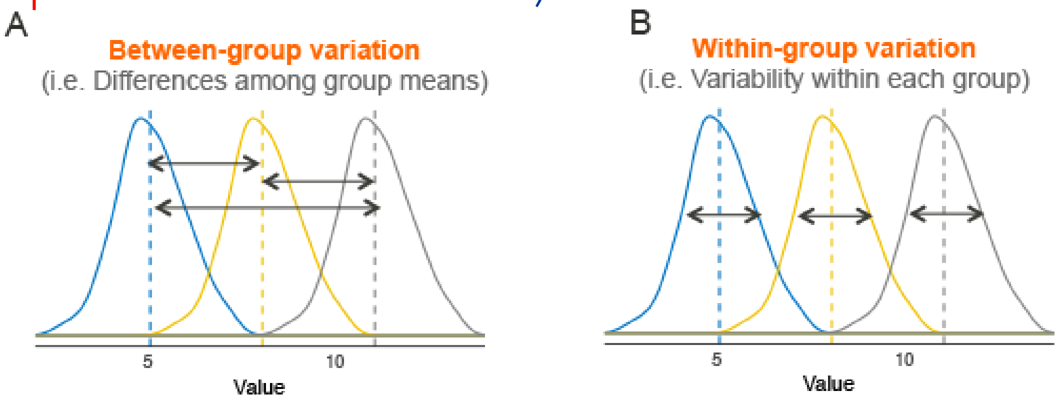
- In R function ***t.test***(*x, y, mu, paired=T*) (package ***stats***)

- **Outputs** : test statistic, difference of means, pvalue, 95% CI of the difference.



STATISTICAL TESTS
MULTIVARIATE TEST : COMPARISON OF MEANS (ANOVA)

- **Goal** : comparer the **means** of a variable between more **than 2 independent groups** (normally distributed variable and equivalent variances)



- Hypotheses :
  - H_0 : the **means** are equivalent
  - H_1 : At least two **means** are different

- Need to perform **post-hoc pairwise tests** in case of overall significant pvalue with **Tukey's tests**

- In function ***anova_test***(*y ~ group) and* ***tukey_hsd(****y ~ group)* (package ***rtatix***)

- **Outputs** : ANOVA table and pairwise comparisons (for post-hoc tests)

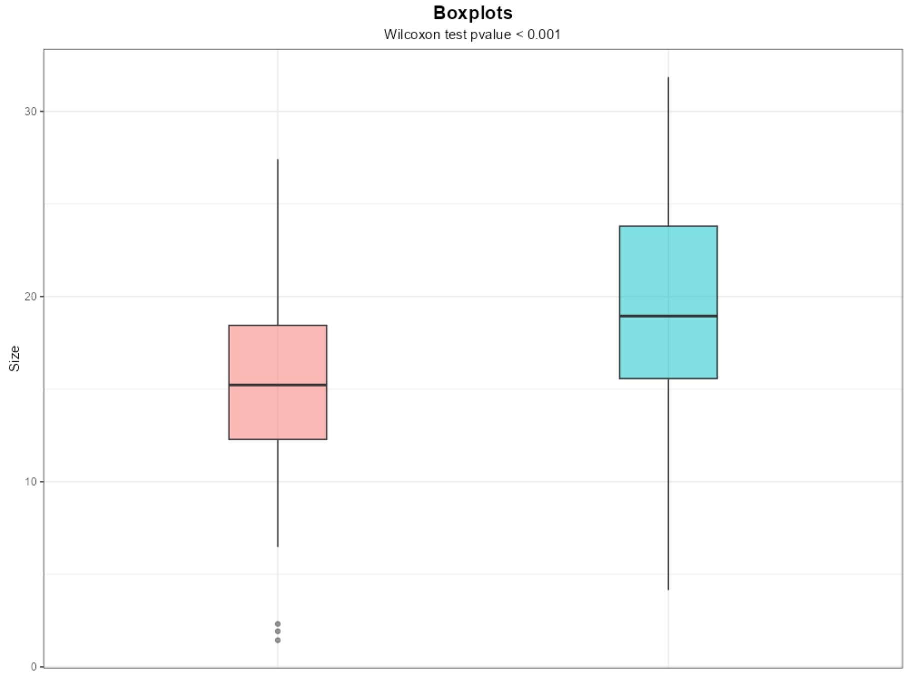
STATISTICAL TESTS
MULTIVARIATE TEST : COMPARISON OF MEDIANS

- **Goal** : compare the medians of a variable between **two independent** groups (non-normally distributed variable)

- **Hypotheses** :
  - ➢ H_0 : the two **medians** are equivalent
  - ➢ H_1 : the two **medians** are different

- In R : function ***wilcoxon.test***(*x, y, mu)* (package ***stats***)

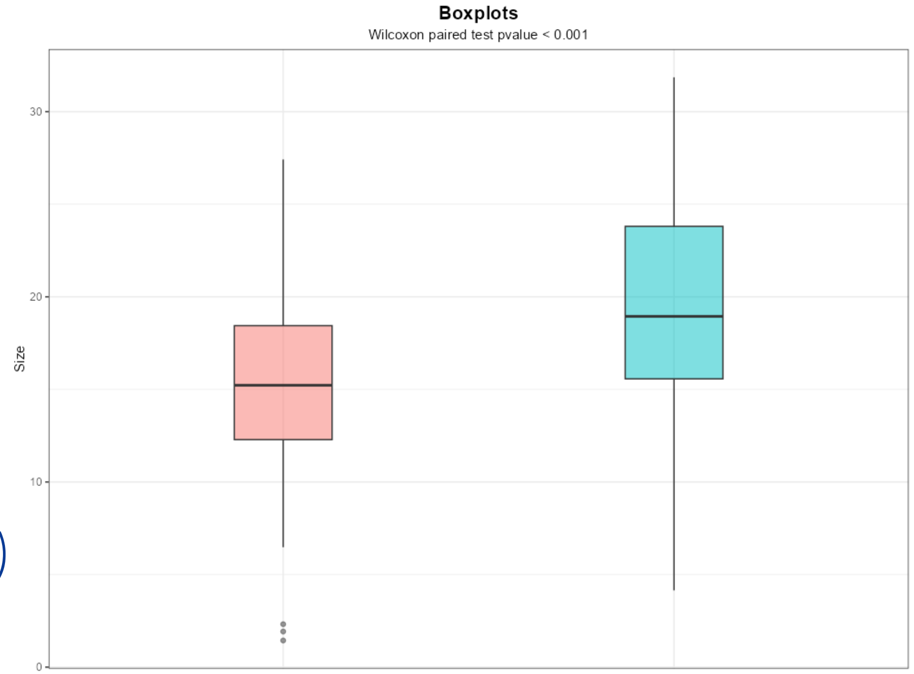
- **Outputs** : test statistic, pvalue.



STATISTICAL TESTS
MULTIVARIATE TEST : COMPARISON OF MEDIANS

- **Goal** : compare the medians of a variable between **two paired** groups (non-normally distributed variable)

- Hypotheses :
  - H_0 : the two **medians** are equivalent
  - H_1 : the two **medians** are different



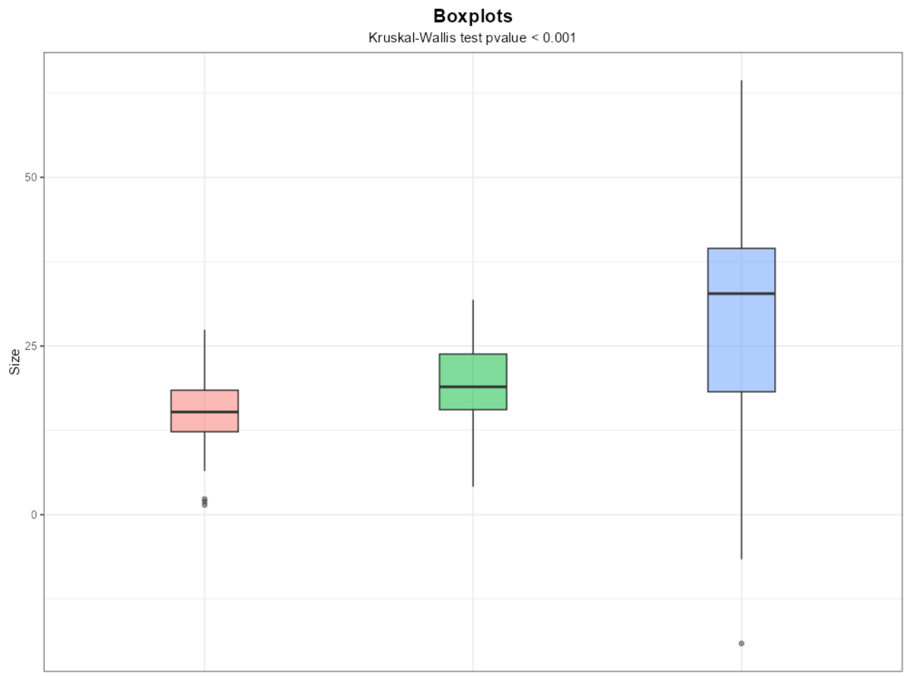
- In R : function ***wilcoxon.test***(*x, y, mu, paired=T*) (package ***stats***)

- **Outputs** : test statistic, pvalue.

STATISTICAL TESTS
MULTIVARIATE TEST : COMPARISON OF MEDIANS

- **Goal** : compare the medians of a variable between **more than two independent** groups (non-normally distributed variable)

- Hypotheses :
  - ➢ H_0 : the **medians** are equivalent
  - ➢ H_1 : At least one **medians** is different



- In R : function ***kruskal.test***(*y ~ group*) and ***pairwise.wilcox.test*** for the post-hoc comparisons between groups (package ***stats***)

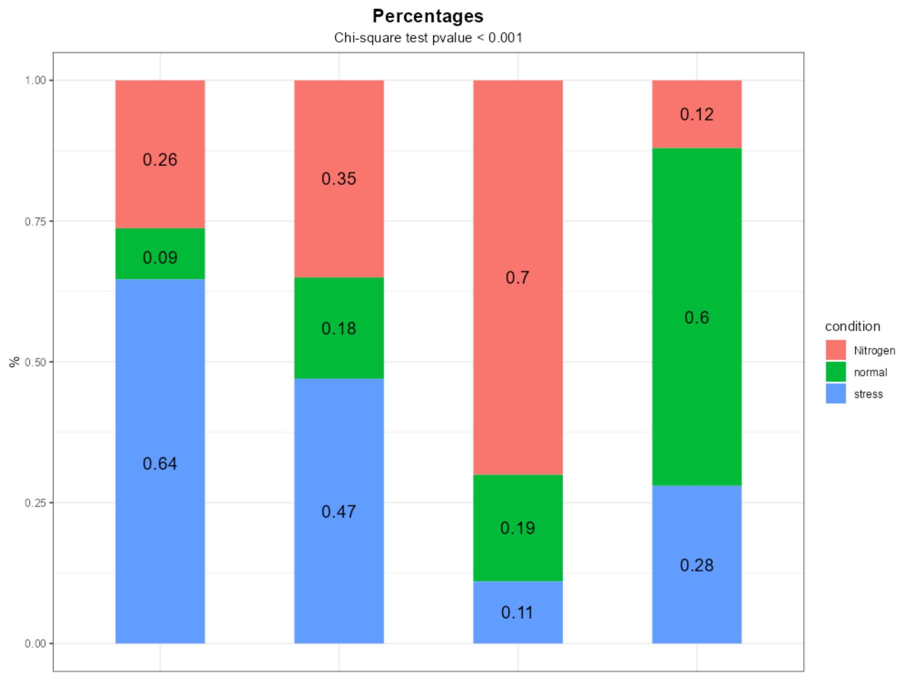
- **Outputs** : test statistic, pvalue and matrix with adjusted pvalues (for post-hoc)

STATISTICAL TESTS
MULTIVARIATE TEST : COMPARISON OF PROPORTIONS

- **Goal** : compare **percentages** of levels of a qualitative variable between **two or more independent groups.**

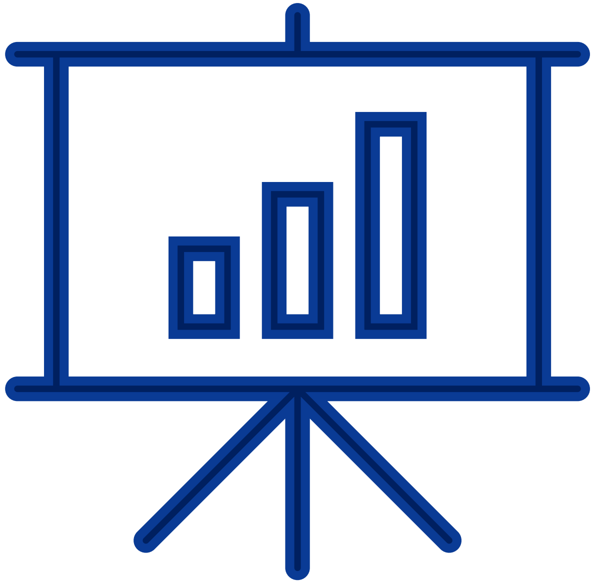
- **Hypotheses** :
  - H_0 : **proportions** are equivalent
  - H_1 : **proportions** are different

- To illustrate: **stacked bar chart**



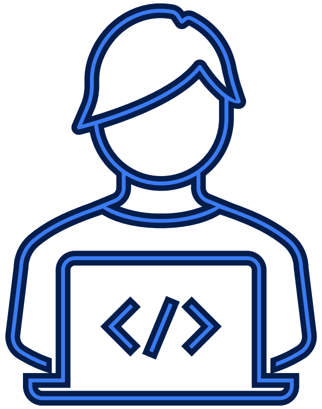
- In R : function ***chisq.test***(*contingency table : 2-ways count table*) and **chisq.posthoc.test** for pairwise comparisons (*on GitHub : ebbertd/chisq.posthoc.test)*

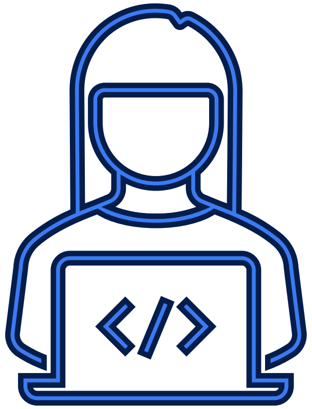
STATISTICAL TESTS
MULTIVARIATE TESTS



Live demo

STATISTICAL TESTS
MULTIVARIATE TESTS





Time to play !
(10 minutes)

UNIVARIATE LINEAR MODELS

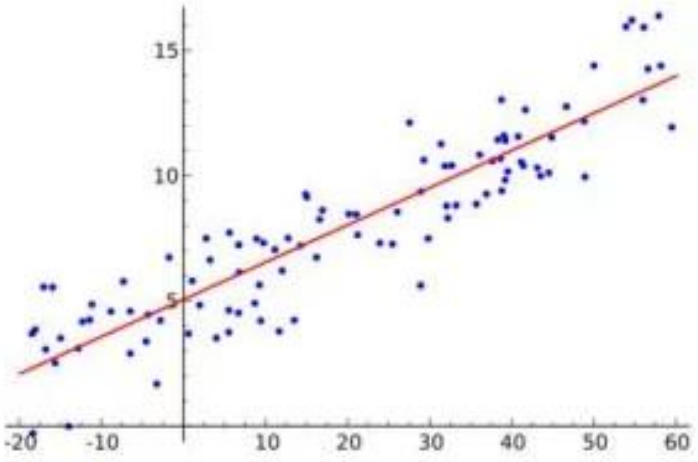
04

UNIVARIATE LINEAR MODELS
HIGHLIGHTS

- **Goal** : explain a continuous variable Y with a continuous variable X

- **Hypothesis to test** : the relationship between the two variables **is linear**

- **Equation** :

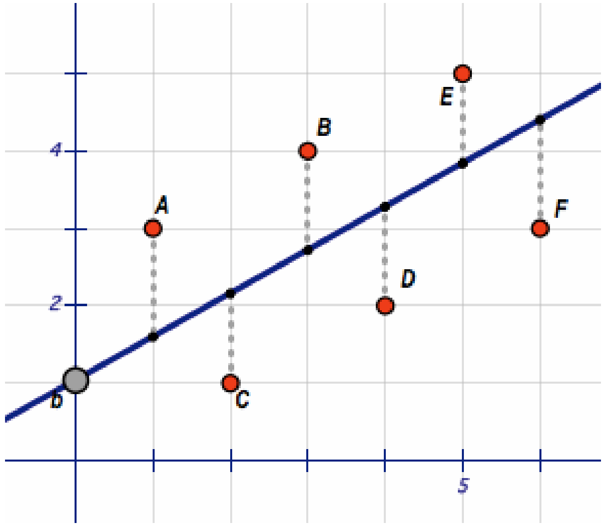


$$Y = a \times X + b + \varepsilon$$

with Y : variable to explain (continuous)
X : explanatory variable (continuous)
a : coefficient of explanatory variable X (slope of the regression)
b : intercept (value of Y when X = 0)
ε : model residuals (proportion of the variability of Y not explained)

UNIVARIATE LINEAR MODELS
LEAST SQUARES

- **Adjustment method : least square algorithm:** plays on parameters a and b in the equation in order to minimize the sum of squares (sum of ε^2) between real values Y and the regression line (estimated values $\hat{Y}$)



- **Equation :**

$$Y = a \times X + b + \varepsilon$$

with
Y : variable to explain (continuous)
X : explanatory variable (continuous)
a : coefficient of explanatory variable X (slope of the regression)
b : intercept (value of Y when X = 0)
ε : model residuals (proportion of the variability of Y not explained)

UNIVARIATE LINEAR MODELS
RESIDUALS

- Residuals embody the **proportion of the variability** of Y not explained by X
- The more **residuals are important**, the least **the model fits the data**
- **Normality of residuals** are a **strong fact** to check !
- In case of **non-normally distributed residuals** :
  - ➢ **Significance of estimates** of the model can be wrong because its calculation assumes normality of data.
  - ➢ Check data quality (residuals) or use **multivariate linear models** or no**n-linear models** is required

UNIVARIATE LINEAR MODELS
COEFFICIENT OF DETERMINATION (R^2)

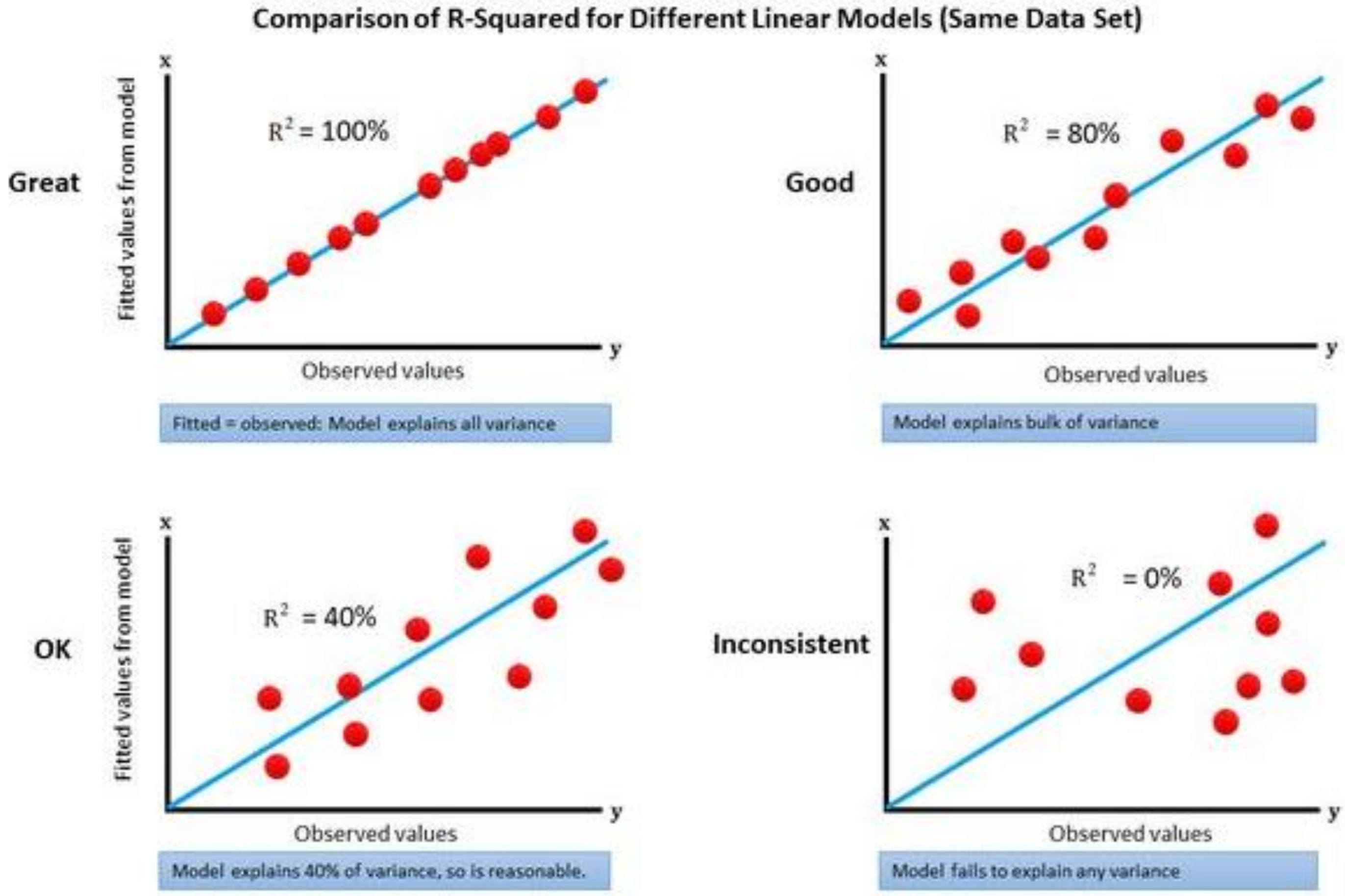
- **Quality model metric** : Coefficient of Determination R^2

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

with y_i: real value for observation $\boldsymbol{i}$
$\hat{y}_i$: predicted value by the model for observation $\boldsymbol{i}$
$\bar{y}$: mean of real values

- R^2 = square of **linear coefficient of correlation**
- Allows to determine the proportion of variability of Y explained by the model (in %)
- **Warning** : a good R^2 does not mean the model has a good **predictive accuracy**!

UNIVARIATE LINEAR MODELS
COEFFICIENT OF DETERMINATION (R^2)



UNIVARIATE LINEAR MODELS
OTHER QUALITY METRICS

- Mean absolute error (MAE) : $\text{MAE} = \frac{1}{N}\sum_1^N |y_i - \hat{y}_i|$

- Mean square error (MSE) : $\text{MSE} = \frac{1}{N}\sum_1^N (y_i - \hat{y}_i)^2$

- Root Mean square error (RMSE) : $\text{RMSE} = \sqrt{MSE}$

- In R : functions *MAE*, *MSE* and *RMSE* available in package *caret*

UNIVARIATE LINEAR MODELS
MODELING WITH R

lm function (*stats* package)

Parameters :
formula = $Y \sim X$ with Y and X are continuous
data = dataset (number of points : N)
subset = train model only on a subset of the dataset
weight = optional vector with the weights of points
na.action = handling of NA values

Results : *lm* object with the following attributes :
coefficients = a (slope) and b (intercept)
residuals = vector of N points with residuals of model
fitted.values = vector of N points with values estimated

UNIVARIATE LINEAR MODELS
MODELING WITH R

- *summary* function : print in the console :
  - The model
  - Descriptive statistics of residuals (minimum, Q1, median, Q3, maximum)
  - Coefficients of the model (estimates, standard-error, statistic and pvalue
  - Quality of model : R^2 and adjusted R^2 (R^2 penalized with the number of predictors)

- *anova* function (*stats* package) : print the **analysis of variance table** (sum of squares)

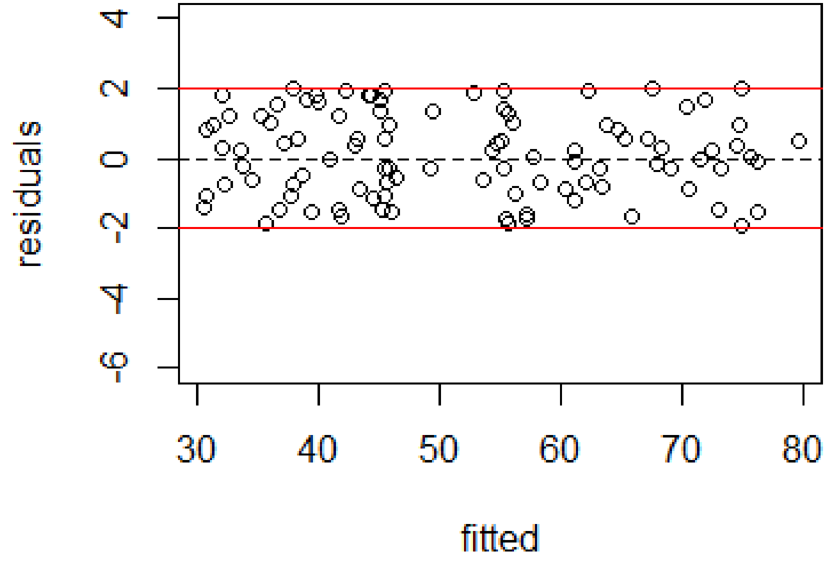
- *predict* : applies the equation on data

- *plot* : displays the residual quality plots (6 plots)

UNIVARIATE LINEAR MODELS
MODELING WITH R

*After launching a **lm** model: assess the **quality of the model** with **residuals checking** with **plot** function:*

1. Residuals vs Fitted values :
2. Residual Q-Q plot
3. Scale-Location
4. Cook's distance : outliers detection

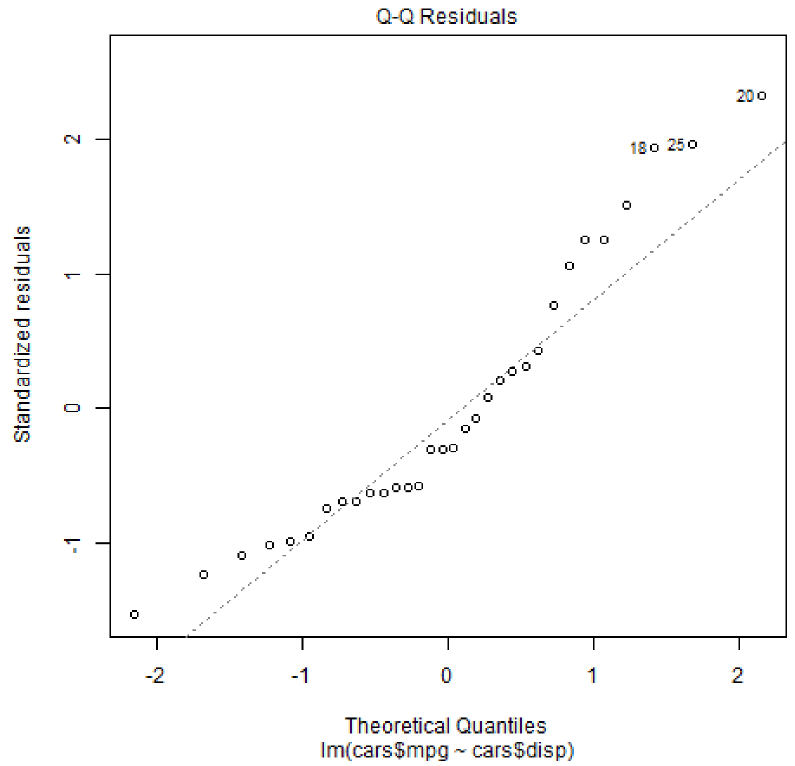


UNIVARIATE LINEAR MODELS

MODELING WITH R

*After launching a **lm** model: assess the **quality of the model** with **residuals checking** with **plot** function:*

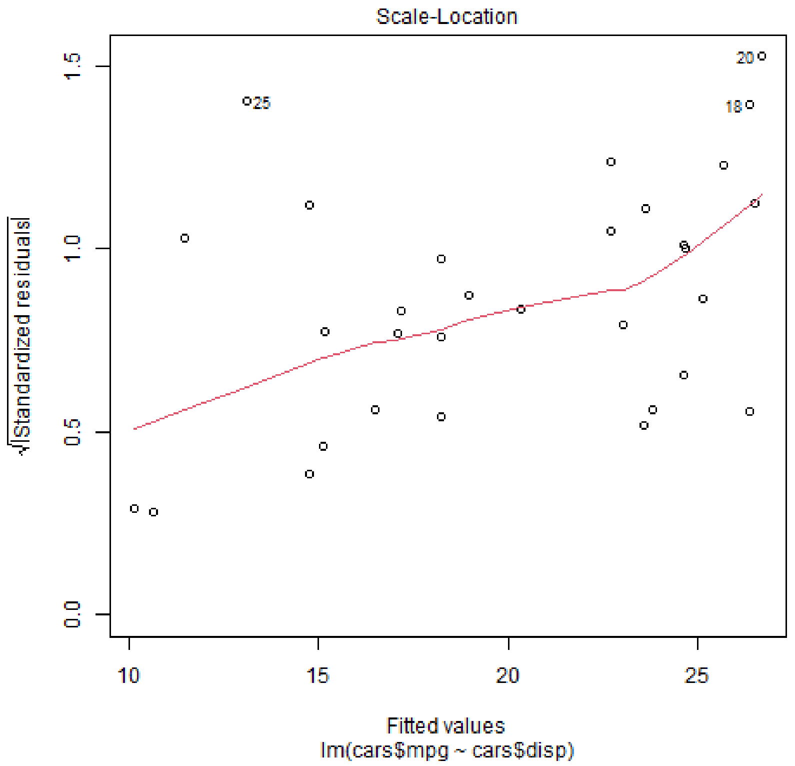
1. Residuals vs Fitted values :
2. **Residual Q-Q plot**
3. Scale-Location
4. Cook's distance : outliers detection



UNIVARIATE LINEAR MODELS
MODELING WITH R

*After launching a **lm** model: assess the **quality of the model** with **residuals checking** with **plot** function:*

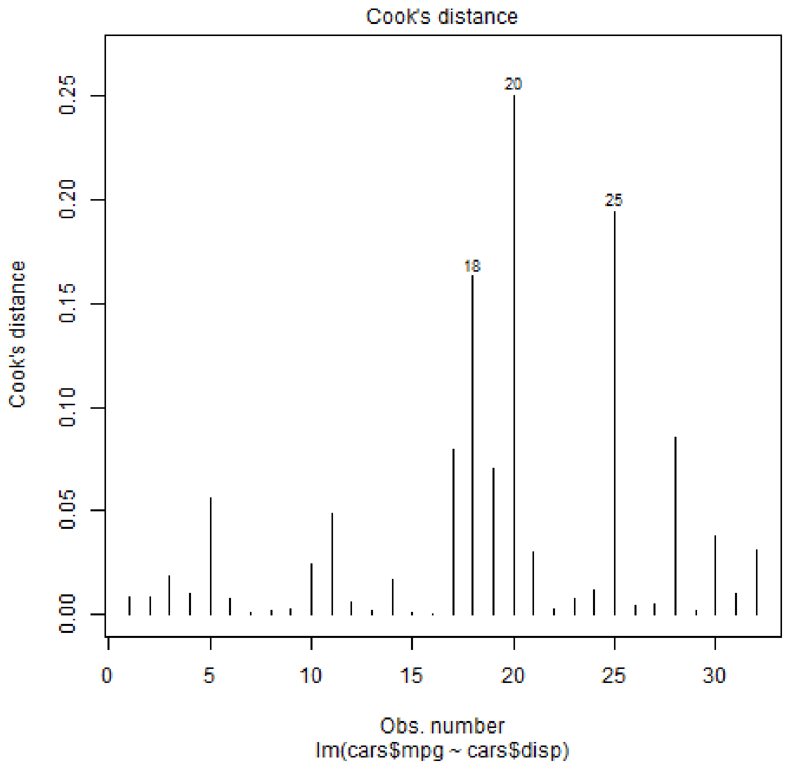
1. Residuals vs Fitted values :
2. Residual Q-Q plot
3. **Scale-Location**
4. Cook's distance : outliers detection



UNIVARIATE LINEAR MODELS
MODELING WITH R

*After launching a **lm** model: assess the **quality of the model** with **residuals checking** with **plot** function:*

1. Residuals vs Fitted values :
2. Residual Q-Q plot
3. Scale-Location
4. Cook's distance : outliers detection



UNIVARIATE LINEAR MODELS

MODELING WITH R

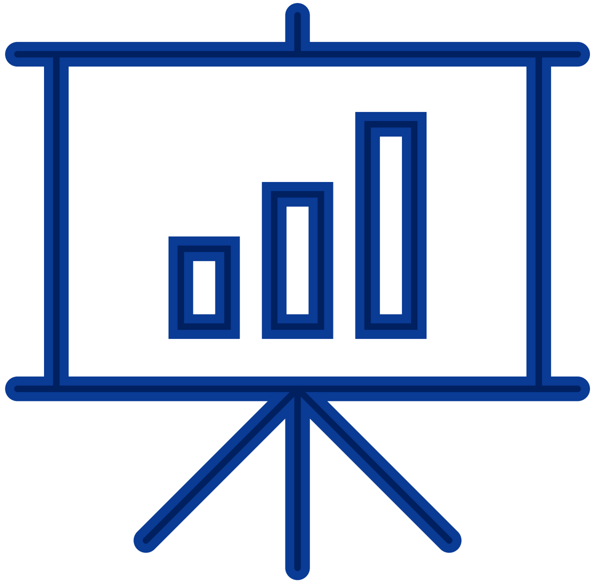
*After launching a **lm** model: assess the **quality of the model** with **residuals checking** with **normality test** :*

1. Plot **histogram** of residuals (use ***residuals*** function or ***model$residuals*** to access to them

2. Plot **Q-Q plot** of residuals with ***ggqqplot*** function (package ***ggpubr***)

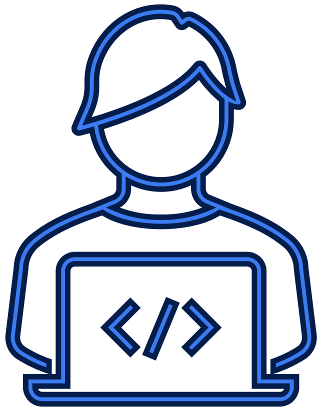
3. Launch a Shapiro-Wilks normality test ***shapiro.test*** (***stats*** package)

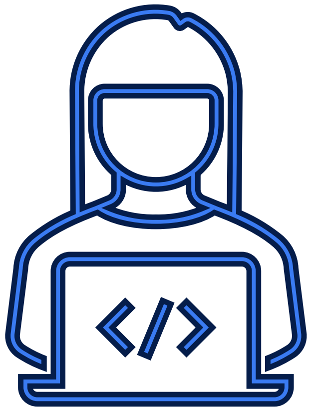
UNIVARIATE LINEAR MODELS



Live demo

UNIVARIATE LINEAR MODELS





Time to play !
(20 minutes)

MULTIVARIATE LINEAR MODELS

05

MULTIVARIATE LINEAR MODELS
HIGHLIGHTS

- **Goal** : explain a continuous variable Y with many continuous variables X_i

- **Hypothesis to test** : the relationship between Y and the X_i variables is linear

- **Equation** : $Y = a_1 \times X_1 + a_2 \times X_2 + \cdots + a_n \times X_n + a_{1,2} \times X_1 X_2 + b + \varepsilon$

  with
  Y : variable to explain (continuous)
  X : explanatory variables (continuous)
  a_i : coefficient of explanatory variable X_i
  $a_{1,2}$: coefficient of interaction between X_1 and X_2
  b : intercept (value of Y when all X_i = 0)
  ε : model residuals (proportion of the variability of Y not explained)

MULTIVARIATE LINEAR MODELS
MODEL QUALITY

- **Adjustment method** : Least Squares method
- **Quality Model metric** : R^2
- Several R^2 in R :
  - **Classical R^2**
  - **Adjusted R^2** : allows to compare the R^2 of different models with different numbers of parameters
  - **Predictive R^2 (also called Q^2)** : ability of the model to predict new values (calculated with PRESS function with cross-validation)

Warning : the more we add parameters the more we increase R^2 (and **overfitting** issue)

MULTIVARIATE LINEAR MODELS
VARIABLE SELECTION

- Several methods of explanatory variables selection :

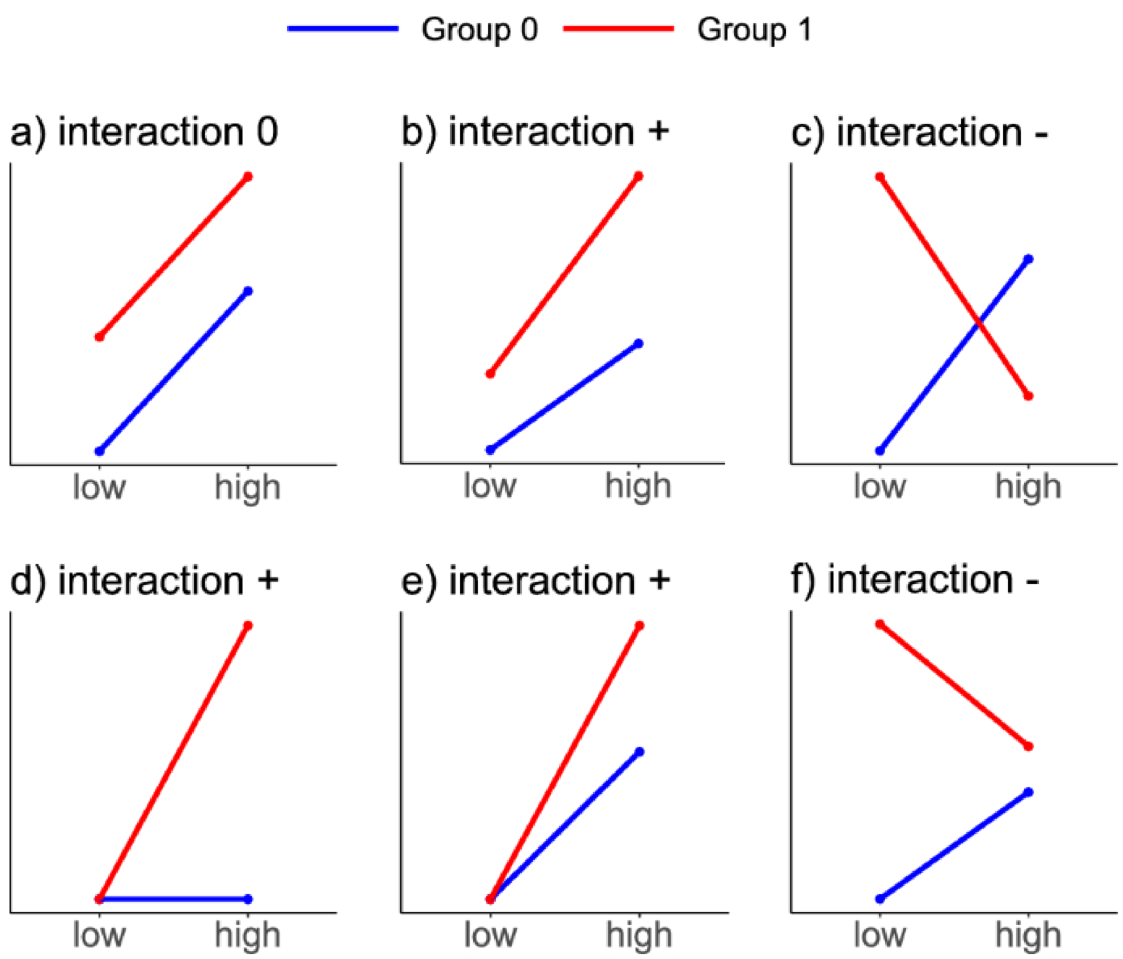
  - Backward method : from the complete model we iteratively remove the least significant variable (based on the pvalue)
  - Forward method : from the null model (containing only the mean of Y we iteratively test each variable add the most significant one (based on the pvalue)
  - Stepwise method : mix of the two other methods

The algorithm stops when it can not find another variable to include in the model (based of a pvalue threshold : 15% or 0.15 by default)

In R : *stepAIC* function included in package *MASS*

MULTIVARIATE LINEAR MODELS
EQUATION TERMS

- Terms in the equation :
  - ➢ **Main effects** : effect of a variable X on Y
  - ➢ **Interactions** : simultaneous effect of variables X_1 and X_2 on Y



MULTIVARIATE LINEAR MODELS

MODELING WITH R

lm function (*stats* package)

Parameters : *formula* = $Y \sim X_1 + X_2 + \cdots + X_n + X_1 \times X_2 \cdots X_{n-1} \times X_n$

with Y and X_i are continuous

data = dataset (number of points : N)

subset = train model only on a subset of the dataset

weight = optional vector with the weights of points

na.action = handling of NA values

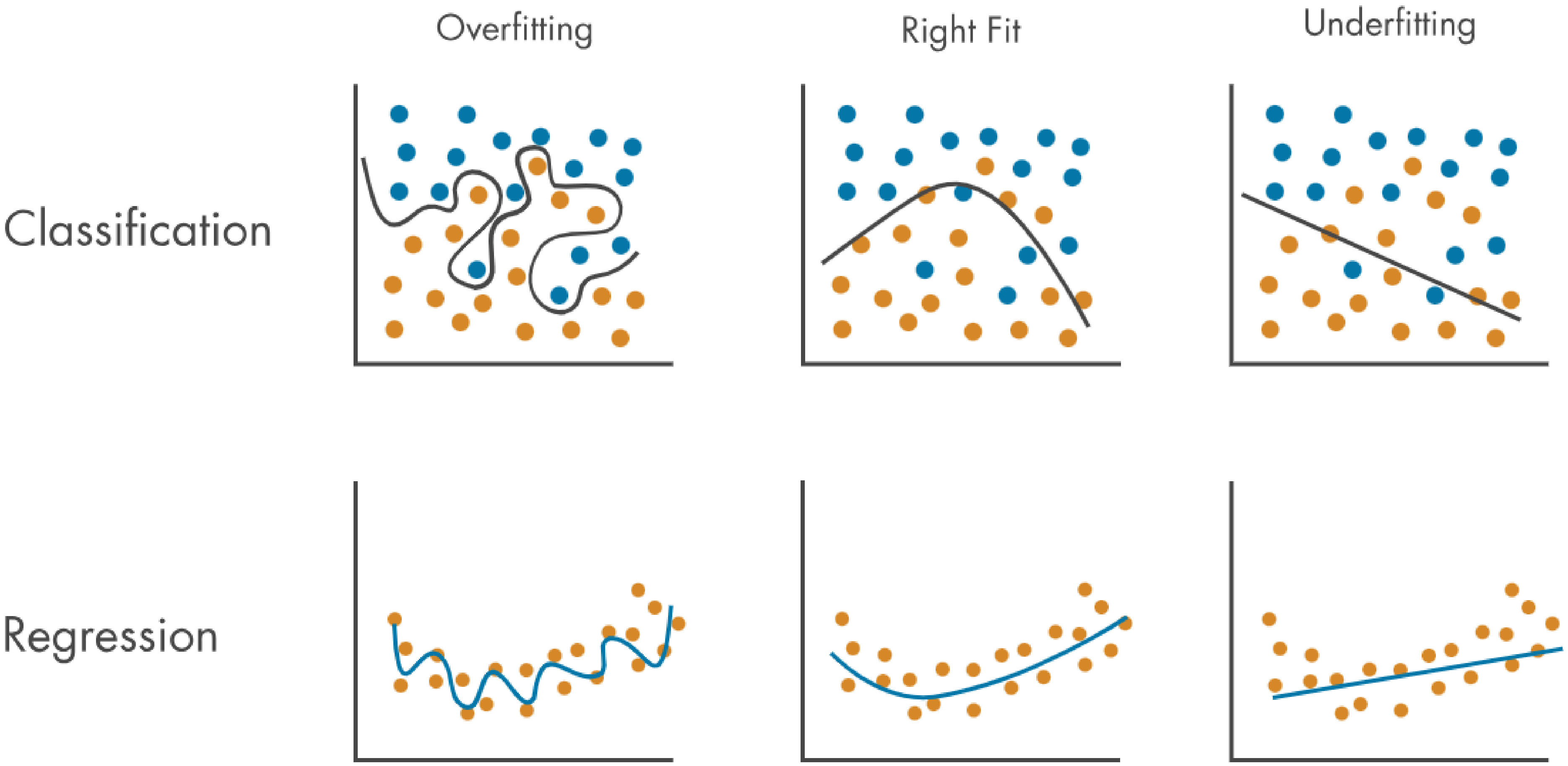
Results : *lm* object with the following attributes :

coefficients = $a_1, \cdots, a_n, a_{1,2}, \cdots, a_{n-1,n}$ and b (intercept)

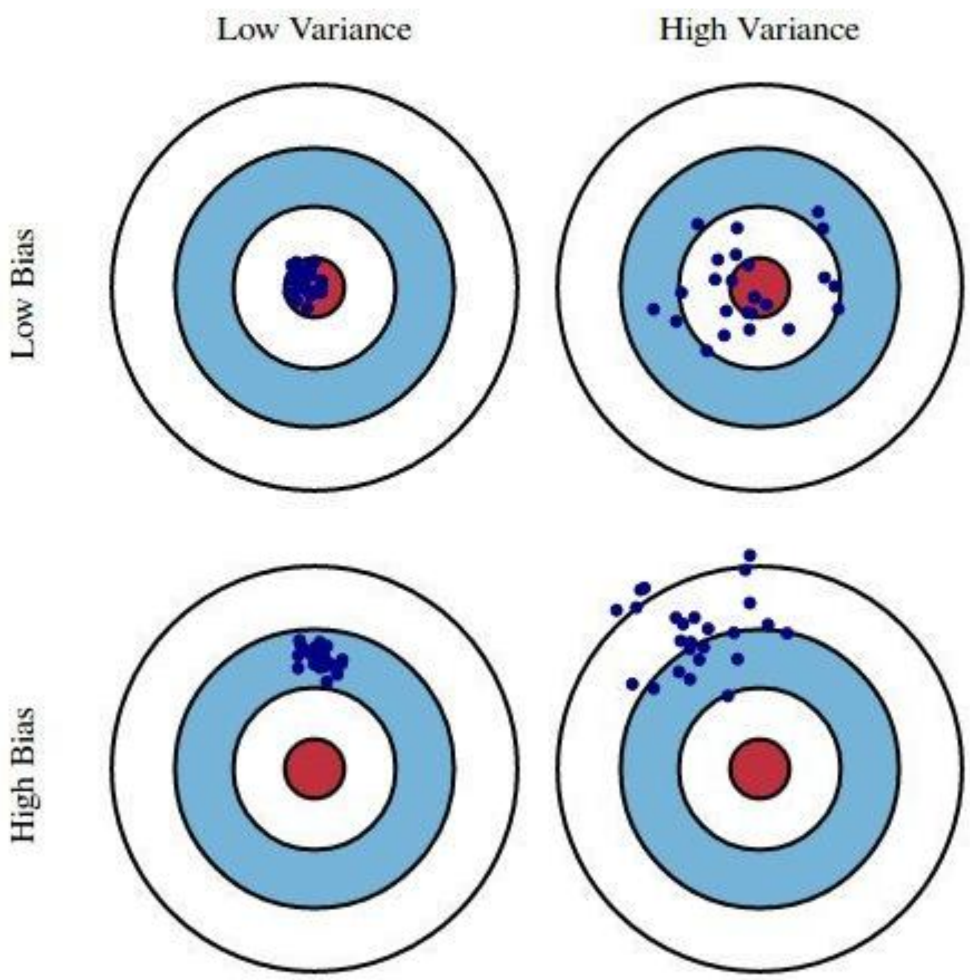
residuals = vector of N points with residuals of model

fitted.values = vector of N points with values estimated

MULTIVARIATE LINEAR MODELS
OVERFITTING ISSUE



MULTIVARIATE LINEAR MODELS
OVERFITTING ISSUE

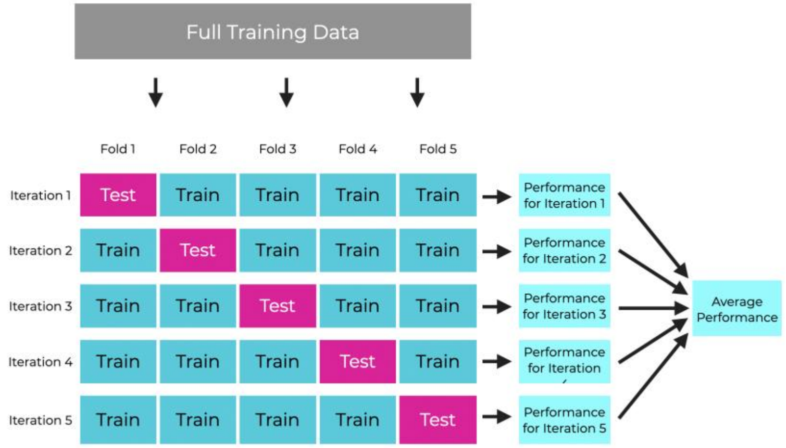


MULTIVARIATE LINEAR MODELS
CROSS VALIDATION

- Statistical procedure often used in Machine Learning which allows to reduce the **overfitting issue**, very harmful during the development of **predictive models**.

- Idea : a model is **trained** on a representative subset of data and **tested** on the remaining data (not seen) several times.

- **Predictive R^2** is calculated with cross-validation : the model is built on all data except one and tested on this datapoint. This step is done N times (N=number of observations). This method is called LOOCV (**L**eave-**O**ne **O**ut **C**ross **V**alidation)



- In R many techniques available in *caret* package

MULTIVARIATE LINEAR MODELS
MULTICOLINEARITY

- Statistical issue where two (or more) explanatory variables are **strongly correlated**

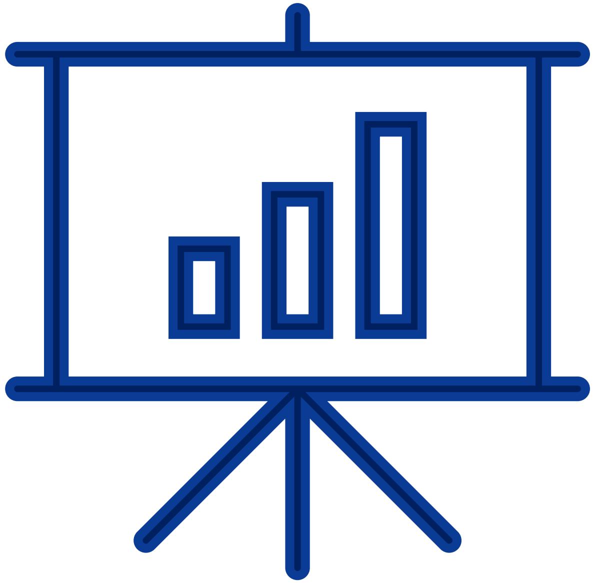
- Including such variables in a model can rise **instability of parameters** of the equation and also cause a **non-convergence of variable selection algorithms**

- This high variability of parameters **does not impact predictive ability** of the model by can **strongly impact significance of factors** (pvalues).

- The **VIF** index (Variance Influence Factor) is useful for highlighting variables which are highly correlated. VIF is available in R with the ***car*** package with ***vif*** function.

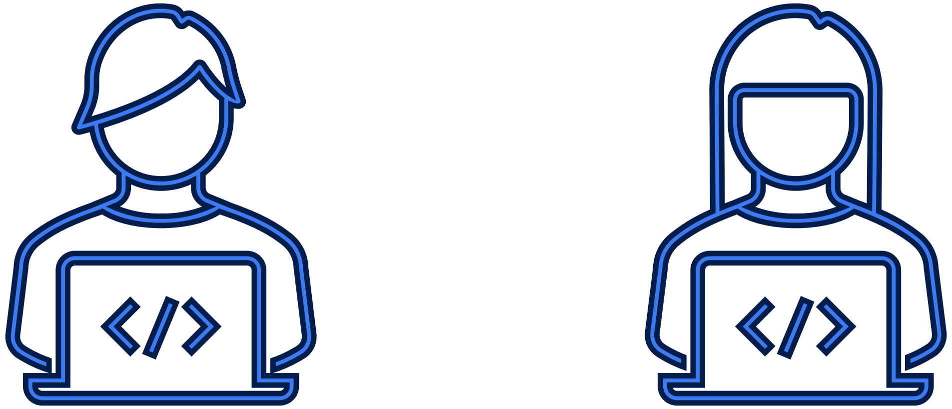
- **Remove a variable when the VIF is > 5.**

MULTIVARIATE LINEAR MODELS



Live demo

MULTIVARIATE LINEAR MODELS



Time to play !
(30 minutes)

QUESTIONS

06

THANK YOU FOR YOUR ATTENTION

SEPTEMBER 2025