# Elements of descriptive and inferential statistics for biology

François-Xavier Lejeune

*f-x.lejeune@icm-institute.org*

Year 2023-2024

# Course outline

- Introduction
- Part 1: Data description
- Part 2: Hypothesis testing
- Part 3: Data modeling

# Introduction

# Why do we need statistics in biology? (1/3)

Statistical methods are necessary to quantify and account for the **variability inherent in biological or biomedical data**, which typically affects the measurements of a variable of interest.

This variability mainly comes from two sources that must be distinguished:

- **Biological variability** due to subjects, organisms, and biological samples => ***biological variation of interest***
- **Technical variability** due to measurement, instrumentation, reagent, variability (unconsciously) introduced by experimenters, sample preparation => ***non-biological interferences (irrelevant variation)***

# Why do we need statistics in biology? (2/3)

Both sources of variations can still be characterized as follows:

1. Biological variability [related to subjects or groups of subjects]

   - **Within variations** for the *measurements taken on the same subject (repeated measurements) or the same group of subjects*

   - **Between variations** for the *measurements taken from several subjects or several groups of subjects* (*e.g.*, differences in age, sex, education, pathology, treatment, biological constants or genetic characteristics...)

2. Metrological variability [related to the measurement protocol]

   - Variations of the **experimental conditions** (*e.g.*, differences in temperature, humidity, luminosity, several people involved in the data collection…)

   - Errors induced by the **measuring device** due, e.g., to mechanical vibrations, calibration issues, equipment aging…

# Why do we need statistics in biology? (3/3)

## To manage the large amount of biological data:

Methods for the analysis and integration of large and heterogeneous data:

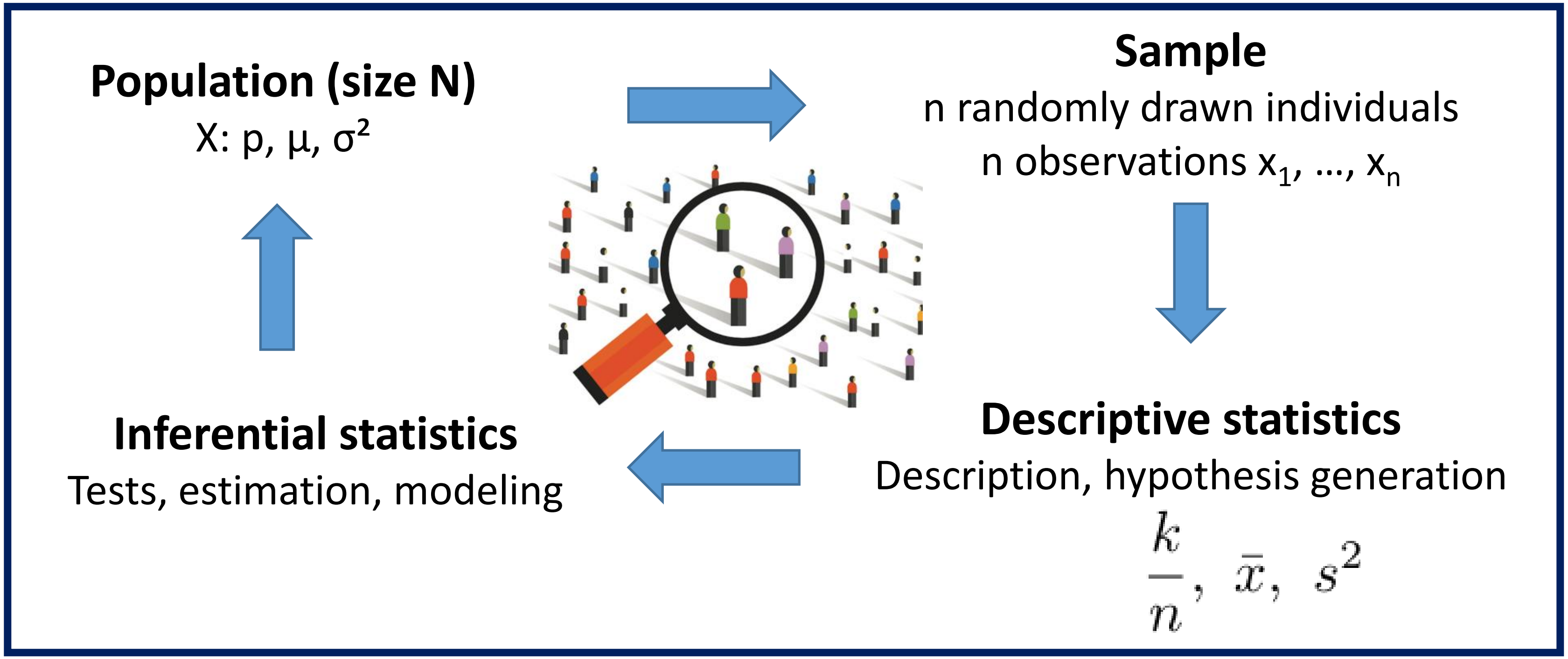
- Clinical, Brain imaging (PET Scan, MRI) data,
- Behavioral data, Electrodermal activity (EDA) recordings,
- Electrophysiological recordings (MEG, EEG, MEA),
- Omics (genomics, transcriptomics, proteomics, metabolomics…),
- Histology (microscopic study of cells and tissues)…

☞ *methods for feature selection and dimensionality reduction…*

# Basic statistics vocabulary

- **Population**: group of individuals or statistical units targeted by the study (size N)
- **Sample**: subset of the population used for the study (size n << N)
- **Variable X**: common trait of individuals that can be observed or measured

**Statistical experiment**: typically involves a randomly drawn sample (***"randomization"*** step) of sufficient size to represent the study population for investigating a phenomenon or testing a hypothesis.



**Population (size N)**

X: p, μ, σ²

**Sample**

n randomly drawn individuals

n observations $x_1, \ldots, x_n$

**Descriptive statistics**

Description, hypothesis generation

$$\frac{k}{n},\ \bar{x},\ s^2$$

**Inferential statistics**

Tests, estimation, modeling

# Choosing the appropriate statistical analysis?

The choice of the statistical methods is essentially guided by

1. **"What are we looking for in the data"**
   ***[☞ biological question or study objective(s)]***

- ***Exploratory approach*** to gain new insights and generate novel hypotheses for further investigation…
- ***Inferential approach*** to test an a priori hypothesis or construct a "predictive model"…

2. **Data characteristics**

- Types and distributions of variables (Quantitative and/or categorical)
- Experimental design: *e.g.* sample sizes available per condition
- Presence of missing values and/or extreme values (***"outliers"***)
- Repeated measurements on the same individual, possibly at several points in space and/or time (***longitudinal*** study)

# Quantitative variables

The variable is quantitative if its values correspond to measurable quantities given by numbers.

**Discrete data = integer values in a countable set**

Examples (counting measures):

- number of relapses in MS patients,
- number of cells per unit area,
- number of mutations in a 10 kb DNA sequence,
- number of words recalled in a memory test…

**Continuous data = infinity of values in a real interval**

Examples:

- weight, height, body mass index, age,
- daily dose of levodopa in a parkinsonian patient,
- blood glucose testing,
- volume of a brain region in brain imaging…

# Categorical variables

The variable is qualitative or categorical (factor) if its values are not quantities measured by numbers, but define a group of categories called modalities or levels.

**Nominal variable = categories without any kind of natural order**

Examples:

- Sex: male/female,
- Smoking status coded as ‘yes’/’no’,
- Blood group: A / B / AB / O…

**Ordinal variable = categories can be ordered**

Examples:

- frequency of an activity: never, rarely, sometimes, often, very often,
- pain severity: none, minimal, moderate, severe, unbearable,
- Alzheimer’s Disease progression: pre-symptomatic stage, mild cognitive impairment, mild AD, moderate AD, severe AD…

# Part 1: Data description

***Summarize and represent graphically the information contained in the data***

# Unidimensional descriptive analysis

Usually, the distribution of a variable is described using

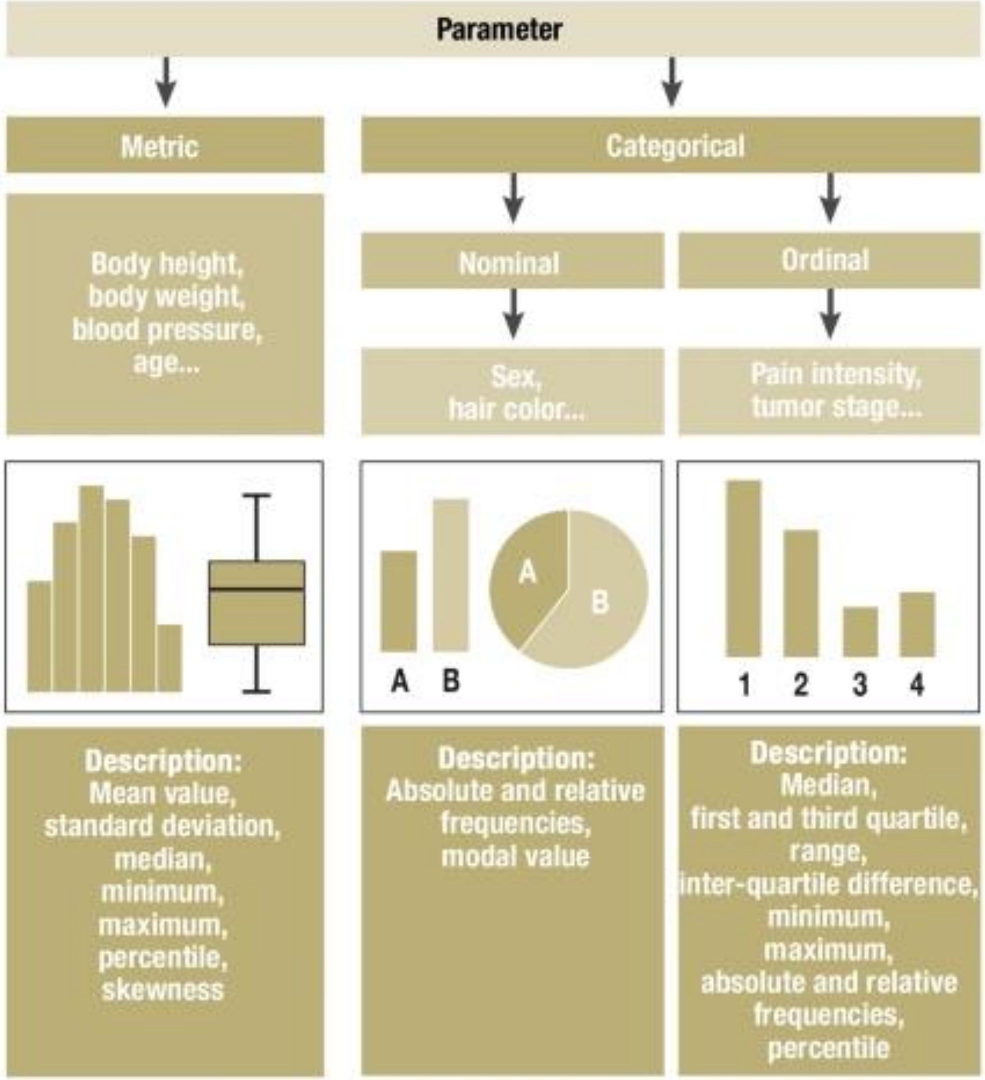
❑ **3 numerical criteria**

- Central tendency
- Dispersion
- Shape of data (skewness + kurtosis)

❑ **1 frequency graph**

Data presentation:

- **Mean ± Standard Deviation**
- **Median with Interquartile Range**
- **Counts and percentages**



Parameter

Metric

Body height, body weight, blood pressure, age...

Description: Mean value, standard deviation, median, minimum, maximum, percentile, skewness

Categorical

Nominal

Sex, hair color...

Description: Absolute and relative frequencies, modal value

Ordinal

Pain intensity, tumor stage...

Description: Median, first and third quartile, range, inter-quartile difference, minimum, maximum, absolute and relative frequencies, percentile

*Figure taken from du Prel, Röhrig, Blettner, Dtsch Arztebl Int 2009*

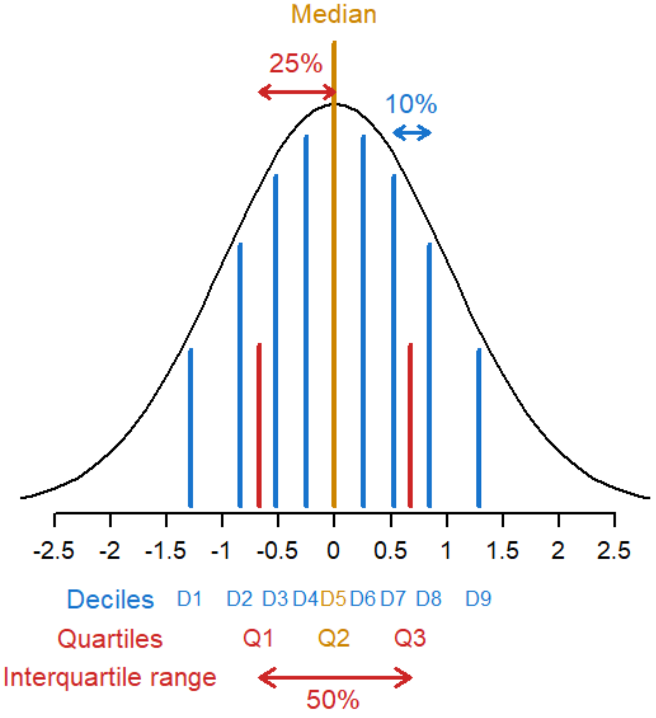
# Notion of quantiles

The notion of **empirical quantile** applies to the ordered values of a quantitative variable.

The quantile of order α ($0 \leq \alpha \leq 1$) then refers to the value $q_\alpha$ of the variable such that a proportion α of the values in the population is less than or equal to $q_\alpha$.

Usual quantiles:

- **Median:** α = 50%
- **Quartiles:** α = 25%, 50%, 75% (Q1, Q2, Q3)
- **Deciles:** α = 10%, 20%, …, 90% (D1, D2, …, D9)
- **Percentiles:** α = 1%, 2%, …, 99% (C1, C2, …, C99)



*Main quantiles of the Normal distribution N(0,1)*

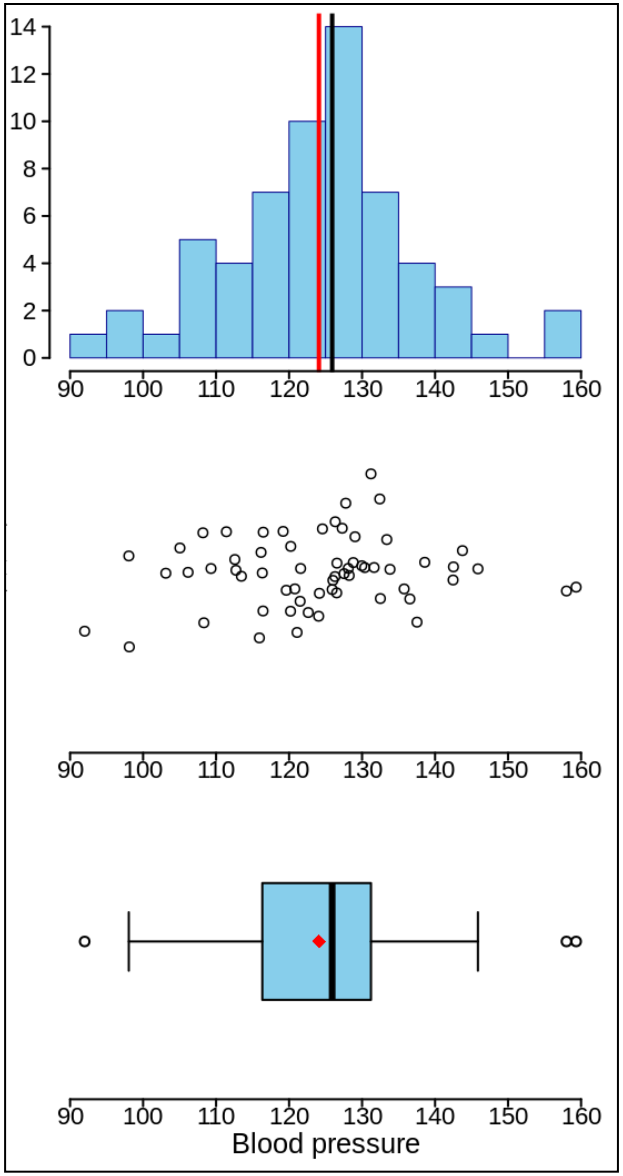
# Numerical summary of a distribution

**Central tendency** = *value around which the observations are distributed*

- Empirical average: $$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$
- Mode
- Median $Q_2$

**Dispersion** = *observations spread around the central tendency*

- Estimated variance: $$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$
- Standard deviation (SD): $$s = \sqrt{s^2}$$
- Range: Max – Min
- Interquartile range: IR = Q3 – Q1
- Coefficient of variation: $$\mathrm{CV} = \frac{s}{\bar{x}}$$



# Standard Deviation vs Standard Error of the Mean

Among the dispersion values, the **standard error of the mean (SEM)** is a criterion often used (sometimes incorrectly) instead of the standard deviation:

$$\mathrm{SEM} = \frac{\mathrm{SD}}{\sqrt{n}}$$

It is therefore important to make the following distinction:

The **SD** of a sample is used to indicate the variability of the observed values within the sample or population (*e.g.*, ages within a group of patients).

Unlike the SD, the **SEM** does NOT reflect the variability of the sample but how much the estimated mean varies among samples as if the study was repeated on several samples of size **n** (standard deviation of the mean...).

☞ **Whatever the statistic used SEM or SD, it must ALWAYS be indicated in a study!!!**

# Probabilistic models

**Probability theory** and **statistics** are generally distinguished as follows:

Probability: aims at defining *mathematical models* (or *theoretical laws of distribution*) governing chance and at studying their properties

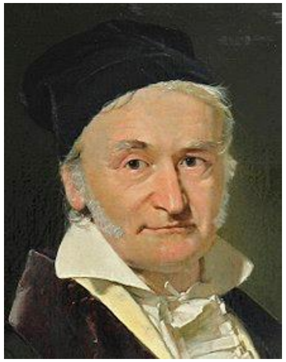
Statistics: aims to compare these theoretical models with real data

Many so-called “parametric” approaches are based on the assumption that the observed data are realizations of random variables whose distribution law is known. In this case, the analysis will consist of selecting, adjusting and validating probabilistic models that can be used to test hypotheses, predict or guide decision-making.

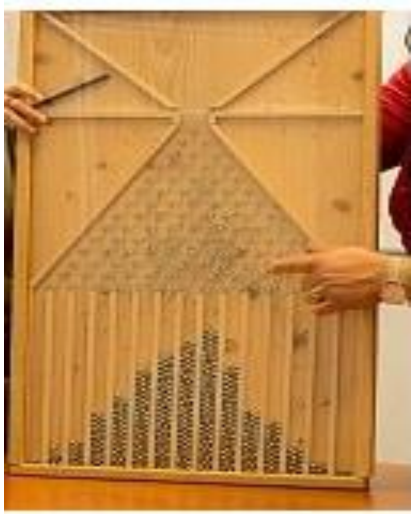
To do this, we have several laws of continuous or discrete distributions commonly used in practice: **uniform**, **normal**, **exponential**, **binomial**, **Poisson**, *etc.*

# Normal distribution (Laplace-Gauss)

Among the probability laws, the **Gaussian distribution** characterized by its “bell-shaped” curve is frequently used in practice because it allows for modeling the variability of many natural phenomena (fasting blood glucose, bacterial division rate, etc.) and the distribution of measurement errors.

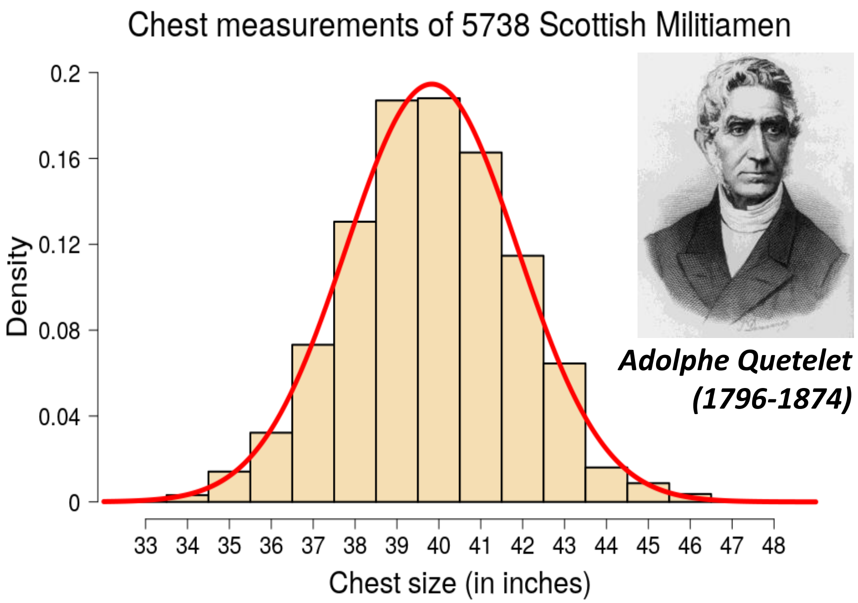


*Carl Friedrich Gauss (1777-1855)*

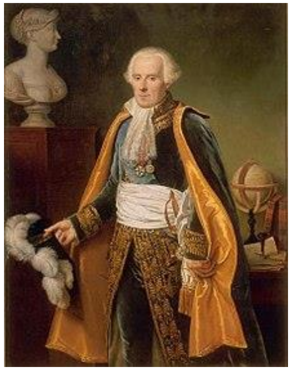


*Example 1. Galton board (flow of balls through a pyramid of nails)*

*Example 2. Data on chest measurements of 5738 Scottish Militiamen (Quetelet 1846)*



*Adolphe Quetelet (1796-1874)*



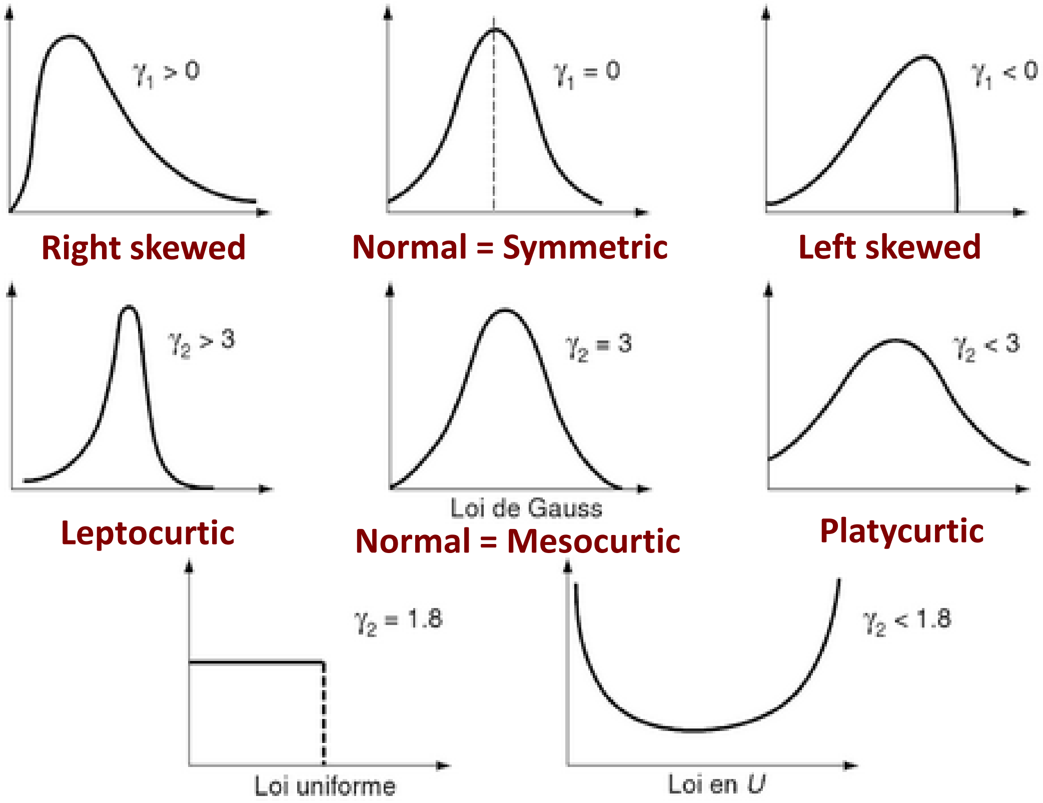
*Pierre-Simon de Laplace (1749-1827)*

***Data: https://www.stat.cmu.edu/StatDat/Datafiles/MilitiamenChests.html***

# Describe the shape of a distribution

The bell density of the Gaussian distribution is also used as a reference to characterize the shape of other continuous probability distributions on the following 2 criteria:

- **Measure of asymmetry (skewness)**
- **Measure of peakedness (kurtosis)**



**Right skewed** **Normal = Symmetric** **Left skewed**

**Leptocurtic** **Normal = Mesocurtic** **Platycurtic**

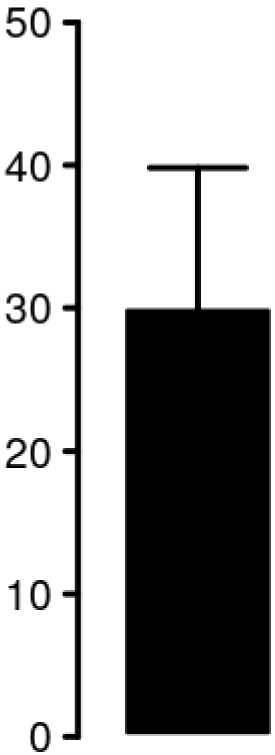
$$\gamma_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^{3/2}}$$

$$\gamma_2 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^{2}}$$

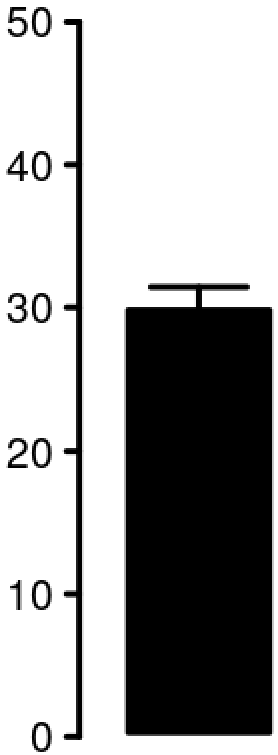
***Figure taken from G. Saporta, Probabilités, analyse des données et statistique, Technip***

# Usual graphical representations

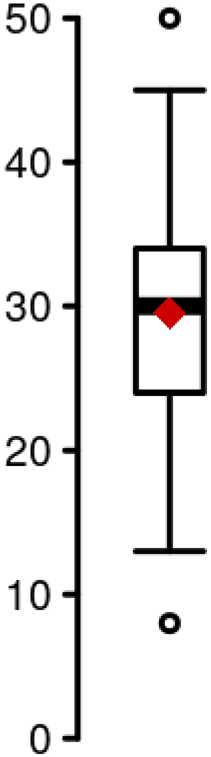
5 possible representations of the distribution of UPDRS-OFF score values of 29 parkinsonian patients: **which one is the most informative?**



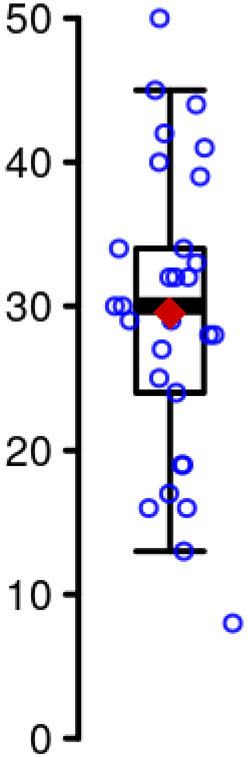
**Plot Mean + SD**



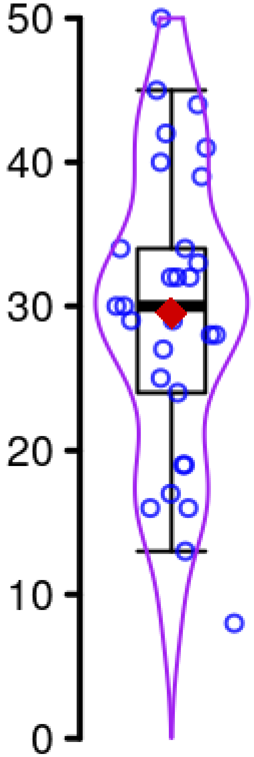
**Plot Mean + SEM**



**Boxplot + mean value**



**Boxplot + mean value + sample values**



**Violin plot + Boxplot + mean value + sample values**

***Source: Nucleipark project***

**UPDRS = Unified Parkinson's Disease Rating *Scale: Parkinson’s disease progression assessment scale; OFF = off treatment***

# 95% confidence intervals

Interval with a 95% probability of containing the parameter of interest:

- **Numerical parameter**
  if X follows a normal distribution:

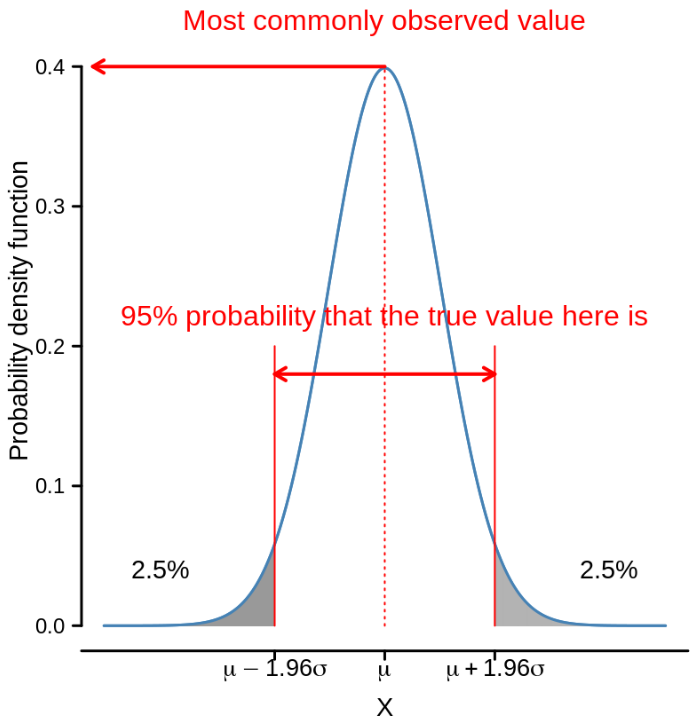
$$\boxed{\bar{x} \pm 1.96 \times \mathrm{SD}}$$

- **Mean**
  whatever the distribution of X, if n > 30:

$$\boxed{\bar{x} \pm \frac{1.96 \times \mathrm{SD}}{\sqrt{n}} \quad \text{or} \quad \bar{x} \pm 1.96 \times \mathrm{SEM}}$$

- **Frequency**
  if n×p, n×(1-p) > 10:

$$\boxed{p \pm 1.96 \times \sqrt{\frac{p\,(1-p)}{n}}}$$



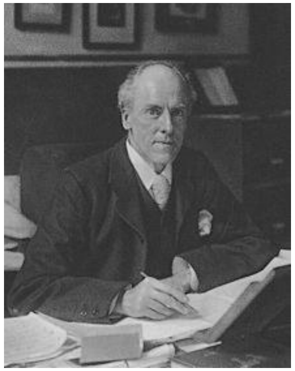
# Two-dimensional descriptive analysis

Bivariate analysis is the ***study of the relationship between two variables X and Y observed on the same sample of individuals***.

Usual methods (numerical indicator + graphic):

- **2 quantitative variables**: correlation + scatterplot with a regression line
- **2 categorical variables:** contingency table + bubble plot, mosaic plot or bivariate barplot
- **1 quantitative variable with 1 categorical variable:** correlation ratio + boxplot

**/!\ Bivariate analysis focuses on the simultaneous variation of two variables, but it does not allow for establishing causality in the relationship!**

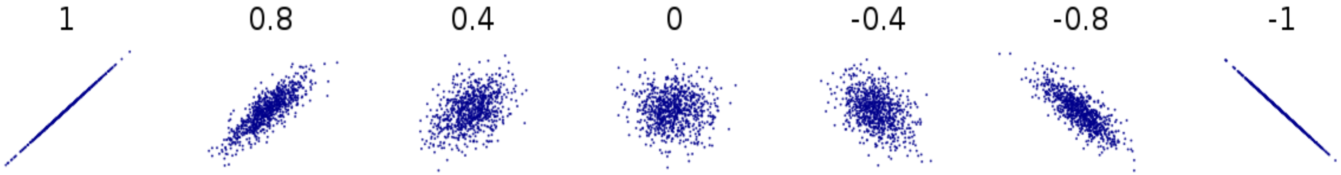


*Karl Pearson (1857-1936) at the origin of statistics applied to biomedicine*

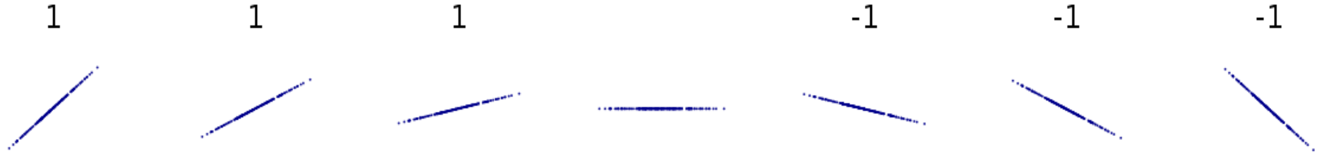
# Correlation and scatterplots

**Application:** The $r_{XY}$ **coefficient of (Bravais-) Pearson correlation** measures the linear relationship between **2 quantitative variables**.

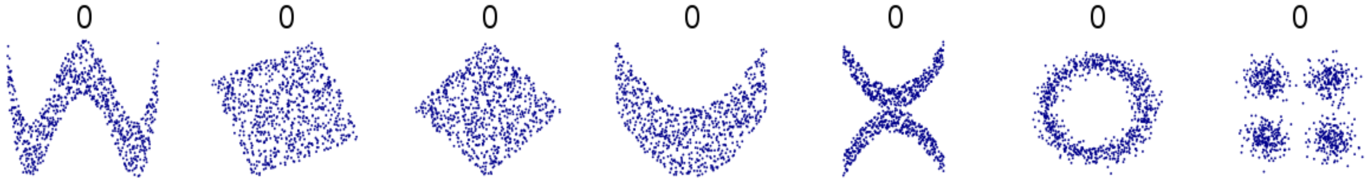
$$\mathrm{r}_{XY} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{N}(x_i-\bar{x})^2} \times \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2}} = \frac{\widehat{\mathrm{Cov}}(X,Y)}{\hat{\sigma}_X \cdot \hat{\sigma}_Y} \quad (-1 \leq \mathrm{r}_{XY} \leq 1)$$



☞ *|$r_{XY}$| = 1 with more or less important linear regression slopes*



☞ *$r_{XY}$ = 0 does not necessarily imply the absence of a (non-linear) link between X and Y*



**2 quantitative variables X and Y**

***http://guessthecorrelation.com/***

# Coefficient ρ of Spearman's rank correlation

*Charles Spearman (1863-1945)*

**Application:** Spearman's rank correlation coefficient ρ is applied to establish:

- the relationship between **2 ordinal qualitative variables**
- a monotonic non-linear relationship between **2 quantitative variables**

**Principle:**

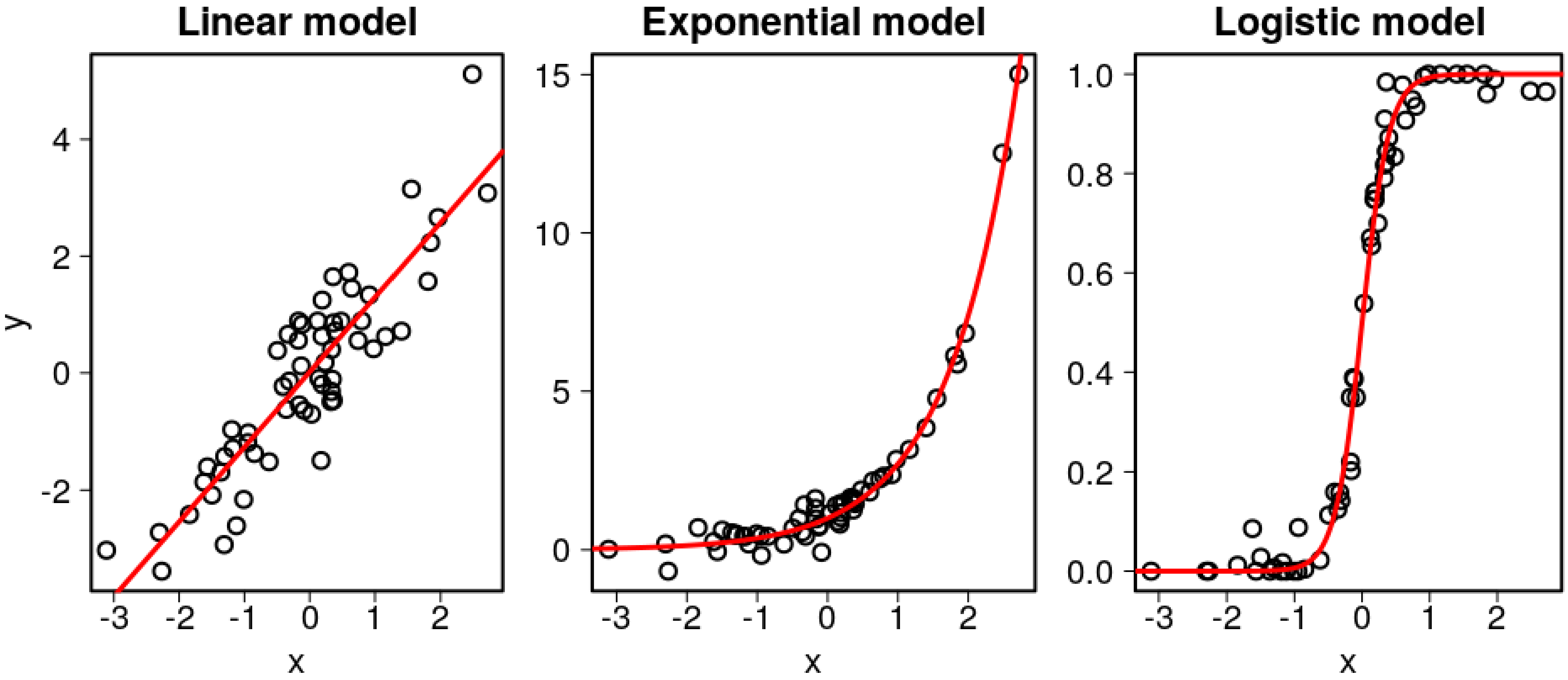
1. the ordered values (or levels) of the variables X and Y are replaced by the ranks noted $x_i$ et $y_i$
2. the Spearman coefficient is then given by the following formula:

$$\boxed{\rho = 1 - \frac{6\sum_{i=1}^{N} d_i^2}{N(N^2-1)} \quad \text{with} \;\; d_i = x_i - y_i \quad (-1 \leq \rho \leq 1)}$$

- ✓ if $\rho$ is close to 0: no relationship between X and Y
- ✓ if $\rho$ is close to -1: strong negative relationship between X and Y
- ✓ if $\rho$ is close to 1: strong positive relationship between X and Y

# Pearson or Spearman?

The example below illustrates the values of Pearson and Spearman correlation coefficients for 3 types of monotonic relationships:



r = 0.897
ρ = 0.861

r = 0.767
ρ = 0.914

r = 0.87
ρ = 0.963

# Contingency table and χ² coefficient

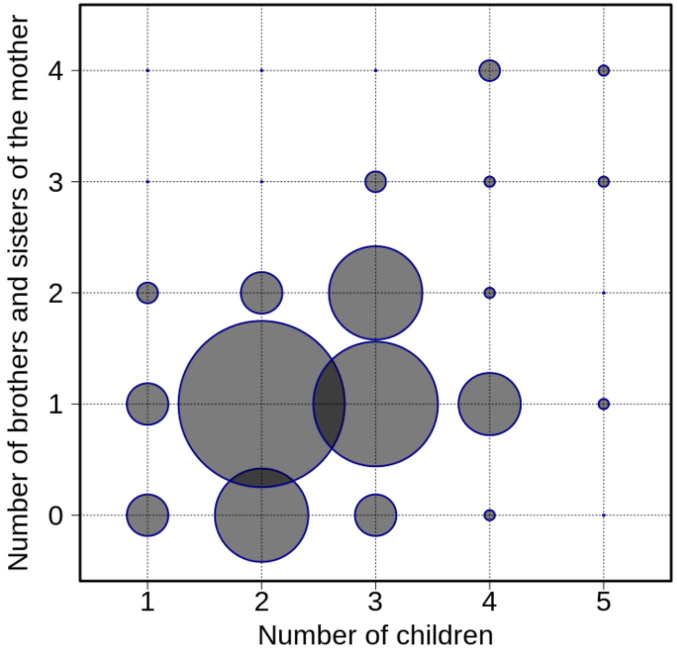
## Contingency table

| X \ Y | $y_1$ | $\cdots$ | $y_h$ | $\cdots$ | $y_c$ | rowsums |
|---|---|---|---|---|---|---|
| $x_1$ | $n_{11}$ | $\cdots$ | $n_{1h}$ | $\cdots$ | $n_{1c}$ | $n_{1+}$ |
| $\vdots$ | $\vdots$ |  | $\vdots$ |  | $\vdots$ | $\vdots$ |
| $x_\ell$ | $n_{\ell 1}$ | $\cdots$ | $n_{\ell h}$ | $\cdots$ | $n_{\ell c}$ | $n_{\ell +}$ |
| $\vdots$ | $\vdots$ |  | $\vdots$ |  | $\vdots$ | $\vdots$ |
| $x_r$ | $n_{r1}$ | $\cdots$ | $n_{rh}$ | $\cdots$ | $n_{rc}$ | $n_{r+}$ |
| colsums | $n_{+1}$ | $\cdots$ | $n_{+h}$ | $\cdots$ | $n_{+c}$ | $n$ |

**Marginal distributions of X and Y**

**Conditional distributions of X / Y = $y_h$ and Y / X = $x_l$**

## Bubble plot



The **coefficient χ²** measures the difference between the "observed" $n_{lh}$ and "theoretical" counts $n_{l+} * n_{+h}$ expected if X and Y are independent:

$$\chi^2 = \sum_{\ell=1}^{r} \sum_{h=1}^{c} \frac{\left(n_{\ell h} - \frac{n_{\ell +} n_{+h}}{n}\right)^2}{\frac{n_{\ell +} n_{+h}}{n}}$$

***"+ χ² is large***
***+ the relationship is strong***
***between X and Y"***

**2 categorical variables X and Y**

# Measures of association based on $\chi^2$

Based on the χ2 coefficient, **3 linkage measures** are useful **for assessing the strength of association between 2 categorical variables**:

- **Cramér's V**

$$\boxed{V = \sqrt{\frac{\chi^2}{n \times \min(r-1, c-1)}}, \qquad 0 \leq V \leq +1}$$

- **Contingency coefficient CC**

$$\boxed{CC = \sqrt{\frac{\chi^2}{\chi^2 + n}}}$$

- **Phi coefficient**

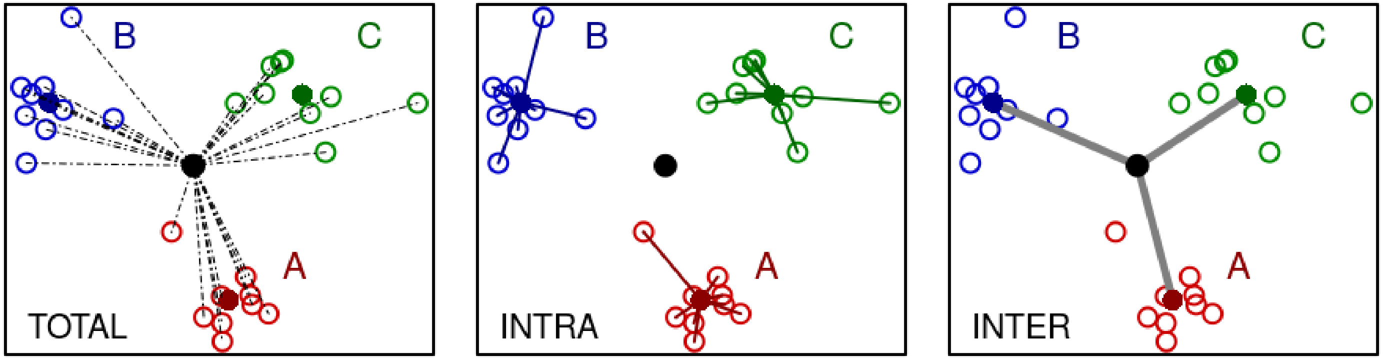
$$\boxed{\phi = \sqrt{\frac{\chi^2}{n}}}$$

**2 categorical variables X and Y**

# Correlation ratio

Let Y be a quantitative “dependent” variable observed among n individuals and G a factor indicating the grouping of individuals into K distinct classes.

***Example for K = 3 classes***



The observations of Y may differ due to two sources of variation:

- Variation within classes: **Intra-class variation**
- Variation between classes: **Inter-class variation** (factor effect)

We then use the **correlation ratio**:

$$\boxed{\eta^2_{Y/G} = \frac{\text{Inter-class variation}}{\text{Total Variation}}}$$

to measure the intensity of the effect of the factor G on the variable Y (value between **0 = no connection** and **1 = perfect connection** between Y and G).

**1 quantitative variable Y with 1 categorical variable G**

# Multidimensional descriptive analysis

In the case of a dataset with $p$ variables (**$p > 3$**), it becomes impossible to visualize the individuals in a **$p$-dimensional space**.

To reduce the dimension, the objective of **factorial methods** is to calculate a limited number of composite variables, called **latent variables** or **principal components**, which are constructed from the original variables in such a way as to summarize the data as well as possible. These methods facilitate the generation of graphical representations of the data (***individuals*** and ***variables***) utilizing these PCs as axes.

**Main factorial methods:**

- Principal Component Analysis (PCA, quantitative variables)
- Multiple Correspondence Analysis (MCA, categorical variables)
- Multiple Factor Analysis (MFA, variables may be numerical or categorical)

**Goals: summarize and visualize multidimensional data**

# Understanding dimension reduction

A common example of **dimension reduction** is the taking of photographs, which takes us from a 3-dimensional space (the one we live in) to a 2-dimensional space (our photo).

Of course, depending on the angle from which we consider our subject, all our photos will not bring the same level of information.

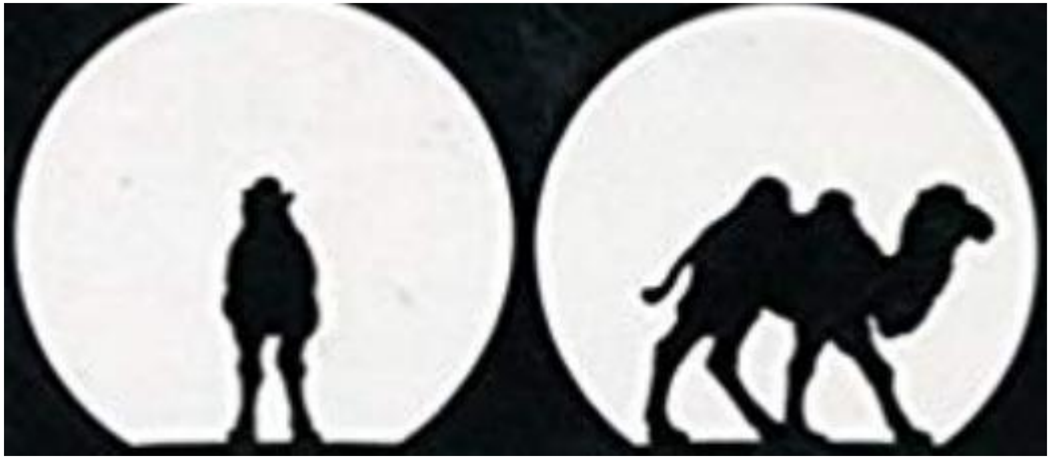
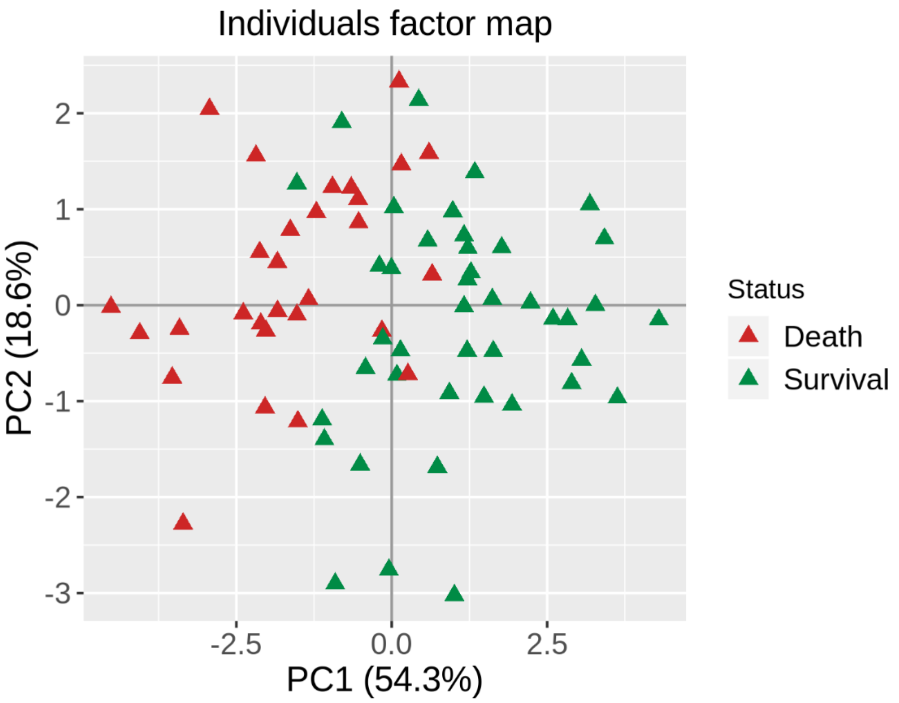
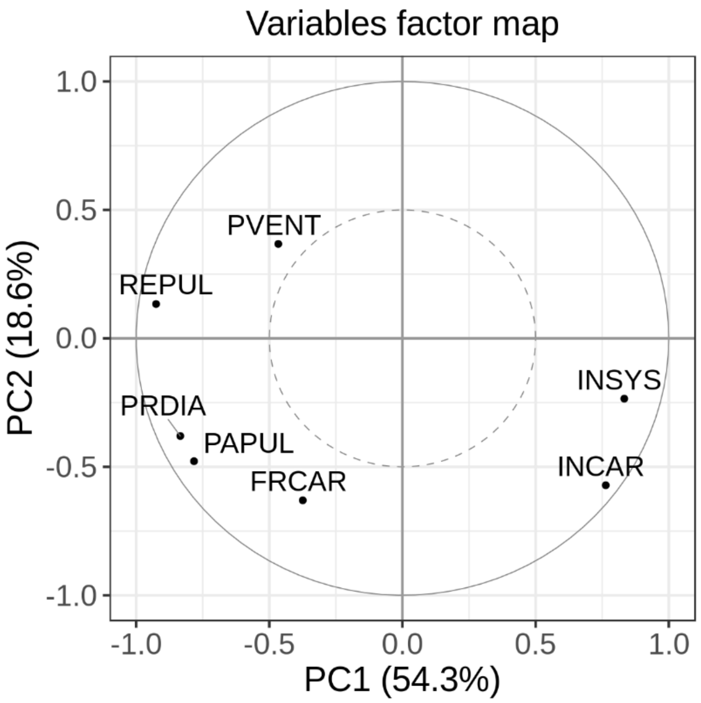


*Figure taken from J.-P. Fénelon: Camel or dromedary?*

# PCA: Example 1

**Goal**: 71 subjects with myocardial infarction (29 deaths, 42 survivors) for whom 7 variables were measured on admission to a cardiology department.





PCA provides a **projection of individuals** in a factorial plane constructed from the 7 cardiovascular variables.

The **correlation circle** shows which variables best explain the variation of the subjects on the 2 axes.

**FRCAR = Cardiac frequency, INCAR = Cardiac index, INSYS = Systolic index, PRDIA = Diastolic pressure, PAPUL = Pulmonary arterial pressure, PVENT = Ventricular pressure, REPUL = Pulmonary resistance**

***Source: J.-P. Nakache***

# PCA: Principle of the method

The projection of multidimensional data on a plane (2D) gives us a distorted vision of reality. The goal of PCA is **to determine reduced dimensional spaces that minimize these distortions**. We can then visualize the data in an “optimal” space, called **factorial plane**, generated by 2 perpendicular lines called **principal components**.

**Computation of the components**:

- Construction of a 1st component T1 so as to ① **minimize** the squares of the distances of the points to T1 and ② **maximize** the dispersion of the points (individuals) projected on T1
- Construction of T2 orthogonal to T1 and maximizing the dispersion on T2
- And so on, in order to capture as much variance as required…

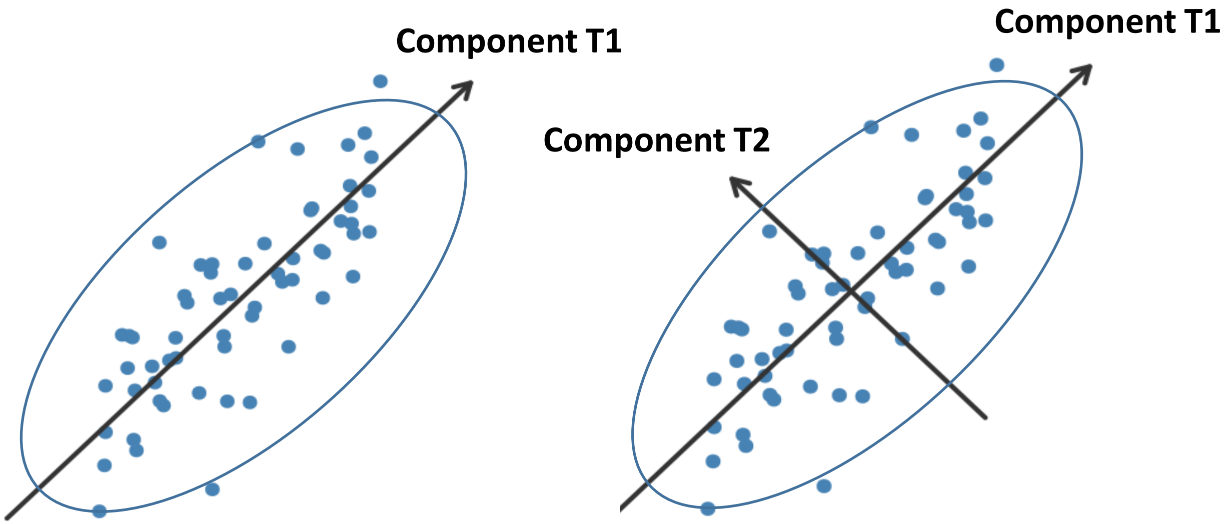


***Figure. The component T1 must “capture” the maximum inertia of the dataset***

# PCA: a spotlight on data

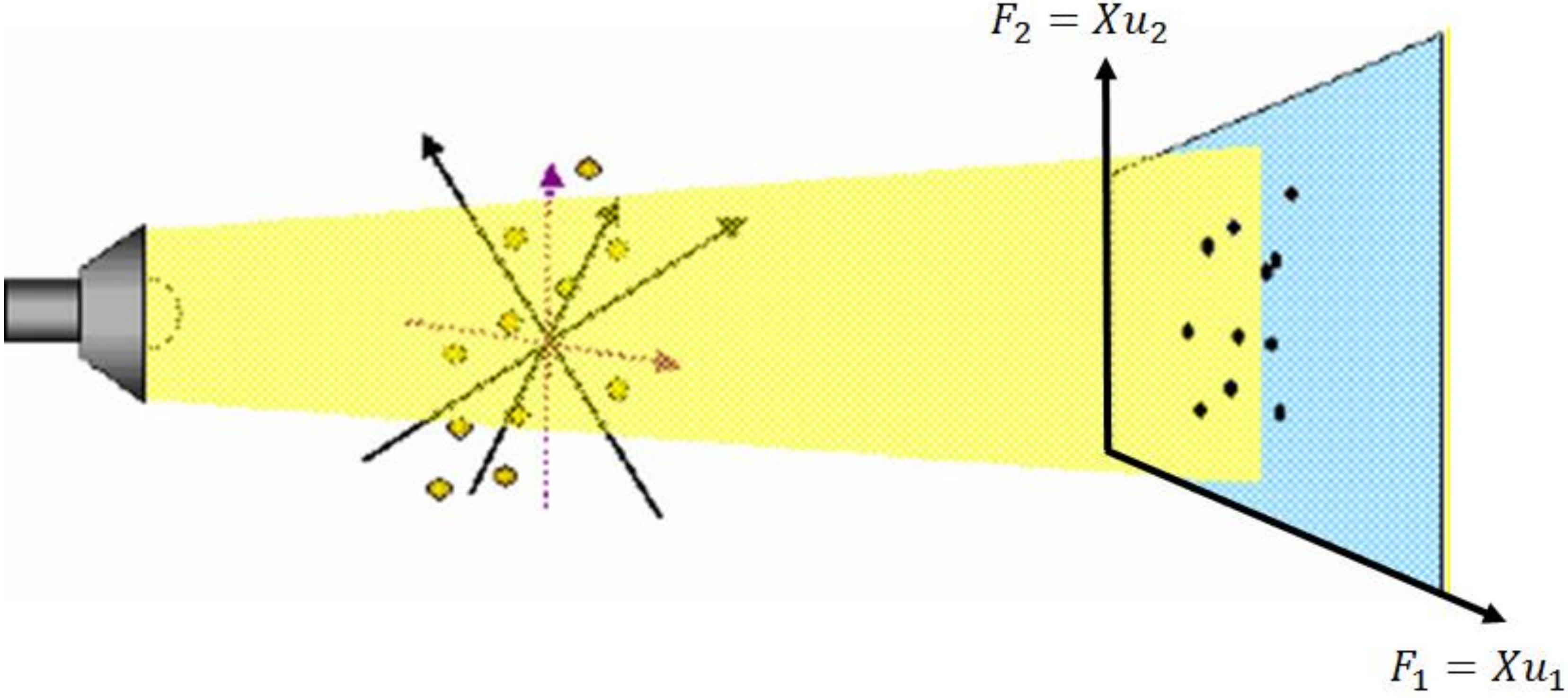


***Image from Umetrics AB, Umeå, Suède,***
***And reproduced in M. Tenenhaus,***
***Statistiques : Méthodes pour décrire, expliquer et prévoir, Dunod, 2007***

# PCA: the method in brief

***(2) Diagonalization step***

***(1) Data standardization***

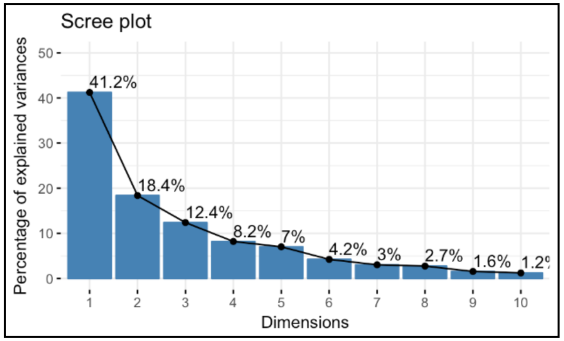
n x p table of centered reduced data

→

Correlation matrix p x p

$$\begin{pmatrix} \lambda_1 & & & 0 \\ & \lambda_2 & & \\ & & \dots & \\ 0 & & & \lambda_p \end{pmatrix}$$

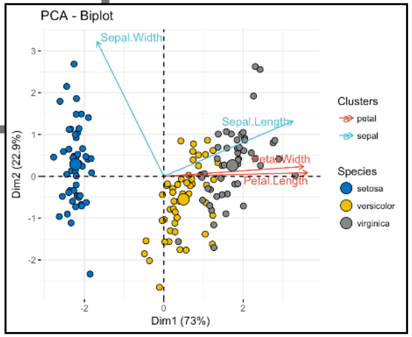
→



p x p matrix of eigenvectors

→

**Matrix n x p' (p' << p) of principal components**



***(3) Projection of the data on the eigenvectors (orthogonal directions of maximum variance)***

# PCA: Example 2

Data represent RNA-Seq expression values from 8 tissues, with multiple biological replicates for each tissue.

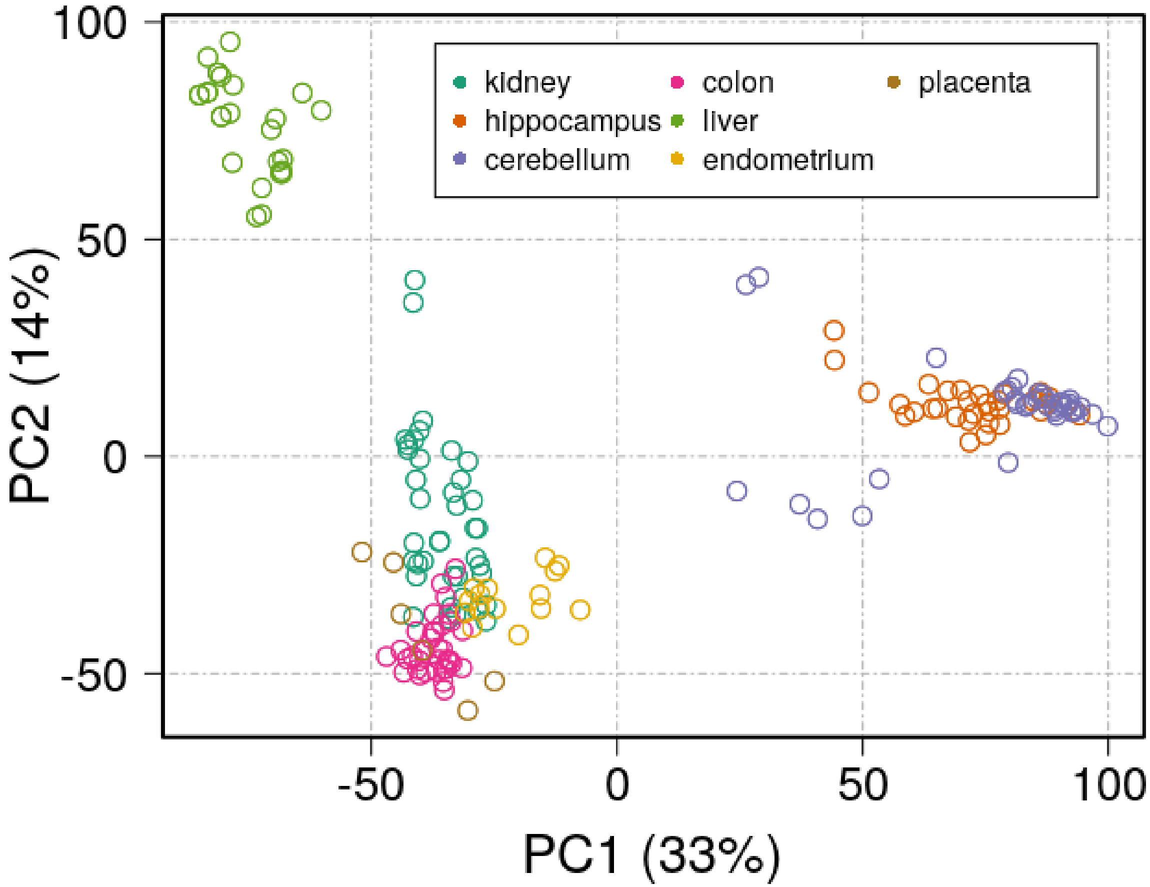


***Figure. By coloring the dots per tissue, the projection of the samples indicates the homogeneity of the expression profiles in each tissue, as well as the differences in expression between the 8 tissues.***

# To go further with PCA

## Principal Component Analysis

Michael Greenacre[1], Patrick J. F. Groenen[2], Trevor Hastie[3], Alfonso Iodice d'Enza[4], Angelos Markos[5], and Elena Tuzhilina[3],

[1] *Universitat Pompeu Fabra and Barcelona School of Management, Barcelona, Spain*
[2] *Erasmus School of Economics, Erasmus University, Rotterdam, The Netherlands*
[3] *Stanford University, Palo Alto, California, USA*
[4] *University of Naples Federico II, Naples, Italy*
[5] *Democritus University of Thrace, Alexandroupolis, Greece*

This is a preprint of an earlier version of the review published in *Nature Reviews Methods Primers.*

**Greenacre, M., Groenen, P.J.F., Hastie, T. *et al.* Principal component analysis. *Nat Rev Methods Primers* 2, 100 (2022). https://doi.org/10.1038/s43586-022-00184-w**

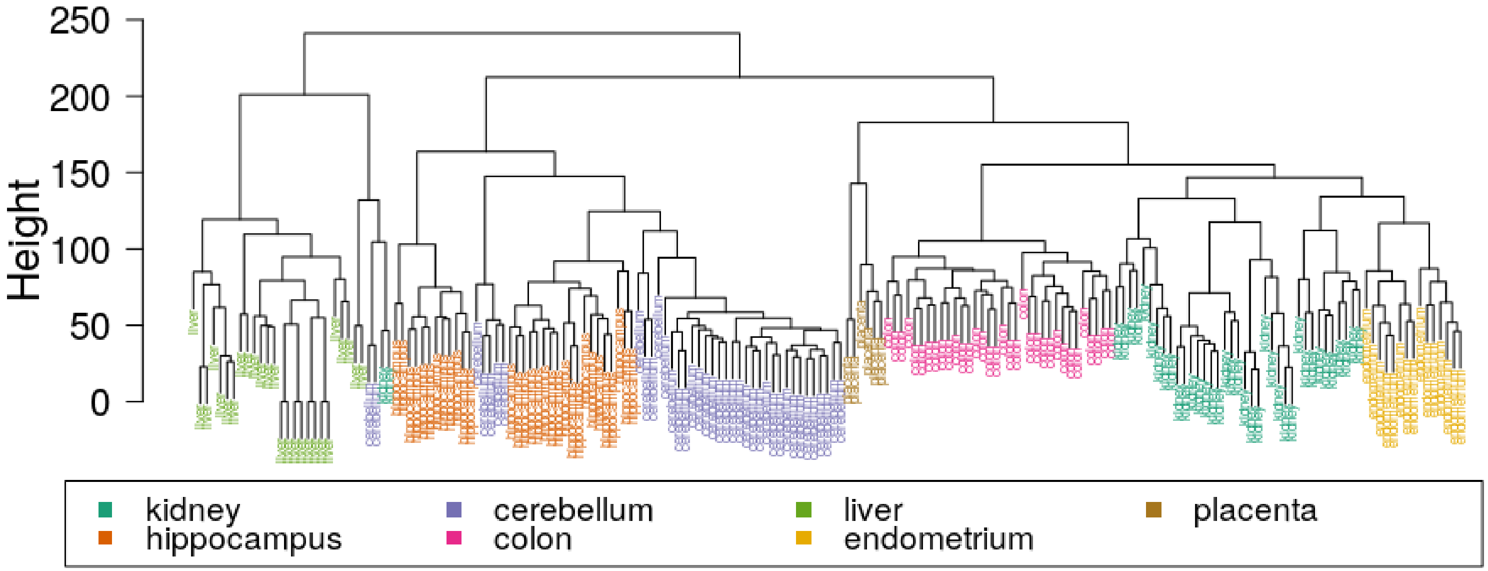
# Hierarchical ascending classification

The **HAC method** aims at gathering individuals in homogeneous and well separated groups (clusters) according to a similarity criteria.

The algorithm is mainly based on 2 criteria:

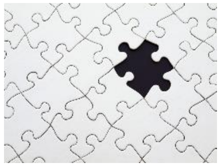
1. **Choice of a distance**: euclidean, max, manhattan…
2. **Aggregation strategy**: diameter (complete), moyenne (average), ward…

**Representation of clusters by a dendrogram**



***Expression data of the 8 tissues: euclidean distance and diameter method***

# Missing data



**Missing data** (MD) are frequent if not inevitable in databases, these can come from various reasons:

- Missing measurements;
- Data measured but lost or not reported;
- Data measured but the value is considered unusable (obvious error of measurement, the value seems aberrant);
- Data not available: *e.g.* “Don’t know” response;
- Censoring case: the value is outside the detection limits of the device;
- Censorship in a survival study:
    - Left cens.: the subject has already experienced the event before the start of the study,
    - Right cens.: the event has not been observed at the end of the study;
- Genetics: punctual absence of genotype (SNPs of some individuals)...

# MD management methods

In some cases, it is feasible to conduct analysis without the need to impute missing data, especially when removing individuals with missing data does not result in a significant loss of available information.

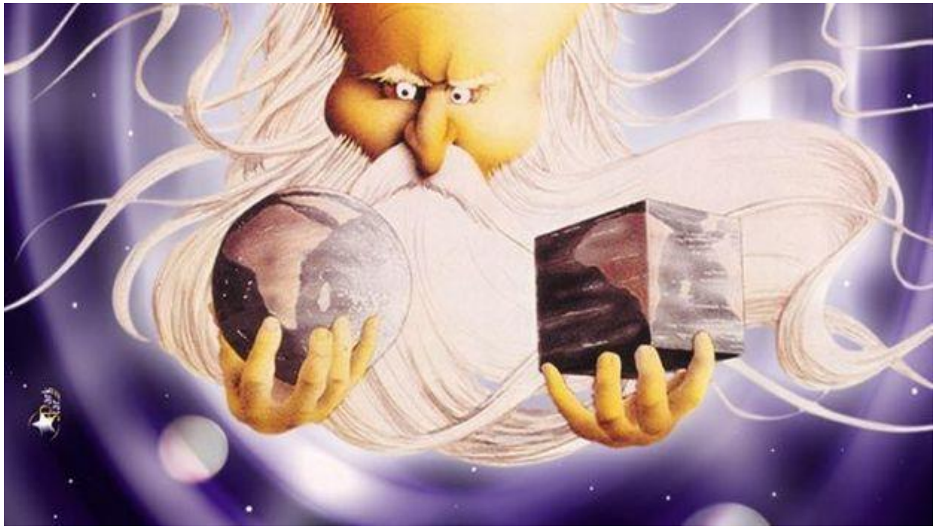
Otherwise, various strategies exist for **MD imputation**:

- **Simple imputation**: The MD is replaced by a unique value, obtained by averaging the k nearest observations (k-NN), by local regression, the NIPALS algorithm, SVD, or the use of random forests...
- **Multiple imputation**: A MD is replaced by several candidate values allowing to take into account in the analysis the additional uncertainty linked to the replacement of the MD
- **Bayesian approach**: It is assumed that the MDs are derived from a prior probability distribution

  *https://www.math.univ-toulouse.fr/~besse/Wikistat/pdf/st-m-app-idm.pdf*

- **Genetics**: reconstruction of missing SNPs by haplotype from a reference population (“most probable” genotype values)

# Part 2: Hypothesis testing



***Choose between two hypotheses***

***Illustration: Monty Python's The Meaning of Life (1983)***

# Principle of statistical tests

A **statistical test** is a decision procedure between two hypotheses concerning one or more samples.

*Example. In a study comparing a "drug" group and a "placebo" group, blood pressures measurements are taken from both groups to determine whether the drug has an effect on blood pressure.*

Hypothesis formulation: If µ1 and µ2 are the mean blood pressure of the 2 "drug" and "placebo" groups, one way to establish the effect of the drug on blood pressure is to demonstrate that µ2 is different from µ1 from the collected observations.

Goal of the test: To ascertain whether the ***observed difference between the two means*** can be attributed to ***chance***, *i.e.* due to sampling fluctuations; or if instead, it reflects a ***real*** difference (sufficiently probable).

# Test hypotheses

The **test hypotheses** transpose the biological question (*e.g.*, the effect of a drug) into two complementary statements using the distribution characteristics of the study population (parameter values, shape of the distribution...), called:

- **Null hypothesis noted H0**: the one that is considered true *a priori*;
- **Alternative hypothesis noted H1**: hypothesis complementary to H0.

The objective of the test is then to decide whether the model described by H0 is "plausible".

Example.

H0: the drug has no influence   against   H1: the drug has an influence

or

H0: $\mu1 = \mu2$   vs   H1: $\mu1 \neq \mu2$

H1 is said to be **two-sided** when there is no need to know the direction of the difference (*i.e.* $\mu1 \neq \mu2$), or **one-sided** if we are interested in a particular direction (*i.e.* $\mu1 < \mu2$ or $\mu1 > \mu2$).

# Examples of formulation for H0 and H1

1. **Comparison of 2 new treatments A and B**

   ***H0***: *A and B have equivalent efficiency*
   ***Two-sided* H1**: *A and B have a different efficiency*

2. **Comparison of treatment A to placebo (inactive product)**

   ***H0***: *Treatment A and placebo are equivalent*
   ***One-sided* H1**: *Treatment A is more effective than placebo*

3. **Effect of an anti-tumor drug related to the presence of a V variant**

   ***H0***: *Independence between the observed effect (+ or -) and the presence of the variant (V-/V+)*
   ***H1***: *Existence of a relationship between the 2 factors*

4. **Comparison of 4 treatments A, B, C and D**

   ***H0***: *The 4 treatments are equivalent*
   ***Two-sided* H1**: *At least one treatment is different from the others*

In a statistical test, “choosing between H0 and H1” is done in order **to avoid 2 types of errors**, called **Type I and Type II errors**.

# Test statistics

Since we have to choose between “H0 and H1”, the conclusion of a test will be based on a **decision variable** called the **test statistic S**. S is a random variable that summarizes the information contained in the sample and whose observed value $\mathbf{s_{obs}}$ can be calculated from the observations.

***Examples of usual S statistics:***

- □ Parametric tests: Student’s **T**, Fisher’s **F** in the ANOVA test…
- □ Non-parametric tests: Mann-Whitney **U**, Kruskal-Wallis **K** statistic…

**To decide…**

☺ The distribution of S is known under H0, which makes it possible to control for the Type I error of “falsely rejecting H0” [FALSE POSITIVE]:
Example. ***Conclude that a treatment is effective when it is in fact ineffective…***

☹ Conversely, the distribution of S is unknown under H1, which makes it difficult to control for the Type II error of “falsely accepting H0” [FALSE NEGATIVE]:
Example. ***Risk of not detecting the efficiency of a treatment…***

# Alpha risk (α) or Type I error

## If H0 is true…

**the Type I error is the Probability of "falsely rejecting H0":**

**$\alpha = P_{H0}$(reject H0), where α and $P_{H0}$ are assumed to be known**

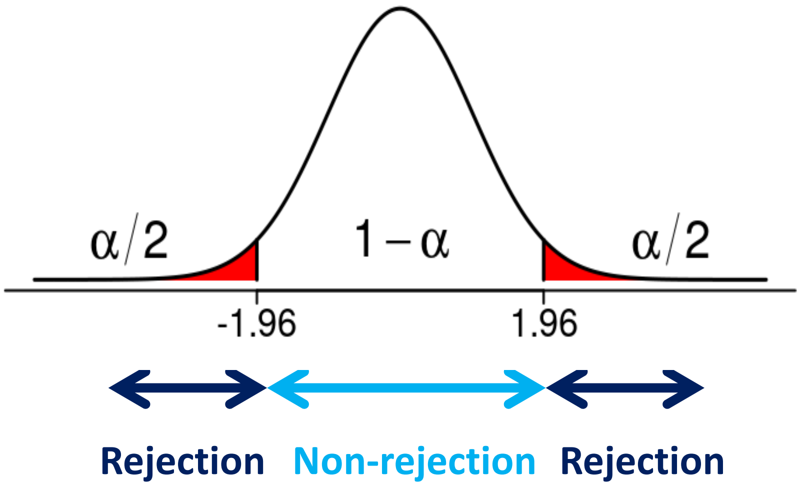
The **significance level α** is set *a priori* by the experimenter (usually 5% or 1%).

Let $\mathbf{s_{obs}}$ be the observed value of the **S** statistic: according to $P_{H0}$ (theoretical distribution of S under H0), $s_{obs}$ can be "probable enough" or "less probable":

For example, if **S** ~ N(0,1):

-1.96 and 1.96 are the quantiles 2,5% and 97.5% of the Normal distribution delimiting the **rejection region**

Decision rule according to the values of $\mathbf{s_{obs}}$ (**two-sided H1**, α = 5%):



**Rejection** **Non-rejection** **Rejection**

# Beta risk (β) or Type II error

## If H1 is true…

**the β error of Type II is the Probability of “falsely accepting H0”:**

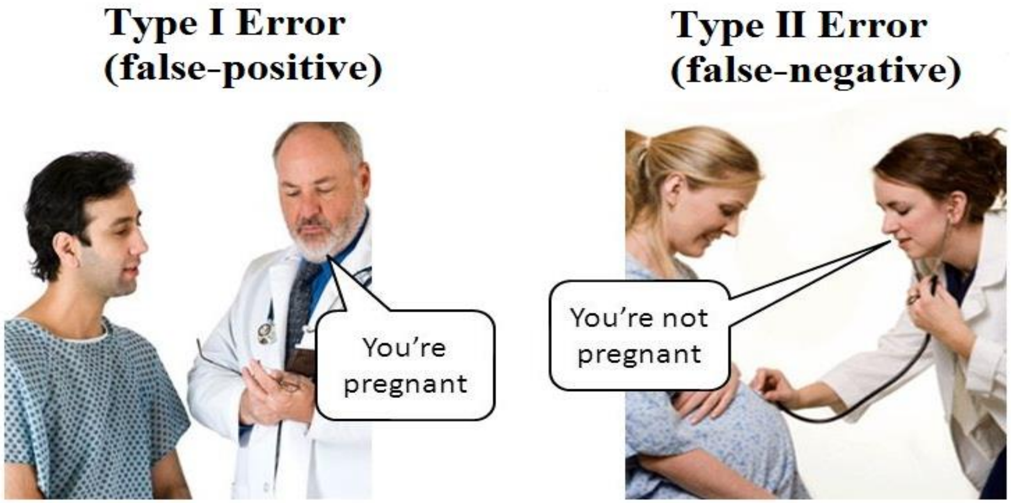
**$\beta = P_{H1}$(reject H1), with $P_{H1}$ unknown and the value β undetermined in general**

The quantity 1−β is the **power** of the test.

| | Truth | |
|---|---|---|
| **Decision** | H0 is TRUE | H1 is TRUE |
| H0 is accepted | Good decision **TN**<br>Confidence level 1 – α | Bad decision **FN**<br>Error β |
| H0 is rejected | Bad decision **FP**<br>Error α | Good decision **TP**<br>Power of the test 1 – β |

# Type I and type II errors

Ideally, a « good test » should minimize both types of error:



***H0: "You're not pregnant"***
***vs***
***H1: "You're pregnant"***

This minimization is actually a compromise to be made as the 2 types of errors are closely related. For example, by varying the value of alpha:

- ***if α = 10% instead of 5%, type I error becomes more probable -> detecting effects is easier, but detection errors occur more frequently***
- ***if α = 1% instead of 5%, type II error becomes more probable -> risk of missing effects, but fewer errors in detection***

While controlling the α error is not a problem, we will see in the following how it is possible to control the β error (dependent on effect and sample sizes).

# Test result report: *p*-value (1/2)

The most common way to report a test result is to indicate its ***p*-value** which

## If H0 is true…

represents the probability of obtaining the observed value of the test statistic (or an even more extreme value):

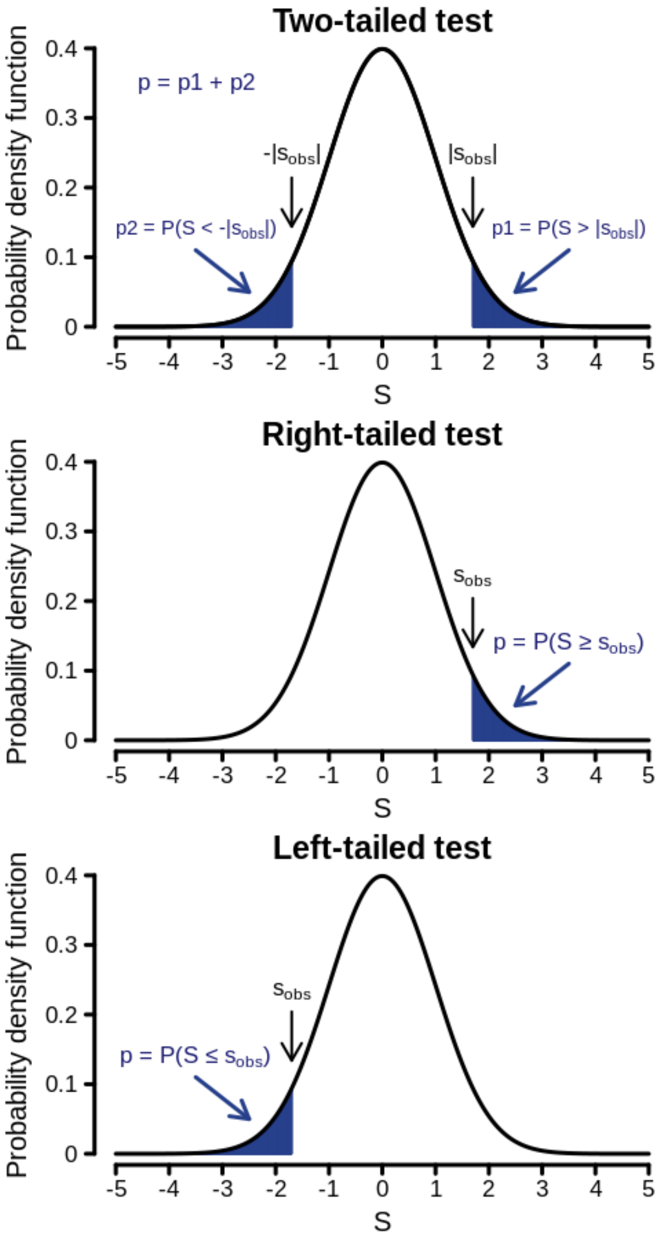
- If the test is two-sided: $p = P_{H0}(|\mathbf{S}| > \mathbf{s_{obs}})$
- If the test is upper-tailed: $p = P_{H0}(\mathbf{S} \geq \mathbf{s_{obs}})$
- If the test is lower-tailed: $p = P_{H0}(\mathbf{S} \leq \mathbf{s_{obs}})$

☞ *The smaller the p-value, the more likely H0 is to be rejected because the obtained value of $s_{obs}$ is considered "too unlikely" when H0 hold.*

Usual thresholds:

- $p < 0.001$: *very high significance* *** ☺
- $p < 0.01$: *high significance* **
- $p < 0.05$: *significance* *
- $p > 0.05$: n.s. ☹

Theoretical distribution of S under H0



CAUTION

# Test result report: *p*-value (2/2)

The interpretation of the test result based exclusively on the p-value can be delicate in practice, if one considers that **the concept of "statistical significance" may not be in agreement with that of "biological significance"**, *i.e.* the real importance that can be given to an observed biological effect.

**Common practical situations:**

1. Difference between 2 groups of subjects, considered "important" by the experimenter, does not pass the threshold of statistical significance simply because of sampling variations measured on ***too small numbers of subjects***, resulting in a "lack of power" of the test.
2. Very small or biologically insignificant effects, considered statistically significant just because of the ***too large number of subjects*** for which "everything becomes significant!"

In both cases, the additional indication of a quantitative measure describing the true magnitude of the observed effect (**effect size**), independently of the size of the population (*e.g.* difference in means), is important for properly balancing the concepts of statistical and biological significance.

# Test result report: Effect size or *when p-value is not enough!*

**In addition to the p-value**, we can look at the distance of an observed value on the sample from its **norm indicated by H0**. This distance is called an **effect** and the importance of this effect can be evaluated through a statistic called **effect size**.

**Absolute effect size:**

- $\mathrm{M_{obs} - M_{H_0}}$ (if for example M is a mean difference)
- $\mathrm{p_{obs} - p_{H_0}}$ (if for example p is a proportion)

**Relative effect size (without units):**

$$\mathrm{d = \frac{M_{obs} - M_{H_0}}{SD_{sample}}} \qquad \text{or} \qquad \text{ratio of proportions: } \mathrm{\frac{p_{obs}}{p_{H_0}}}$$

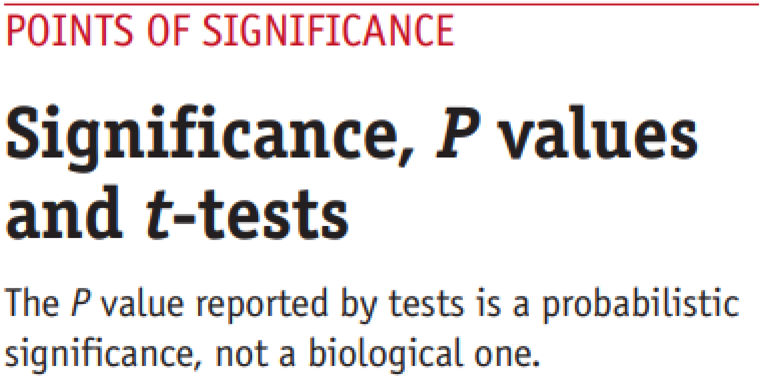
For the **Cohen's d**, the effect size can be described by using the following scale: "weak" effect around 0.2, "moderate" around 0.5 and "strong" around 0.8.

As it is a unitless value, the effect size can also be used in **meta-analyses** combining effect sizes from different studies in integrative studies.

**Other types of effect sizes:** Correlation, odds ratios...

***(Sullivan & Feinn, J Grad Med Educ 2012)***

# To go further with *p*-values, significance and effect size

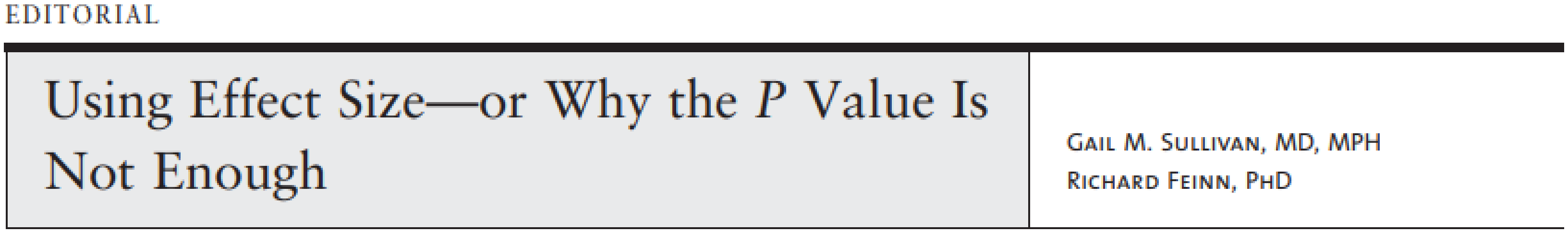


POINTS OF SIGNIFICANCE

**Significance, *P* values and *t*-tests**

The *P* value reported by tests is a probabilistic significance, not a biological one.

**Krzywinski, M., Altman, N. Significance, *P* values and *t*-tests. *Nat Methods* 10, 1041–1042 (2013). https://doi.org/10.1038/nmeth.2698**



EDITORIAL

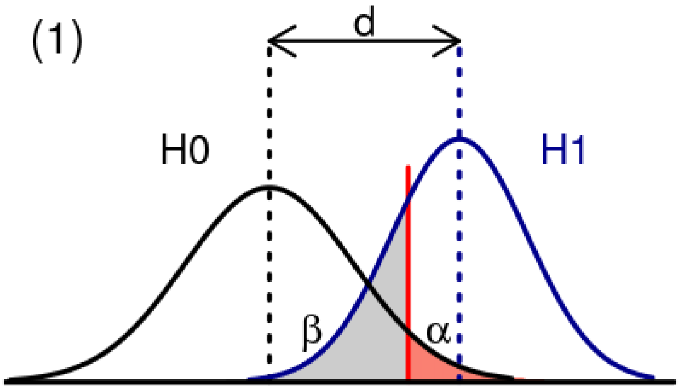
Using Effect Size—or Why the *P* Value Is Not Enough

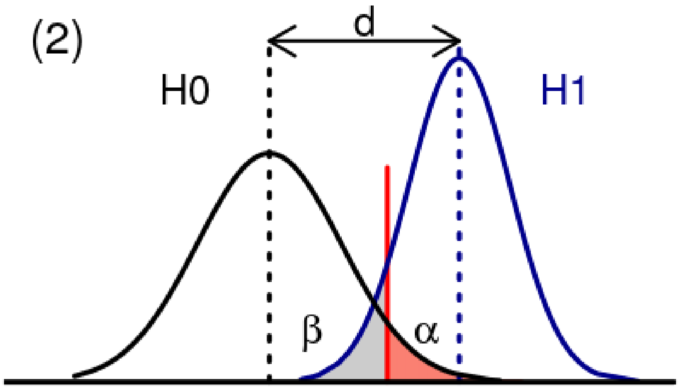
Gail M. Sullivan, MD, MPH
Richard Feinn, PhD

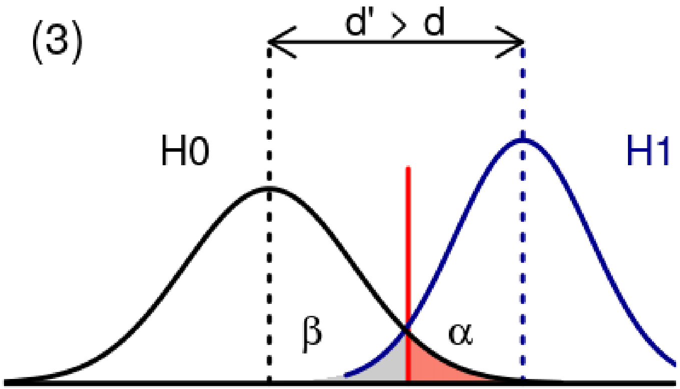
**Sullivan GM, Feinn RS. Using effect size—or why the P. value is not enough. J Grad Med Educ. 2012;4(3):279–282. doi:10.4300/JGME-D-12-00156.1.**

# Power of a test

The **power of a test (1-β)** is its capacity to detect deviations from the null hypothesis. While α is fixed, the power (and type II error β) will depend on the sample size, the dispersion and the size of the difference (or effect size d) “that is supposed to exist”.







***(1) vs (2) same effect size d: power increases with the number of subjects***

***(1) vs (3) same number of subjects: power increases with effect size***

# Sample size

To conduct a successful test, choosing an appropriate **sample size** is crucial for achieving sufficient power:

☹ A size that is **too small** will tend to give larger standard deviations that can lead to missing effects when they really exist (lack of power).

☹ Conversely, a size that is **too large** will tend to detect tiny “statistically significant” effects (systematic rejection of H0), even for effects too small to make biological sense.

It is therefore essential to determine the best suitable sample size for the experiment, ***“somewhere between too many and not enough…”***. For this, calculation rules exist to obtain the sample size according to:

- the effect size,
- the dispersion,
- the required power level

# Power and sample size

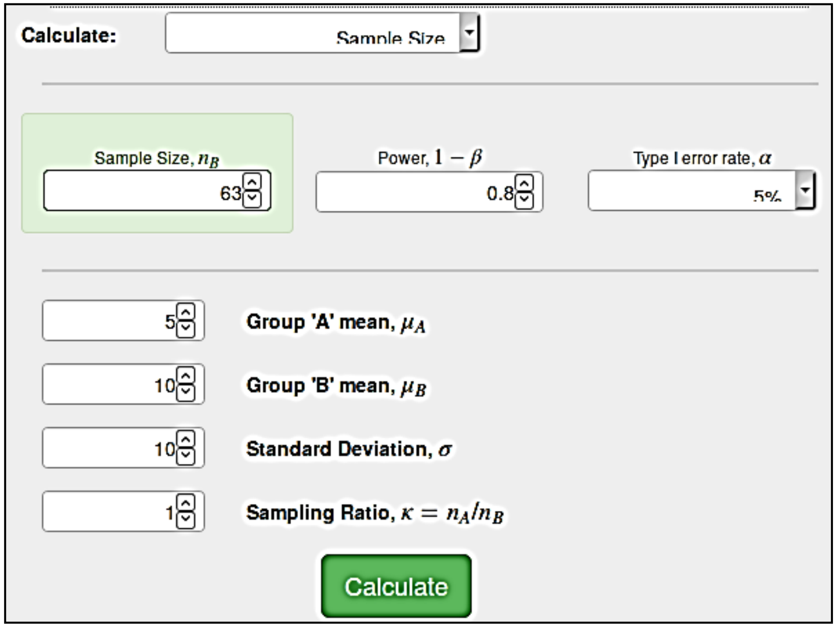
**Power** is important because it tells us how likely it is that the test will identify a significant difference or effect when it actually exists.

To conduct an experiment with a high power level (often 1 – β = 0.8), a prior knowledge of the effect size and dispersion is necessary for determining the appropriate **sample size**. For this purpose, a **pilot study with a small sample size** may be useful to indicate these values, in preparation for a larger scale study.

Explicit mathematical formulas are available for most of the standard parametric tests facilitating the calculations of power or sample sizes.

These calculations can be easily performed using **online calculators** wherein you simply need to choose the desired test and input the anticipated parameter values.

***Chow S, Shao J, Wang H. 2008. Sample Size Calculations in Clinical Research. 2nd Ed. Chapman & Hall/CRC Biostatistics Series.***



Calculate: Sample Size

Sample Size, $n_B$: 63

Power, $1-\beta$: 0.8

Type I error rate, $\alpha$: 5%

5 Group 'A' mean, $\mu_A$

10 Group 'B' mean, $\mu_B$

10 Standard Deviation, $\sigma$

1 Sampling Ratio, $\kappa = n_A/n_B$

http://powerandsamplesize.com/

# Conducting a statistical test

## Main steps of a test:

1. Choose the **appropriate test** for the question asked and the type of data
2. Establish **H0** and **H1** (the 2 hypotheses must be mutually exclusive and include all possibilities)
3. Set the **significance level** of the test (α)
4. Calculation of the **test statistic** (decision variable whose theoretical distribution is known under H0) ***-> STAT SOFTWARE***
5. Calculation of the ***p*-value** ***-> STAT SOFTWARE***
6. Conclude

The choices (steps 1, 2 and 3) as well as the final interpretation (step 6) remain at the discretion of the experimenter. Never forget that the result of a test always includes a dose of uncertainty.

☞ **WE WILL NEVER KNOW IF WE MADE THE RIGHT DECISION!**

# Multiple comparisons problem (1/2)

The problem of **multiple comparisons** arises when a statistical analysis involves testing several hypotheses at once.

The multiplication of tests on the same data set then leads to an increase in the risk of being wrong by highlighting significant differences that are only due to chance (case of FALSE POSITIVES).

From a statistical point of view, we say that the **overall alpha risk** is increasing. In general, the alpha risk for a single test is set at 5%. However, when conducting k tests, the global alpha risk becomes: $\boxed{\alpha_{global} = 1 - (1 - \alpha)^k}$

For $\alpha$ = 5%, we have thus: $\alpha_{global}$ = 0.487 (k=13) et $\alpha_{global}$ = 0.512 (k=14)

*This implies that over 13 comparisons, there is more than a 50% chance of finding one or more significant differences just by chance!!!* ☹

☞ **Corrections exist to achieve a global error rate of 5% for all tests performed.**

# Multiple comparisons problem (2/2)

Les devises Shadok
EN ESSAYANT CONTINUELLEMENT ON FINIT PAR RÉUSSIR. DONC: PLUS ÇA RATE, PLUS ON A DE CHANCES QUE ÇA MARCHE.

Shadok proverb

*"If you keep trying, you will finally succeed. So: The more it fails, the more likely it will work."*

# Corrections of multiple tests

**Bonferroni**: *the p-values only increase with the number of tests performed*

*Carlo Bonferroni (1892-1960)*

- FWER type correction (Familywise Error Rate)
- **Very conservative** procedure (most stringent)
- Calculation:

$$\boxed{p_{\mathrm{adj}} = \min(p \times nbp, 1)}$$

*nbp*: number of tests

**Benjamini-Hochberg**: *p-values increase with their number and the rate of non-significant p-values*

*Yoav Benjamini (1949-)*

- FDR type correction (False Discovery Rate)
- **Not very conservative** (better suited for selecting traits of “potential interest”)
- Commonly used in differential expression analysis
- Calculation:

$$\boxed{p_{(i)}^{adj} = \min\left\{\min_{j \geq i}\left\{\frac{nbp \times p_{(j)}}{j}\right\}, 1\right\}}$$

*i*: rank of *p* in the p-values ordered in ascending order

# 1/ Parametric tests

For normally distributed observations, or given a sample size large enough of independent data with same distribution for establishing asymptotic normality of the sample mean (**central limit theorem** with n > 30):

## 1 sample

- Compare the sample mean to a theoretical value (variance assumed known, **Gauss**)
- Compare the sample mean to a theoretical value (unknown and estimated variance, **Student**)
- Compare a proportion to a theoretical value (**one-sample binomial tes**t)

## 2 independent samples

- Comparison of 2 means (equal variances or sufficiently large samples, **Student**)
- Comparison of 2 variances (**Fisher**)
- Comparison of 2 proportions (**Chi-square test for equality of proportions with Yates continuity correction**)

**2 paired samples:** same sample observed at 2 different times or under 2 different conditions (**Student's paired t-test**)

**More than 2 samples:** **One-way ANOVA**, **Bartlett**'s and **Levene**'s tests to compare several variances

***Taken from https://www.math.univ-toulouse.fr/~besse/Wikistat/pdf/st-l-inf-tests.pdf***

# Student’s t-test (n < 30)

*William Sealy Gosset “Student” (1876-1937)*

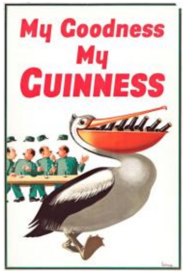
**Goal:**

- Compare the mean of a sample to a theoretical mean
- Compare the means of 2 samples
- Compare the means of 2 paired series
- Testing a correlation coefficient

**Principle of the test (equal variances):**

**/!\ Welch’s t-test for 2 samples of unequal variances**

1. Estimate the mean and standard deviation of each sample
2. Calculate the value of the statistic $$t_0 = \frac{\bar{x}_1 - \bar{x}_2}{s_p \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$
3. Determine the number of df = $n_1 + n_2 - 2$ to extract the critical value of the Student distribution corresponding to the risk level α
4. Compare $t_0$ to the critical value and conclude



*As the publication of the work would have required the prior agreement of the Guinness Company, William Gosset, statistician and head brewer, published his work in statistics under the pen name “Student”.*

# One-way ANOVA (1/2)

***Ronald Fisher (1890-1962)***

For a variable X measured on n individuals grouped in p groups, the ANOVA consists in constructing the following hypothesis test:

$$\begin{cases} \mathrm{H}_0 : \mu_1 = \mu_2 = \ldots = \mu_p = \mu \\ \mathrm{H}_1 : \exists \mu_j \neq \mu \end{cases}$$

*“All groups are equal” vs “At least 1 group is different from the others”*

where $\mu_1, \mu_2, \ldots, \mu_p$ are the means of the $p$ groups and $\mu$ is the overall mean.

## Assumptions of One-way ANOVA

Observations should satisfy the following 3 assumptions:

- Data **independence**
- **Normality** of the distribution in the groups
- **Homoscedasticity**: equal variance in the different groups (Bartlett’s test)

# One-way ANOVA (2/2)

Following the **principle of variance decomposition**: Total variance = Within-class variance + Between-class variance, the **F-statistic** evaluates the ratio between the “explained” variance (INTER) and the residual variance (INTRA):

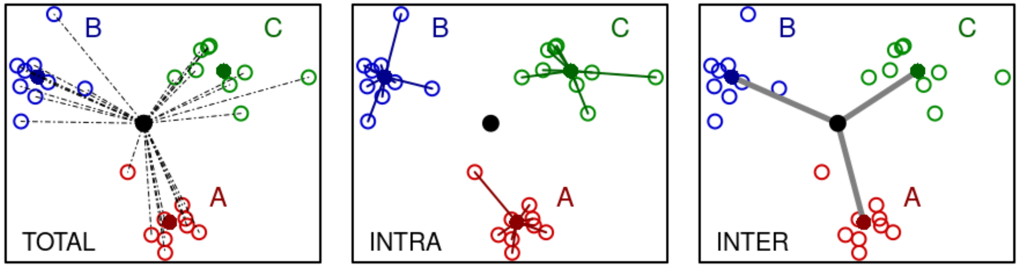
$$\sum_{j=1}^{p}\sum_{i=1}^{n_j}(x_{ij}-\bar{x})^2 = \sum_{j=1}^{p}\sum_{i=1}^{n_j}(x_{ij}-\bar{x}_j)^2 + \sum_{j=1}^{p}(\bar{x}_j-\bar{x})^2$$

$$\text{SST }(n-1\ ddl) = \text{SSE }(n-p\ ddl) + \text{SSM }(p-1\ ddl)$$

Under $H_0$:

$$F = \frac{\text{SSM}/(p-1)}{\text{SSE}/(n-p)}$$

$$\sim Fisher(p-1, n-p)$$



The observed value of *F* is then compared to the critical value of the Fisher distribution with p-1 and n-p df corresponding to the risk level α to conclude.

Note that the ANOVA test can detect a difference among the means, but it does not tell which groups are different from others!

☞ ***Post-hoc tests are required to perform multiple pairwise comparisons!***

# Post-hoc tests of ANOVA

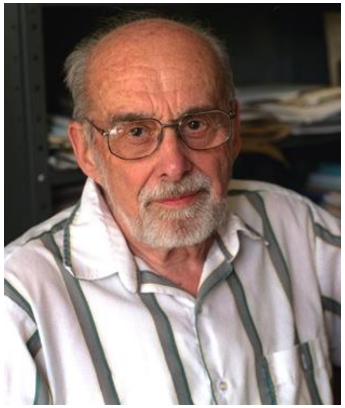
After a conclusive ANOVA test with "at least one group different from the others", a **post-hoc test** is a multiple comparison test that determines significant differences between the groups 2 by 2.

***John Tukey (1915-2000)***

2 types of post-hoc tests are commonly used:

- **Tukey HSD (honestly significant difference)**: Pairwise multiple comparisons comparing all the group means between them (*i.e.*, k×(k-1)/2 possible comparisons with k groups).

- **Dunnett**: Pairwise multiple comparisons comparing all means of the experimental groups to the mean of a given control group.



***Charles Dunnett (1921-2007)***

# Example: One-way ANOVA (1/4)

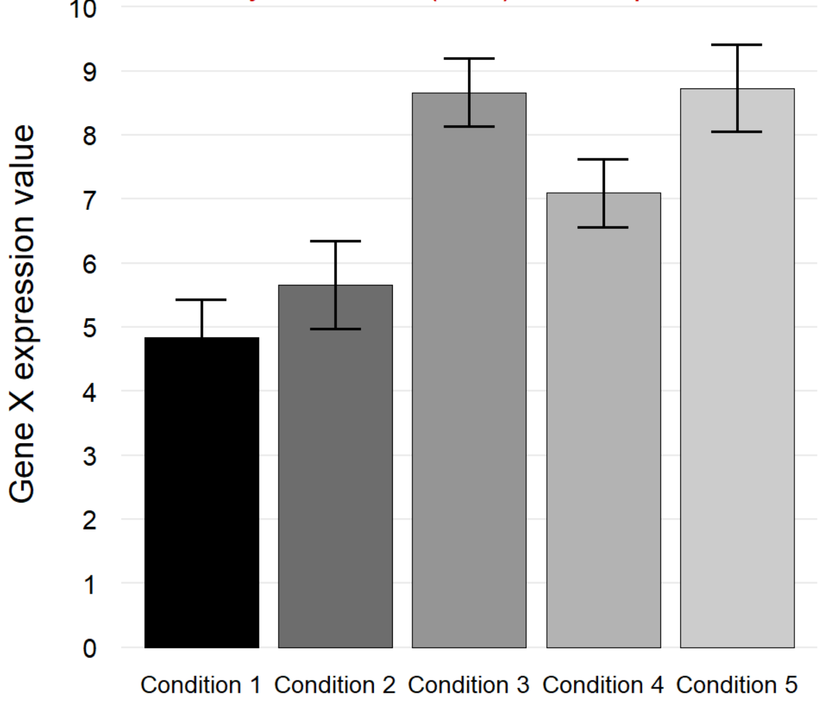
Example on simulated data comparing expression levels of a gene X under 5 different conditions.

```
summary(aov(Expression~Condition, data=dataset))
```

```
            Df  Sum Sq  Mean Sq  F value   Pr(>F)
Condition    4   147.0    36.75    8.267   2.7e-05 ***
Residuals   55   244.5     4.45
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residuals   55  244.5    4.45
```

One-way ANOVA: F(4,55) = 8.27, p = 2.7e-5



*Error bars indicate standard errors of the mean*

# Example: One-way ANOVA (2/4)

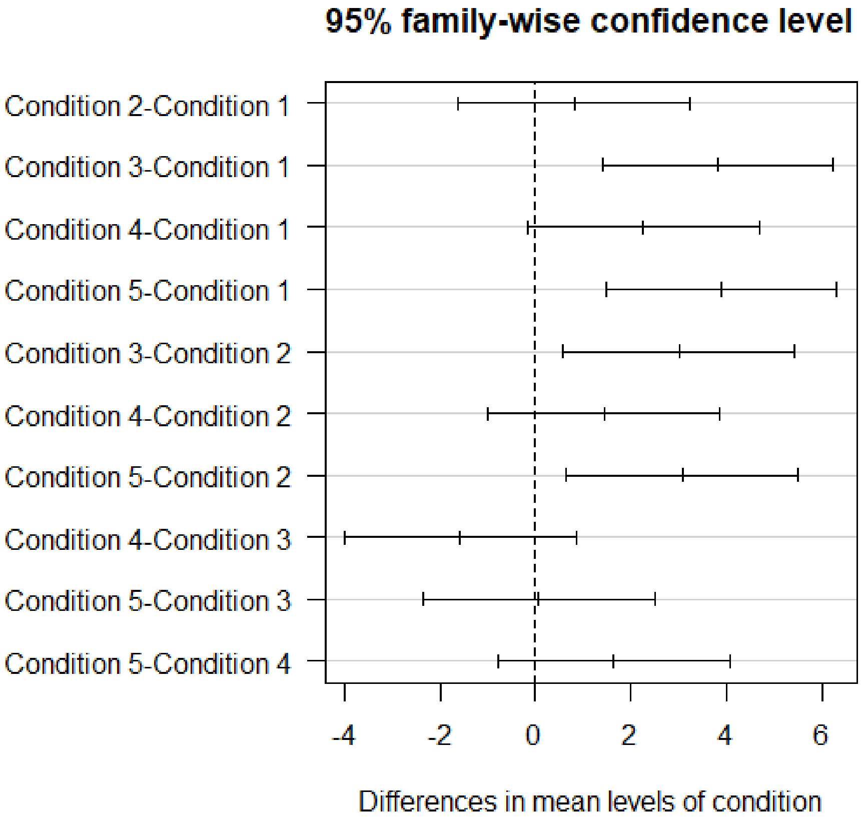
As the ANOVA test is conclusive, the Tukey HSD post-hoc test is performed for the pairwise comparisons of all conditions:

```
TukeyHSD(aov(Expression~Condition, data=dataset))
```

```
Tukey multiple comparisons of means
  95% family-wise confidence level

Fit: aov(formula = lm(Expression~Condition, data = dataset))

$condition
                           diff     lwr     upr     p adj
Condition 2-Condition  1  0.820  -1.608   3.247   0.87494
Condition 3-Condition 1   3.823   1.396   6.251   0.00041
Condition 4-Condition 1   2.256  -0.172   4.683   0.08044
Condition 5-Condition 1   3.896   1.468   6.323   0.00031
Condition 3-Condition 2   3.004   0.576   5.431   0.00820
Condition 4-Condition 2   1.436  -0.992   3.864   0.46158
Condition 5-Condition 2   3.076   0.648   5.504   0.00640
Condition 4-Condition 3  -1.567  -3.995   0.860   0.37221
Condition 5-Condition 3   0.072  -2.355   2.500   0.99999
Condition 5-Condition 4   1.640  -0.788   4.068   0.32684
```



# Example: One-way ANOVA (3/4)

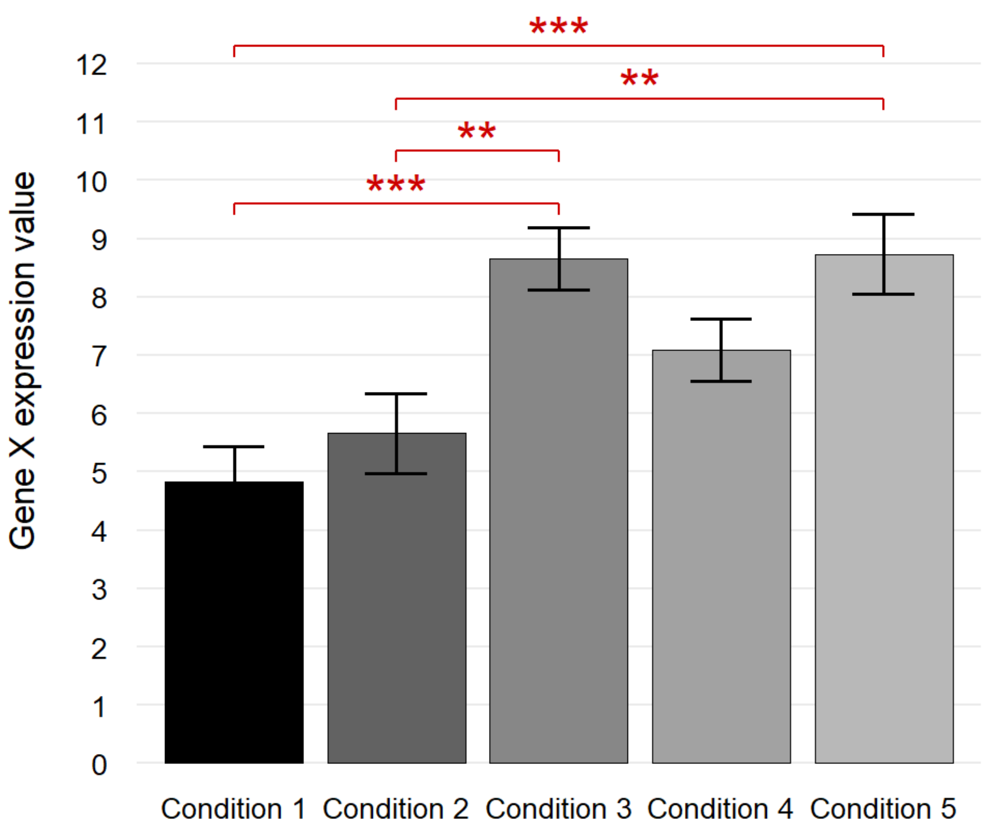
As the ANOVA test is conclusive, the Tukey HSD post-hoc test is performed for the pairwise comparisons of all conditions:

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = lm(Expression~Condition, data = dataset)

$condition
                             diff     lwr     upr    p adj
Condition 2-Condition  1    0.820  -1.608   3.247  0.87494
Condition 3-Condition 1     3.823   1.396   6.251  0.00041
Condition 4-Condition 1     2.256  -0.172   4.683  0.08044
Condition 5-Condition 1     3.896   1.468   6.323  0.00031
Condition 3-Condition 2     3.004   0.576   5.431  0.00820
Condition 4-Condition 2     1.436  -0.992   3.864  0.46158
Condition 5-Condition 2     3.076   0.648   5.504  0.00640
Condition 4-Condition 3    -1.567  -3.995   0.860  0.37221
Condition 5-Condition 3     0.072  -2.355   2.500  0.99999
Condition 5-Condition 4     1.640  -0.788   4.068  0.32684
```

One-way ANOVA: F(4,55) = 8.27, p = 2.7e-5



*Error bars indicate standard errors of the mean*

# Example: One-way ANOVA (4/4)

What should have been done before the ANOVA test: check the conditions of use of the test!

1. **Independence of the groups**: It is assumed that the experiment is conducted on independent observations within and between groups. ☑

2. **Normality of groups:** the Shapiro-Wilk test is applied to the 5 conditions

```
tapply(dataset$Expression, dataset$Condition, shapiro.test)
```

The smallest of the 5 p-values obtained is **0.1426** ☑

*Non-rejection of the normality hypothesis*

3. **Homogeneity of variances**: the Bartlett's test does not indicate a significant difference in variance between groups.

```
bartlett.test(Expression~Condition, data=dataset)
```

☑ *Non-rejection of the homogeneity hypothesis*

```
Bartlett test of homogeneity of variances

data:  Expression by Condition
Bartlett's K-squared = 1.2807, df = 4, p-value = 0.8646
```

# 2/ Tests for goodness-of-fit and independence

**Tests of goodness-of-fit to a given distribution**

- Good fit of the distribution of observations to a known distribution law (**Kolmogorov-Smirnov**)
- Normality of a distribution (**Kolmogorov-Smirnov**, **Shapiro-Wilk**)

**Test for homogeneity**

- Comparison of the distribution of levels of a categorical variable between several samples (**Chi-squared** or **Fisher's exact test**)

**Test of independence**

- Study of the joint distribution of 2 categorical variables (**Chi-squared** or **Fisher's exact test**)

# $\chi^2$ test of homogeneity

**2 categorical variables:** The notion of mean and variance no longer exists. We then try to compare 2 or more distributions observed on the samples to determine *"if they come from the same population"*.

**Example:** Distributions de **4 categories (l = 4)** compared on **3 groups of individuals (m = 3)**. Under $H_0$, *"the distributions are all the same and identical to the distribution observed on the whole samples"*.

$$\chi^2 = \sum_{i=1}^{l}\sum_{j=1}^{m} \frac{\left(n_{ij} - \frac{n_{i+}n_{+j}}{n}\right)^2}{\frac{n_{i+}n_{+j}}{n}}$$

*Under $H_0$, $\chi^2$ approximately follows a $\chi^2((l-1)*(m-1))$ distribution when **n ≥ 30** and all the **counts ≥ 5**.*

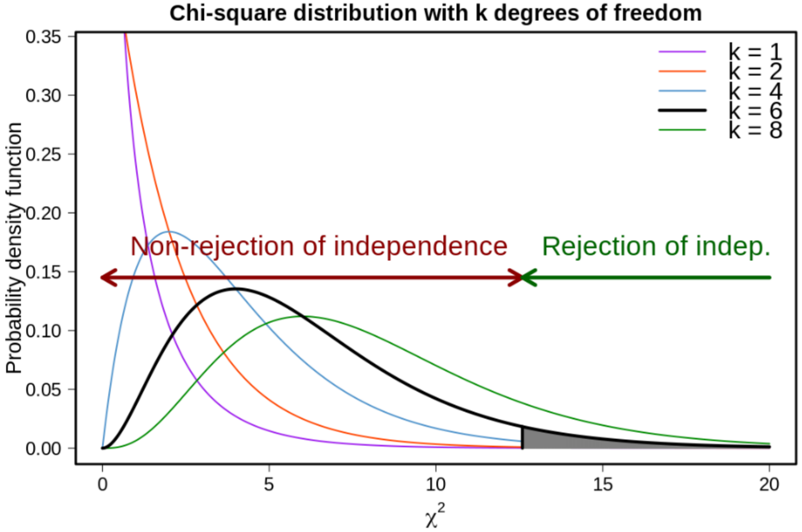
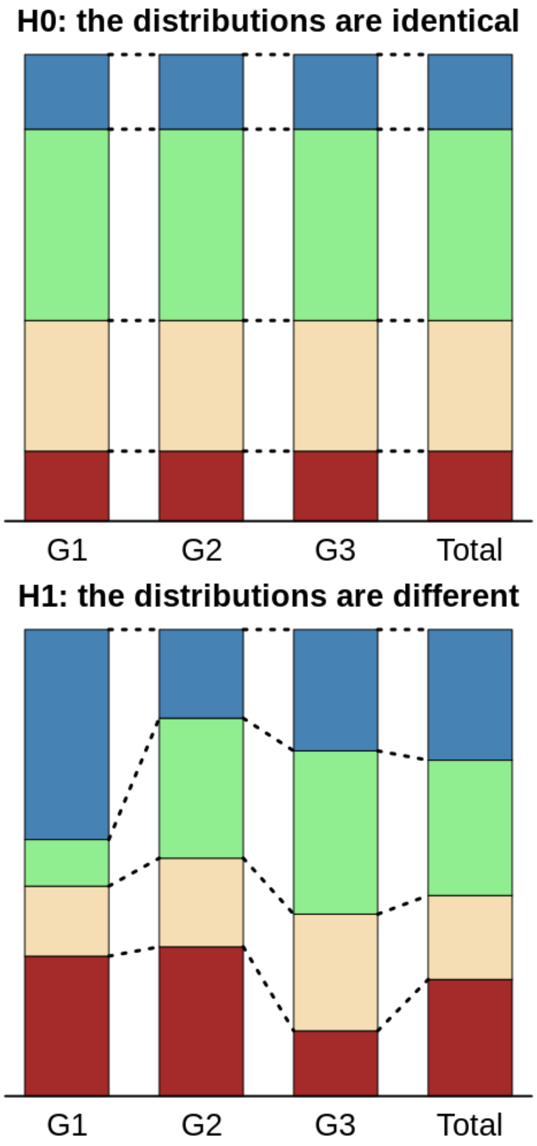


**Fig. Rejection region of the example with α = 5% and 6 df (black curve)**



# Example: $\chi^2$ test of independence

**Cooper study* of the efficacy of Zidovudine (AZT)** in a cohort of 936 asymptomatic HIV-positive subjects with CD4+ T-cell counts >400 mm3. Disease progression over 3 years was defined as the onset of AIDS symptoms or a significant decrease in CD4+ cells.

| TREATMENT | PROGRESSION | NO PROGRESSION | TOTAL |
|---|---|---|---|
| **AZT** | **76** | **399** | **475** |
| **Placebo** | **129** | **332** | **461** |
| **Total** | **205** | **731** | **936** |

**2 categorical variables with 2 levels**:

□ TREATMENT (AZT or Placebo)
□ DISEASE PROGRESSION (YES or NO)

**H0: The variables TREATMENT and PROGRESSION are independent**

```
chisq.test(Cooper)
```

```
Pearson's Chi-squared test with Yates' continuity correction
data: Cooper
X-squared = 18.944, df = 1, p-value = 1.346e-05
```

***Conclusion:***
***Rejection of the null hypothesis of independence indicating a probable treatment effect***

*Cooper et al. (N Engl J Med. 1993)*

# Application in genomics: Enrichment testing (1/2)

**Problem:** In a given set of genes, we want to know if a biological function “f” is more represented than in any other set of genes of the same size obtained “at random” from a random draw in the whole genome.

Examples of biological function: **Gene Ontology** (GO) term, **metabolic pathway** (KEGG or Reactome), or any other list of genes associated with a biological function of interest...

Under H0: There is no relationship between “f” and the gene set selected by the experiment, *i.e.* “No evidence that the biological function particularly characterizes the gene list”.

The functional enrichment test is based on a **Fisher’s exact test** using the **hypergeometric (discrete) distribution** described by the following 3 parameters:

- Size **N** of the population (reference or “background” genome)
- Size **n** of the studied gene set
- Probability ***p*** of a favorable event in the population (*i.e.*, of randomly drawing a gene associated with “f”)

# Application in genomics: Enrichment testing (2/2)

Let **N** be the (known) size of the genome, **E** the number of genes (known) in the genome belonging to “f” such that ***p*** = E/N, and **n** the size of the studied sample, the **enrichment** test gives *a posteriori* the probability under H0 of having obtained in the sample a number equal to or greater than the actually observed number **e** of genes associated with “f”:

$$P(X \geq e) = 1 - \sum_{k=1}^{e-1} \frac{\binom{E}{k}\binom{N-E}{n-k}}{\binom{N}{n}}$$

If $P(X \geq e) < 5\%$ : “f” is over-represented in the sample, and the sample studied is said to be “enriched” for the function “f”.

If the enrichment test is performed for several functions represented in the sample, this represents several hypotheses being tested and the p-values must be corrected to control the false positive rate (Bonferroni or Benjamini-Hochberg).

```
e = 5; n = 150; E = 28; N = 2700
phyper(e-1, E, N-E, n, lower.tail= FALSE)
fisher.test(matrix(c(e, E-e, n-e, N-E-n+e), 2, 2),
                              alternative='greater')$p.value
```

# 3/ Normality tests

**Normality assumptions** are often required in statistical analyses:

- Confidence intervals
- Student's t-test, ANOVA
- Linear regression (normality of residuals)
- *Etc.*

It is then necessary to **check these hypotheses** either by a **graphical approach**:

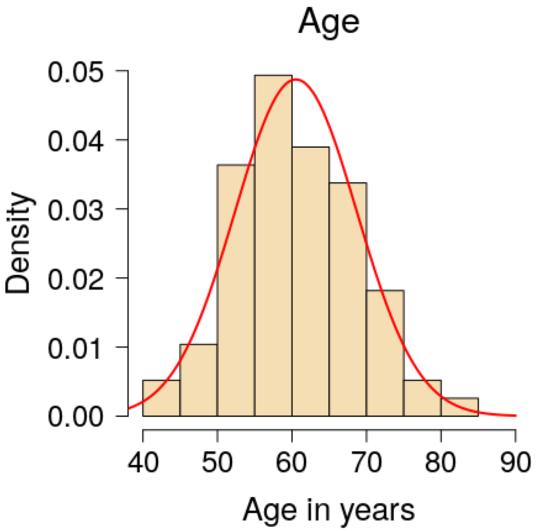
- Superposition of the Gaussian density on the histogram of observations
- Quantile-Quantile plot (Henry's line)

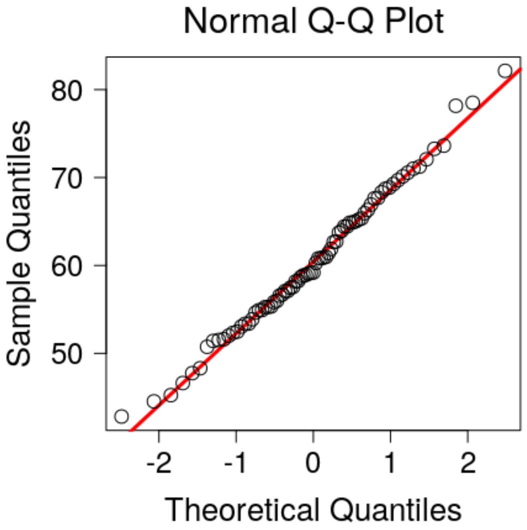
or by using a **goodness-of-fit test to the normal distribution**:

- Shapiro-Wilk test (n < 50)
- Kolmogorov-Smirnov test (n > 50)

# Example: Normality tests

Ages and cognitive scores (MMSE) of 77 subjects in a clinical study (simulated data):





```
ks.test(age,
"pnorm",
mean(age),
sd(age))
```

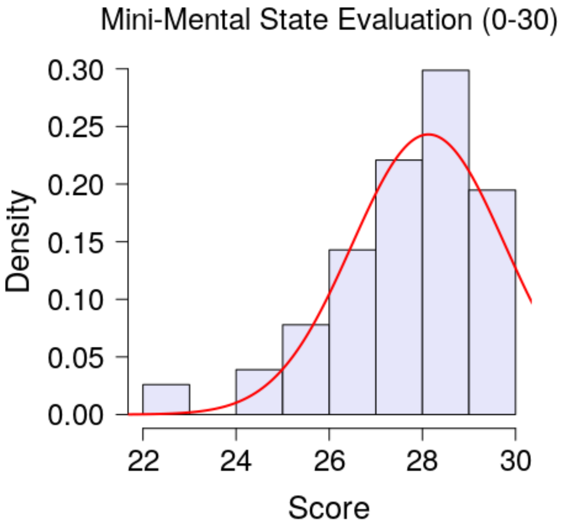
**One-sample Kolmogorov-Smirnov test**

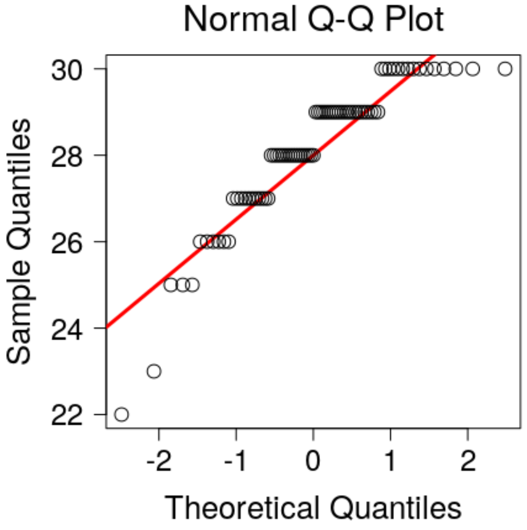
**data: age**

D = 0.060703, **p-value = 0.9226**

alternative hypothesis: two-sided

**Non rejection of the null hypothesis of normality**





```
ks.test(mmse,
"pnorm",
mean(mmse),
sd(mmse))
```

**One-sample Kolmogorov-Smirnov test**

**data: mmse**

D = 0.19552, **p-value = 0.005551**

alternative hypothesis: two-sided

**Rejection of the null hypothesis of normality**

# 4/ Non-parametric tests

A **nonparametric test** (NP) is a test that does not require any assumptions about the distribution of the data (**distribution-free test**). It is generally based on the study of the ranks of the observations, without depending on the means and variances estimated in the original data.

**Pros:**

- Applicable when certain conditions required for a parametric test are not met (*e.g.* normality, equality of variances...)
- Tests suitable for small samples (n < 30)
- Tests suitable for ordinal variables (*e.g.* degree of satisfaction)

**Cons:**

- When the distribution conditions are well satisfied: NP tests are less powerful than parametric tests
- Difficult to interpret because we no longer compare parameters such as means, proportions or variances...

☞ ***Most parametric tests have equivalent non-parametric tests.***

# Nonparametric rank tests

In the case of small samples and non-Gaussian distributions (see goodness-of-fit tests), one test strategy is to **replace the values of the observations by their ranks**:

**Comparison of 2 independent samples:**

- **Wilcoxon-Mann-Whitney test**
  (also called **Mann-Whitney U test** or **Wilcoxon sum-rank test**)

**Comparison of 2 paired samples:**

- **Wilcoxon signed-rank test**

**Comparison of more than 2 samples:**

- **Kruskal-Wallis test (+ Dunn's post hoc test)**

***Taken from***
***https://www.math.univ-toulouse.fr/~besse/Wikistat/pdf/st-l-inf-tests.pdf***

# Example: Wilcoxon-Mann-Whitney test

**Goal**: We want **to compare 2 groups A and B of patients** whose cytorachy (presence of cells in cerebrospinal fluid measured in number of cells per µL) was assessed.

| groupe | B | B | B | A | B | A | B | B | B | B | B | B | A | B | B | A | B | B | A |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| cytorachy | 5 | 6 | 7 | 8 | 8 | 9 | 10 | 10 | 11 | 14 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 26 |
| rank | 1 | 2 | 3 | 4.5 | 4.5 | 6 | 7.5 | 7.5 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 |

| group | B | A | B | B | A | A | A | A | A | B | B | A | A | A | B | A | A | A | A |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| cytorachy | 27 | 34 | 35 | 40 | 41 | 45 | 49 | 84 | 85 | 92 | 100 | 154 | 160 | 173 | 200 | 348 | 480 | 560 | 612 |
| rank | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 | 34 | 35 | 36 | 37 | 38 |

**Principle of the test:** *Under H0 the values of A (red) and B (blue) ordered on the rows are homogeneously mixed (groups A and B are equivalent).*

```
wilcox.test(valA,valB)
```

```
Wilcoxon rank sum test with continuity correction
data: valA and valB
W = 280.5, p-value = 0.003457
alternative hypothesis: true location shift is not equal to 0
```

***Example from T. Ancelle, Statistique épidémiologie 3ème édition, Maloine, 2012***

# Nonparametric permutation tests

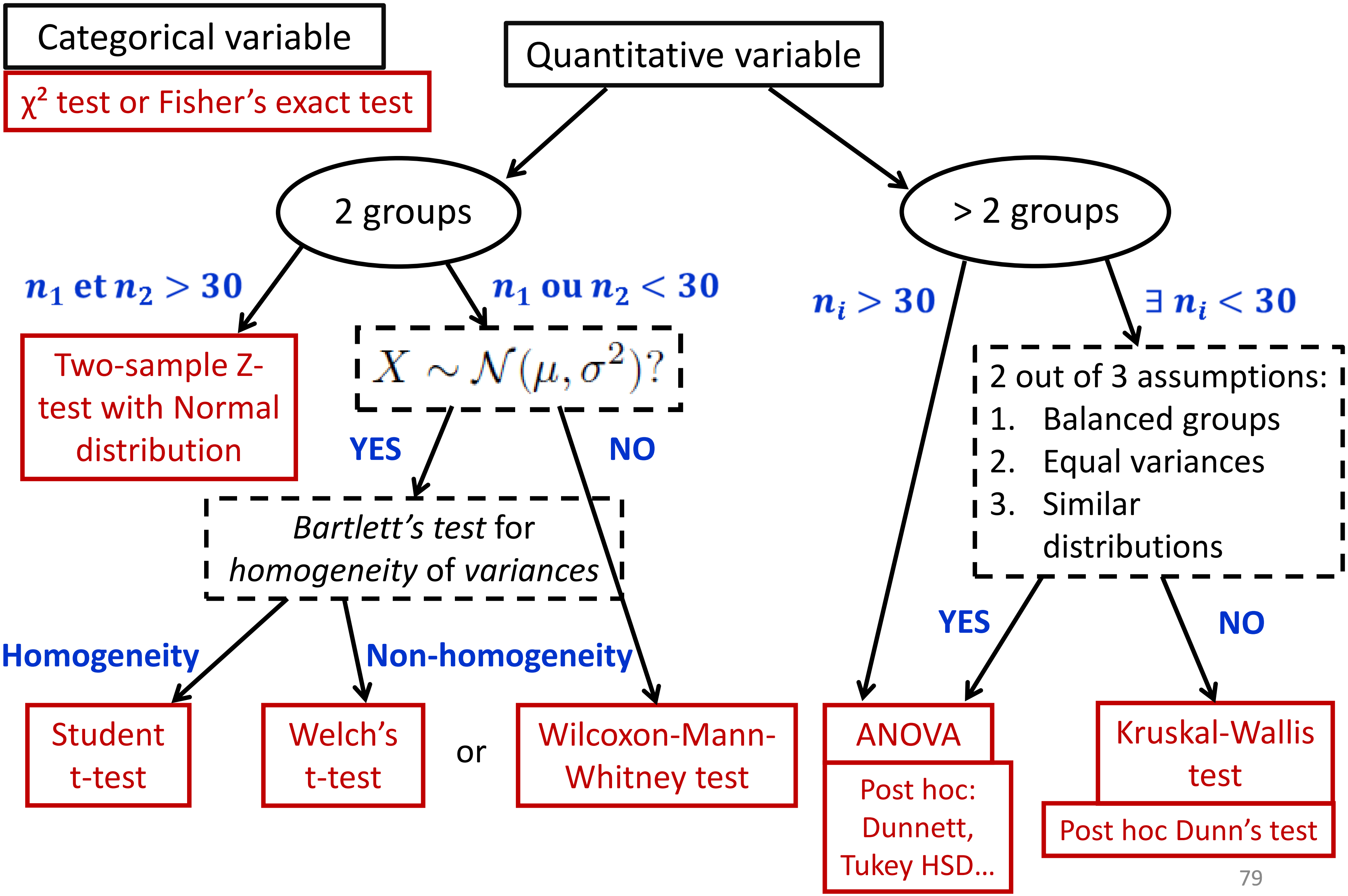
In the case of non-Gaussian distributions, **permutation tests** are robust approaches based on the **resampling of observations**. Thus, the significance will no longer be based on the theoretical distribution of the statistic under H0, but on an **empirical distribution** to be calculated from a large number of permutations (100, 1000 or more).

**Example:** to compare the means of 2 groups by a two-sided permutation test, the hypothesis H0 of "equality of the 2 means" is equivalent to the assumption that *"all observations are interchangeable between the 2 groups"*.

The test procedure is then as follows:

1. Calculate the true (non-permuted) value $T_0$ of the test statistic
2. Generate a large number of random permutations of individuals between the 2 groups
3. Calculate the value of the test statistic T for each permutation
4. Determine the distribution and quantiles 2.5% et 97.5% of the "permuted" Ts under $H_0$
5. Compare $T_0$ to empirical quantiles and conclude

# Comparison of groups



Categorical variable

$\chi^2$ test or Fisher's exact test

Quantitative variable

2 groups

$n_1$ et $n_2 > 30$

Two-sample Z-test with Normal distribution

$n_1$ ou $n_2 < 30$

$X \sim \mathcal{N}(\mu, \sigma^2)$?

YES

*Bartlett's test* for *homogeneity* of *variances*

Homogeneity

Student t-test

Non-homogeneity

Welch's t-test

or

NO

Wilcoxon-Mann-Whitney test

> 2 groups

$n_i > 30$

ANOVA

Post hoc: Dunnett, Tukey HSD…

$\exists\ n_i < 30$

2 out of 3 assumptions:

1. Balanced groups
2. Equal variances
3. Similar distributions

YES

NO

Kruskal-Wallis test

Post hoc Dunn's test

# Part 3: Data modeling

# Regression and modeling

In this section on **regression models**, we are particularly interested in the problem of a "dependent" variable Y ("to be explained"), which is observed alongside one or more "independent" variables X ("explanatory variables" or "regressors") within the same individuals.

The **modeling problem** can then be defined as the search for a simplified representation of Y using the variables X in order to describe it, explain it or predict its values.

From a mathematical point of view, the aim is to determine a function f according to a predefined criterion of fit, capable of approximating the values of Y from the observed values of X:

$$Y = \widehat{f}(X) + \varepsilon$$

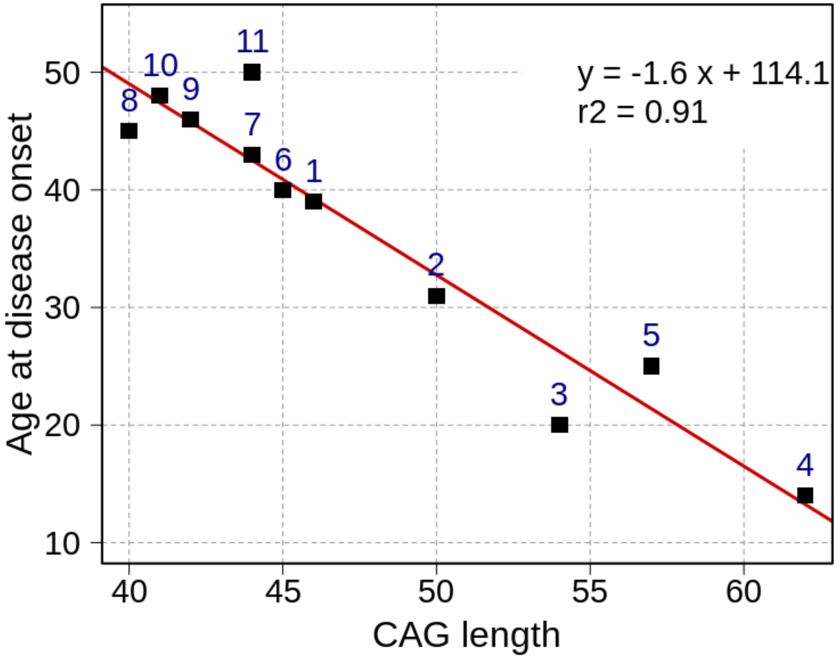
where *ε* represents the noise or the (random) error term.

# Simple linear regression

**Simple regression example:** we aim to explain the age of onset of the first symptoms of 11 patients with spinocerebellar ataxia SCA1 from the polyglutamine length (CAG repeats).

**Explanatory variable** **Variable to be explained**

| ID | CAG length | Age at onset |
|---|---|---|
| **1** | 46 | 39 |
| **2** | 50 | 31 |
| **3** | 54 | 20 |
| **4** | 62 | 14 |
| **5** | 57 | 25 |
| **6** | 45 | 40 |
| **7** | 44 | 43 |
| **8** | 40 | 45 |
| **9** | 42 | 46 |
| **10** | 41 | 48 |
| **11** | 44 | 50 |



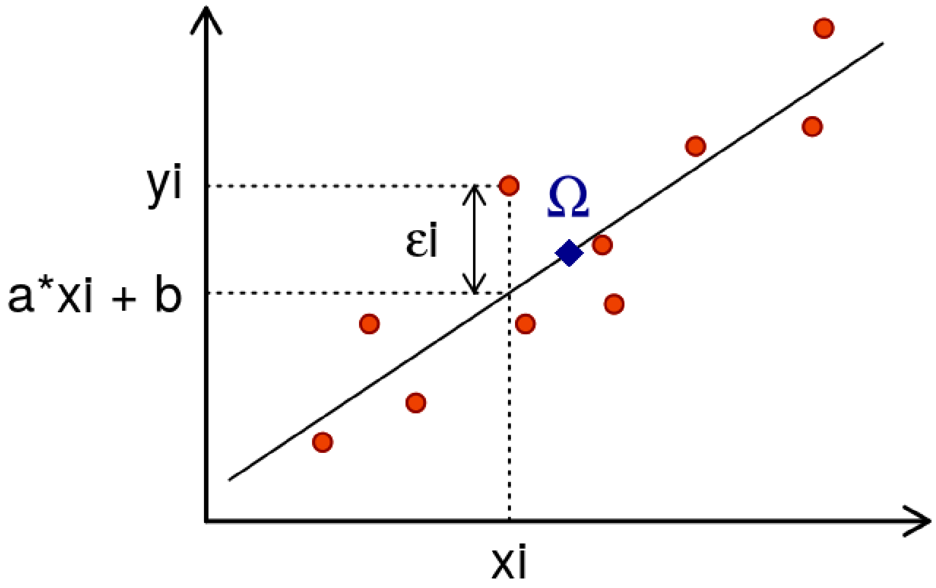
**Simple regression model:** $\boxed{y_i = a \times x_i + b + \varepsilon_i}$

$\varepsilon_i$ : independent identically distributed error terms from the Normal distribution $\mathcal{N}(0, \sigma_\varepsilon^2)$

**Problem: How to estimate ɑ, b and $\sigma_\varepsilon$?**

***Source: BIOSCA project***

# Ordinary least squares (OLS) estimators



The **OLS criterion** consists in finding the values of a and b that minimize the sum of the squared errors:

$$
\begin{aligned}
S &= \sum_{i=1}^{n} \varepsilon_i^2 \\
&= \sum_{i=1}^{n} \left(y_i - a \times x_i - b\right)^2
\end{aligned}
$$

**OLS estimates:**

$$
\begin{cases}
\hat{a} = \dfrac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2} = \dfrac{\widehat{\mathrm{Cov}}(X,Y)}{\widehat{\mathrm{Var}}(X)} = \mathrm{r}_{XY}\,\dfrac{\hat{\sigma}_Y}{\hat{\sigma}_X} \qquad \text{(slope of the regression line)} \\
\hat{b} = \bar{y} - \hat{a} \times \bar{x} \qquad \text{(the line goes through the center of gravity } \Omega \text{ of the scatterplot)}
\end{cases}
$$

**Predicted values of $y_i$:** $\hat{y}_i = \hat{a} \times x_i + \hat{b}$ **Residuals:** $\hat{\varepsilon}_i = y_i - \hat{y}_i$

# Analysis of variance and coefficient of determination

As the OLS criterion $0 \leq S \leq +\infty$, we need a more meaningful criterion to check the quality of the regression (**"goodness of fit"**).

For this, we use **the variance decomposition formula**:

$$\boxed{\begin{aligned} \sum_{i=1}^{n}(y_i - \bar{y})^2 &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \\ \mathrm{SST} &= \mathrm{SSE} + \mathrm{SSM} \end{aligned}}$$

***SST****: Sum of Squares Total*

***SSE****: Sum of Squares Error*

**SSM**: *Sum of Squares Model*

We then define the **coefficient *R*² of determination**:

$$\boxed{R^2 = \frac{SSM}{SST} = 1 - \frac{SSE}{SST}}$$

*R*² close to 1, the model is excellent

*R*² close to 0, the model is useless!

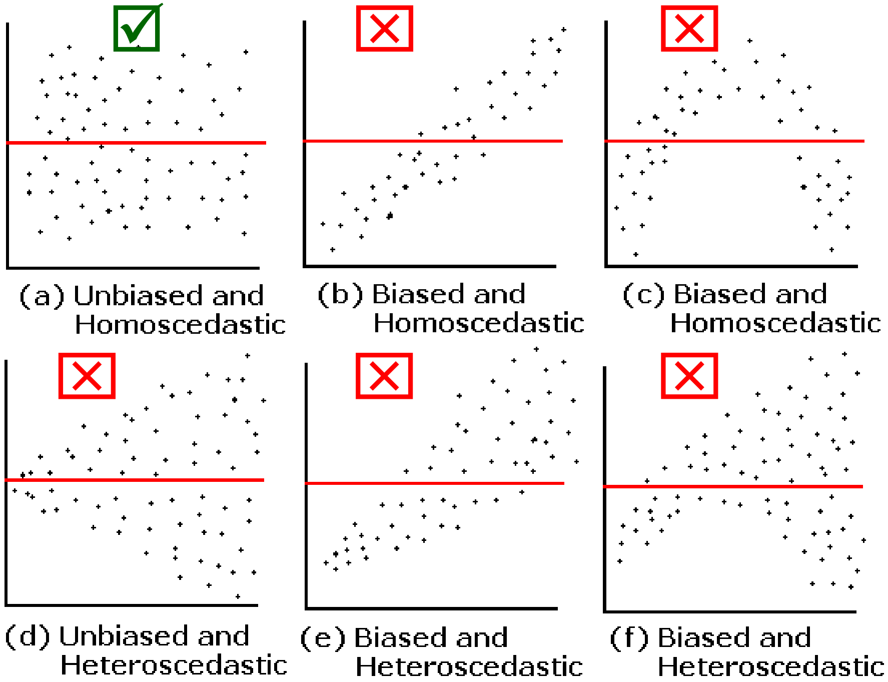
and the **multiple linear correlation coefficient**: $R = \sqrt{R^2}$

In case of a simple model with 1 regressor, we have: $\mathrm{r}_{XY} = sign(\hat{a}) \times R$

# Validation of the model

In order to **validate the regression model**, it remains to check the hypotheses on the residuals (i.i.d. and Gaussian terms). This verification is usually done by visual examination of graphs

- Plot of residuals versus predicted values: the scatterplot should not show any particular structure (independence, homoscedasticity/constant variance, normality);
- Normality: Histogram and QQ-plot (points aligned on a straight line);
- Independence: autocorrelation of residuals (acf), Durbin Watson test…



(a) Unbiased and Homoscedastic

(b) Biased and Homoscedastic

(c) Biased and Homoscedastic

(d) Unbiased and Heteroscedastic

(e) Biased and Heteroscedastic

(f) Biased and Heteroscedastic

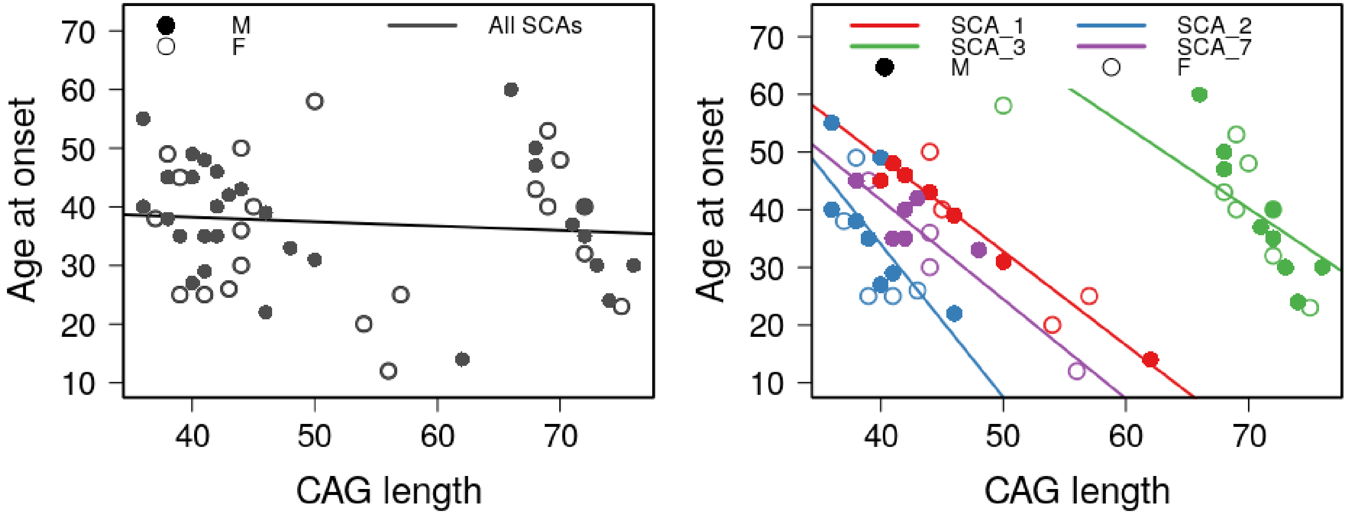
***Figures. Plots of residuals (y axis) vs predicted values (x axis).* *Case (a) validates the linear regression. Cases (b-f) indicate either a heteroskedasticity issue (d-e-f) or a dependency structure that is not taken into account by the model (b-c-e-f).***

***Solutions: use of “robust” OLS estimators, data transformation, data fitting by other types of non-linear models.***

***J. Faraway, Linear Models with R, Chapman & Hall 2004***

# Interaction effect: beware of appearances...

Not taking into account the different genetic forms of SCA would lead to the (somewhat hasty) conclusion that the length of polyglutamine does not influence the age of onset of the disease!



| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| Model 1: All SCAs | -0.075 | 0.112 | -0.67 | 0.50597 (ns) |
| Model 2: SCA_1 | -1.626 | 0.168 | -9.679 | 0.000005 (***) |
| Model 3: SCA_2 | -2.653 | 0.826 | -3.213 | 0.008267 (**) |
| Model 4: SCA_3 | -1.426 | 0.331 | -4.315 | 0.000613 (***) |
| Model 5: SCA_7 | -1.722 | 0.257 | -6.703 | 0.000152 (***) |

*Table: Estimation and significance of regression slopes*

*Source: BIOSCA project*

# Linear model

For a single **quantitative dependent variable** linked with **one or more quantitative and/or categorical explanatory variables**, it is worth noting that linear regression and ANOVA are special cases of the same statistical model called **linear model**:

a) 1 quantitative explanatory variable: **simple linear regression**

b) ≥ 2 quantitative explanatory variables: **multiple linear regression**

c) 1 categorical variable: **t-test (2 levels)** or **one-way ANOVA**

d) ≥ 2 categorical variables: **multi-way ANOVA**

e) Combination of quantitative and categorical variables: **ANCOVA**

In general, the use of the linear model also allows us to assess the Y-X relationship in the presence of *covariates* (quantitative) or *cofactors* (qualitative). An "adjusted" model facilitates the disentanglement of the effects of covariates or cofactors from the variable of interest, determined with an **adjusted p-value**.

# Simple cases of non-linear regressions

Models based on the “exponential” and “power” functions are 2 simple cases of **non-linear relationships**, as they can be **linearized** using the logarithm function:

$$\underline{\text{Power model}}\text{ of type: }\boxed{y = B \cdot x^{\alpha}},\ (B, \alpha) \in \mathbb{R}^2.$$

$$\text{Logarithm transformation: } \ln y = \ln B + \alpha \ln x$$

$$\Leftrightarrow Y = \alpha X + \beta \text{ by setting } Y = \ln y,\ X = \ln x \text{ and } \beta = \ln B.$$

$$\underline{\text{Exponential model}}\text{ of type: }\boxed{y = B \cdot \mathrm{e}^{\alpha x}},\ (B, \alpha) \in \mathbb{R}^2.$$

$$\text{Logarithm transformation: } \ln y = \ln B + \alpha x$$

$$\Leftrightarrow Y = \alpha X + \beta \text{ by setting } Y = \ln y,\ X = x \text{ and } \beta = \ln B.$$

# Generalized linear model

***How to predict a variable Y with discrete or categorical values from an explanatory variable X that can be either discrete or continuous?***

The **generalized linear model** (GLM) includes the linear regression model, as well as other interesting models allowing, for example, to predict a ***dependent variable Y with discrete values*** from 1 or more continuous or discrete explanatory variables.

**Example 1 – Y binary variable 0/1:** we wish to explain the absence (Y = 0) or presence (Y = 1) of coronary heart disease as a function of age in 100 subjects.

**Example 2 – Y counting variable:** we wish to quantify the evolution of a number of bacteria as a function of time.

**Example 3 – Y variable of life span:** we wish to study the life span in weeks from diagnosis to death as a function of the white blood cell count (log10 scale) at baseline for 33 patients with leukemia.

# Formulation of GLM

Generalized linear models are composed of **3 components**:

- **Random component**: Dependent variable Y associated with a probability distribution
- **Deterministic component**: Linear predictor or linear combination of explanatory variables $X_1$, …, $X_k$: $\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k$
- **Link function**: Function *g* describing the functional relationship between the linear predictor and the mathematical expectation of the dependent variable Y

**GLM:** $$\boxed{g\left(\mathbb{E}(y|x_1, \ldots, x_k)\right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k}$$

| Distribution | Type of data | Type of GLM | Link function (g) |
|---|---|---|---|
| Normal | Var. ~ Normal | Linear model | Identity: g(y) = y |
| Poisson | Count | Log-linear model | Log: g(y) = log(y) |
| Binomial | Percentage | Logistic regression | Logit: g(y) = log(y/(1-y)) |
| Gamma | Duration | Gamma model with inverse link function | Inverse: g(y) = 1/y |

# Estimation of parameters

Unlike the case of linear regression where the estimation of the coefficients is based on the ordinary least squares method ***(i.e. minimizing the squared differences between the observed response and the predicted response)***, the parameters of the GLM are determined by another estimation method called the **maximum likelihood method**.

**Principle of the maximum likelihood method:**

Find the parameter values that maximize the **"probability that the observed values of Y are realized conditionally on the assumed known parameters"**:

$$\mathrm{Prob}(y_1, \ldots, y_n | \beta_0, \beta_1, \ldots, \beta_k)$$

The software calculates the estimates using an iterative algorithm (Fisher scoring or Newton-Raphson) where starting from an initial (fixed) value of the parameters, the value is updated at each iteration until convergence of the algorithm.

# Case of binary logistic regression

The interpretation of the logistic model is relatively simple because its expression involves an “Odds” value corresponding to the ***ratio of the chances between events Y = 1 and Y = 0 knowing the value of X***:

$$\ln\left(\frac{\mathbb{P}(Y=1|X=x)}{1-\mathbb{P}(Y=1|X=x)}\right)=\beta_0+\beta_1 x \iff \frac{\mathbb{P}(Y=1|X=x)}{\mathbb{P}(Y=0|X=x)}=\mathrm{e}^{\beta_0+\beta_1 x}$$

The coefficient $\mathrm{e}^{\beta 1}$ can then be interpreted in terms of the **Odds-Ratio** (OR) reflecting the change in the ratio of the chances of the event Y = 1 versus Y = 0 occurring when X goes from x to x + 1:

$$OR=\frac{\mathbb{P}(Y=1|X=x+1)/\mathbb{P}(Y=0|X=x+1)}{\mathbb{P}(Y=1|X=x)/\mathbb{P}(Y=0|X=x)}=\mathrm{e}^{\beta_1}$$

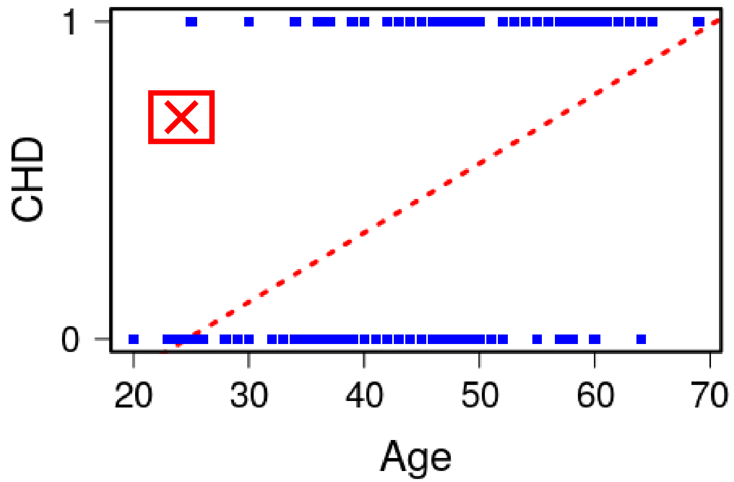
3 tests are available to evaluate the contribution of the variable X to the model:

- Wald test
- Likelihood ratio test (LRT)
- Score test

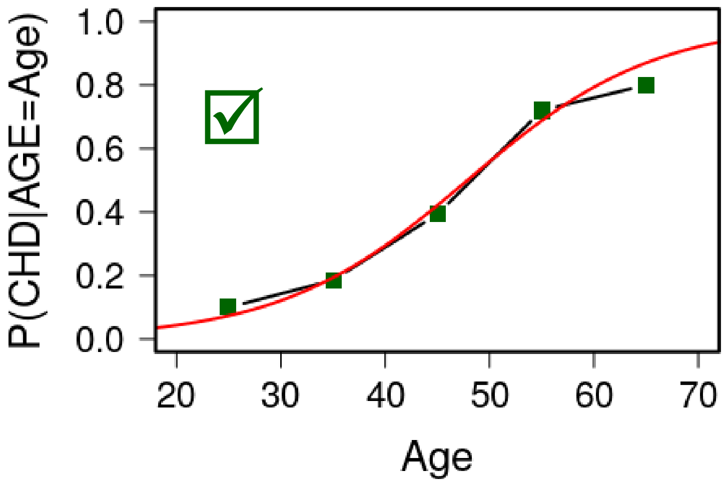
$$\begin{cases} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{cases}$$

# Example 1: Binary logistic regression

Goal: to study the relationship between age and the presence (1) or absence (0) of coronary heart disease (CHD) in a population of 100 individuals.



*Left figure: no linear fit is possible between age and the binary CHD variable.*



*Right figure: the sigmoidal shape (S-curve) of the relationship between age and CHD conditional expectation is well fitted by the logit function.*

```
summary(glm(dat$CHD ~ dat$AGE,
family = binomial(link = logit)))
```

$$\mathbb{P}(\mathrm{CHD} = 1|\mathrm{AGE} = \mathrm{Age}_i) = \frac{\exp(-5.31 + 0.11 \times \mathrm{Age}_i)}{1 - \exp(-5.31 + 0.11 \times \mathrm{Age}_i)}$$

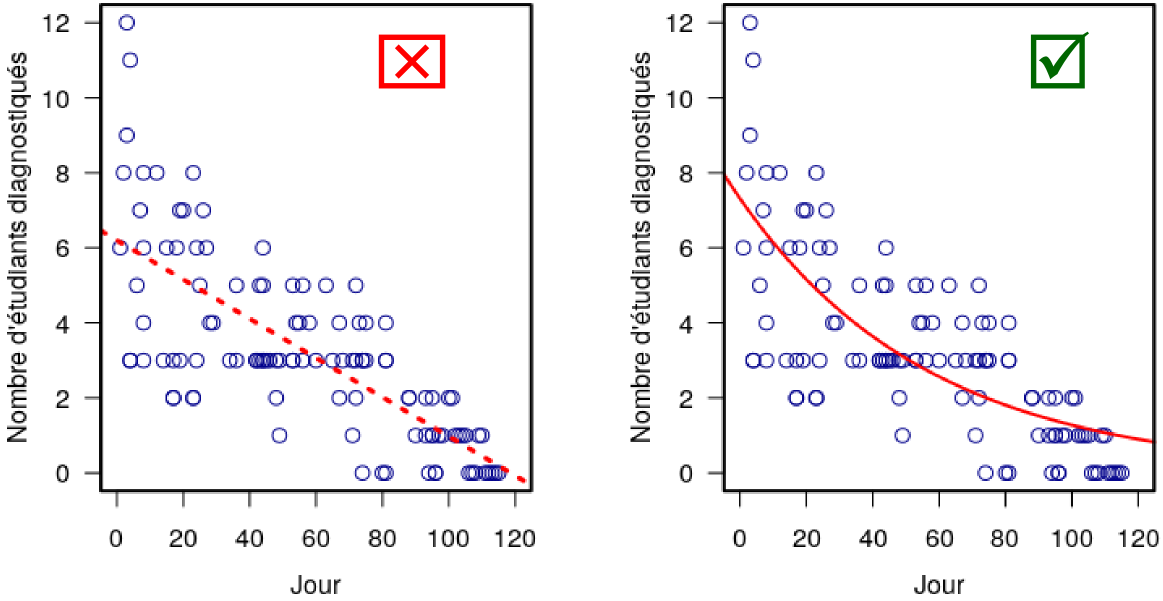
Wald test

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.30945    1.13365  -4.683 2.82e-06 ***
dat$AGE      0.11092    0.02406   4.610 4.02e-06 ***
```

*Example from Hosmer and Lemeshow, Applied Logistic Regression, Wiley, 2nd edition (2000)*

# Example 2: Poisson regression

Goal: to study the number of students diagnosed with an infectious disease since the first day of the epidemic.



```
summary(glm(formula = Students ~ Days, family = poisson, data = cases))
```

**Coefficients:**

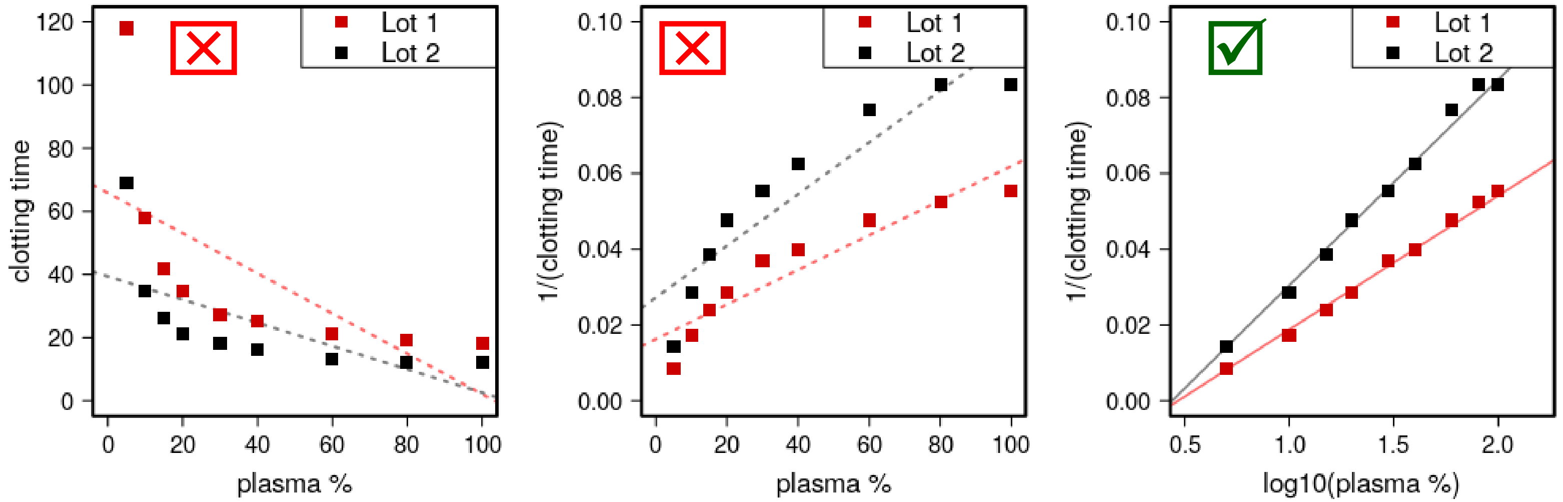
```
             Estimate   Std. Error   z value   Pr(>|z|)
(Intercept)  1.990235   0.083935      23.71    <2e-16 ***
Days        -0.017463   0.001727     -10.11    <2e-16 ***
```

**The average count of new cases as a function of time is described by the model:**

$$\boxed{\hat{\mu}_t = e^{1.99-0.02\times t}}$$

# Example 3: Gamma regression

Goal: to study the blood clotting time (sec) when thromboplastin (2 batches tested) is added to plasma at 9 different concentrations (%).



***Figures. Steps of modeling by a Gamma model using the inverse link function.***

```
summary(glm(lot1 ~ log10(u), data = clotting, family = Gamma(link = inverse))
```

**Lot 1 - Coefficients:**

```
             Estimate  Std. Error   t value  Pr(>|t|)
(Intercept) -0.0165544  0.0009275  -17.85 4.28e-07 ***
log10(u)     0.0353288  0.0009555   36.98 2.75e-09 ***
```

**Lot 2 - Coefficients:**

```
             Estimate  Std. Error   t value  Pr(>|t|)
(Intercept) -0.023908   0.001326  -18.02 4.00e-07 ***
log10(u)     0.054339   0.001328   40.91 1.36e-09 ***
```

**Lot 1:** 1/E(clotting|x) = -0.017+0.035 x **Lot 2:** 1/E(clotting|x) = -0.024 + 0.054 x x = log10(plasma)

***McCullagh P. and Nelder, J. A. (1989) Generalized Linear Models. London: Chapman and Hall***
***Data from Hurn et al (1945), J Lab & Clin Med.***

# Concept of repeated measures

- *Repetition of measurements of a variable on the same individuals in a sample. It may involve successive data collected over time (**longitudinal studies**).*

The difficulty with repeated measures is that they introduce a correlation between the measures coming from the same individual (source of intra-subject variability). In a data modeling problem, the **autocorrelation between the residuals** resulting from these repeated measurements goes against the fundamental assumption of independence of errors on which the linear model is based.

**Pros of repeated measurements:**

- useful to study the dynamics of a phenomenon
- more observations available for the same number of individuals

**Cons (loss of independence of observations):**

- 2 observations from different individuals are independent, but 2 observations from the same individual are not independent!

# Mixed effects models

In addition to the case of repeated measurements, one can be faced with situations where several observations are “naturally grouped” in a study.

***Example. Family studies involving several members of the same family, studies conducted on twins, multi-center clinical studies where data are collected on patients from different hospitals, same lines of animals, etc.***

Compared to the classical (generalized) linear model, the use of more complicated models, called **mixed-effects models**, allows the combination of “fixed” effects with “random” effects to take into account the correlation of individuals “grouped” together (*e.g.*, patients from the same center), the intra-individual correlation of longitudinal data, or a mixture of the two (measurements over time for the same subjects from different groups of subjects).

# Fixed and random effects

It is difficult to find consensual definitions for the **fixed and random effects** of a modeling problem, so a simplified presentation is given here to help in applications.

**Fixed effect:** the data come from all possible levels of a qualitative variable whose specific impact on the response variable is to be studied.

Example: in a study comparing performance on a cognitive test (response variable) between a group of patients and a group of healthy subjects, the group factor is a fixed effect whose impact on the response variable we wish to study.

**Random effect:** the data come only from a random sample of all possible levels of a qualitative variable (typically a grouping factor) whose specific impact on the response variable is not of interest, but whose effect must be controlled for in the model.

Previous example: if the data come from 5 hospitals, the center may constitute a random effect, ① because the study does not involve all hospitals in France, and ② because some experimental conditions could vary from one center to another.

# From modeling to prediction…

To proceed from modeling to **prediction**, we must first study the **generalization property** of the model; that is, the ability of the model to make robust predictions on new data.

To evaluate this capacity, we need 2 independent data sets:

- **1 training set**: dataset used to “train the model”
- **1 test set**: independent prediction dataset to evaluate the prediction error of the model (according to a predefined criteria)

The learning step must be done in such a way as to avoid as much as possible the two phenomena of **overfitting** and **underfitting**.

Overfitting reflects a situation where the model fits the training data too well, making it less flexible to apply to new data. Conversely, underfitting occurs when the model is not complex enough to correctly describe the relationship between the variables based on the training data.

**TO BE CONTINUED…**

**With the biostatistics team of the Brain Institute:**

Sana Rebbah

Baptiste Crinière-Boizet

Gaspard Martet