Éléments de statistiques descriptives et inférentielles pour la biologie

François-Xavier Lejeune
f-x.lejeune@icm-institute.org

Année 2023-2024

Plan du cours

- **Introduction**
- **Partie 1 : Décrire**
- **Partie 2 : Tester**
- **Partie 3 : Modéliser**

Introduction

Pourquoi faire des statistiques en biologie ? (1/2)

Variabilité = manque de reproductibilité des mesures :

1. Variabilité métrologique [liée au protocole de mesure]
    - Variations des **conditions expérimentales** (ex. différences de température, d'humidité, de luminosité, plusieurs expérimentateurs impliqués dans le recueil des mesures…)
    - Erreurs induites par l'**appareil de mesure** utilisé
2. Variabilité biologique [liée au sujet]
    - **Variabilité intra-individuelle** = *variations de mesures effectuées sur un même sujet (mesures répétées)*
    - **Variabilité inter-individuelle** = *variations de mesures provenant de plusieurs sujets* (ex. différences d'âge, de sexe, de taille, de poids, de pathologie, de constantes biologiques ou de caractéristiques génétiques…)

Pourquoi faire des statistiques en biologie ? (2/2)

Quantité importante des données biologiques :

Besoins de méthodes adaptées à l'analyse et l'intégration de données massives et hétérogènes :

- Clinique, Imagerie (PET Scan, IRM),
- Données comportementales, enregistrements de l'activité électrodermale,
- Enregistrements électrophysiologiques (MEG, EEG, MEA),
- Données omiques (génomique, transcriptomique, protéomique, métabolomique…),
- Histologie (microscopie des tissus biologiques)…

☞ *problèmes de réduction de dimension, sélection de variables…*

Vocabulaire

- **Population** : ensemble des individus ou unités statistiques visés par l'étude (taille N)
- **Échantillon** : sous-ensemble de la population utilisé pour l'étude (taille n << N)
- **Variable X** : trait commun aux individus pouvant être observé ou mesuré

Expérience statistique : elle implique généralement un échantillon tiré de façon aléatoire (étape de ***« randomisation »***) et en nombre suffisamment élevé pour être représentatif de la population visée afin d'étudier un phénomène ou tester une hypothèse.

Population (taille N)

X: p, μ, σ^2

Échantillon

n individus tirés aléatoirement

n observations x_1, …, x_n

Statistique inférentielle

Tests, estimation, modélisation

Statistiques descriptives

Description, génération d'hypothèses

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

Choix de l'analyse statistique appropriée ?

Le choix des méthodes d'analyse dépend essentiellement de

1. **« Ce que l'on cherche dans les données »**
   [☞ question biologique ou objectif(s) d'une étude]

- **Approche exploratoire** pour acquérir de nouvelles connaissances et générer de nouvelles hypothèses à tester…
- **Approche inférentielle** pour vérifier une hypothèse *a priori* ou construire un « modèle prédictif »…

2. **Caractéristiques des données**

- Variables quantitatives et/ou qualitatives + distributions
- Design expérimental : ex. tailles d'échantillon par condition
- Présence de valeurs manquantes et/ou valeurs extrêmes (***« outliers »***)
- Mesures répétées sur le même individu, éventuellement à plusieurs points de l'espace et/ou du temps (étude ***longitudinale***)

Variables quantitatives

La variable est quantitative si ses valeurs correspondent à des quantités mesurables données par des nombres.

Valeurs discrètes = valeurs entières dans un ensemble dénombrable

Exemples (mesures de comptage) :

- nombre de poussées chez les patients atteints de SEP,
- nombre de cellules par unité de surface,
- nombre de mutations dans une séquence d'ADN de 10 kb,
- nombre de mots rappelés à un test de mémoire…

Valeurs continues = infinité de valeurs dans un intervalle réel

Exemples :

- poids, taille, âge,
- dose quotidienne de lévodopa d'un patient parkinsonien,
- dosage sanguin de la glycémie,
- volume d'une région du cerveau en imagerie cérébrale…

Variables qualitatives

La variable est qualitative ou catégorielle (facteur) si ses valeurs ne sont pas des quantités mesurables par des nombres, mais définissent un groupe de catégories appelées modalités ou niveaux.

Variable nominale = pas d'ordre naturel entre les différentes modalités

Exemples :

- sexe : homme/femme,
- statut : fumeur/non fumeur,
- groupe sanguin : A / B / AB / O…

Variable ordinale = les modalités peuvent être ordonnées

Exemples :

- fréquence d'une activité : jamais, rarement, parfois, souvent, très souvent,
- niveau de sévérité d'une douleur : aucune, minime, modérée, sévère, insupportable,
- stade de la maladie d'Alzheimer : de précoce à terminal…

Partie 1 : DÉCRIRE

Résumer et représenter graphiquement l'information contenue dans les données

Analyse descriptive unidimensionnelle

Généralement, on décrit la distribution d'une variable à l'aide de

- **3 critères numériques**
  - Tendance centrale
  - Dispersion
  - Forme (skewness + kurtosis)
- **1 graphique de fréquences**

Présentation des données :

- **Moyenne ± Déviation standard** (distribution symétrique ou normale)
- **Médiane et écart interquartile** (distribution asymétrique)
- **Fréquences et pourcentages** (données qualitatives)

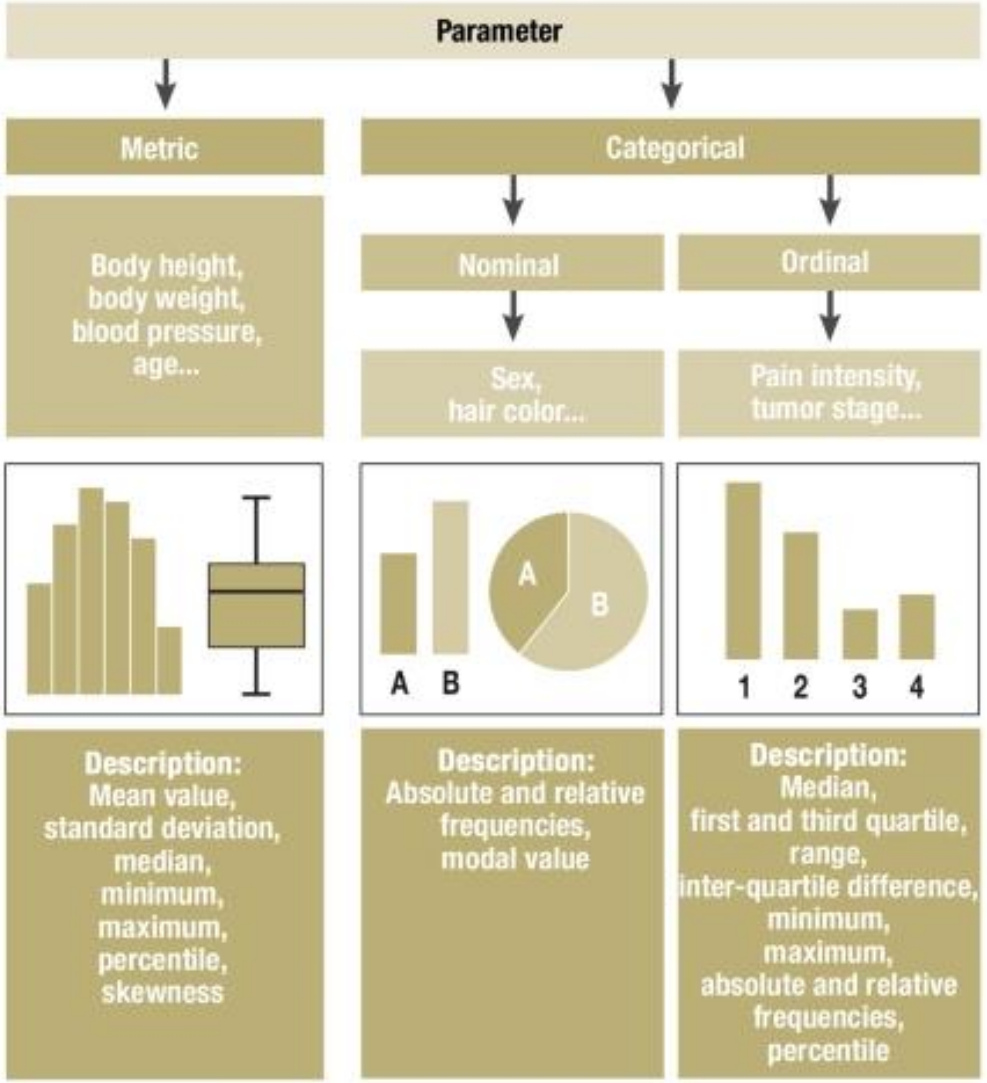


Figure extraite de du Prel, Röhrig, Blettner, Dtsch Arztebl Int 2009

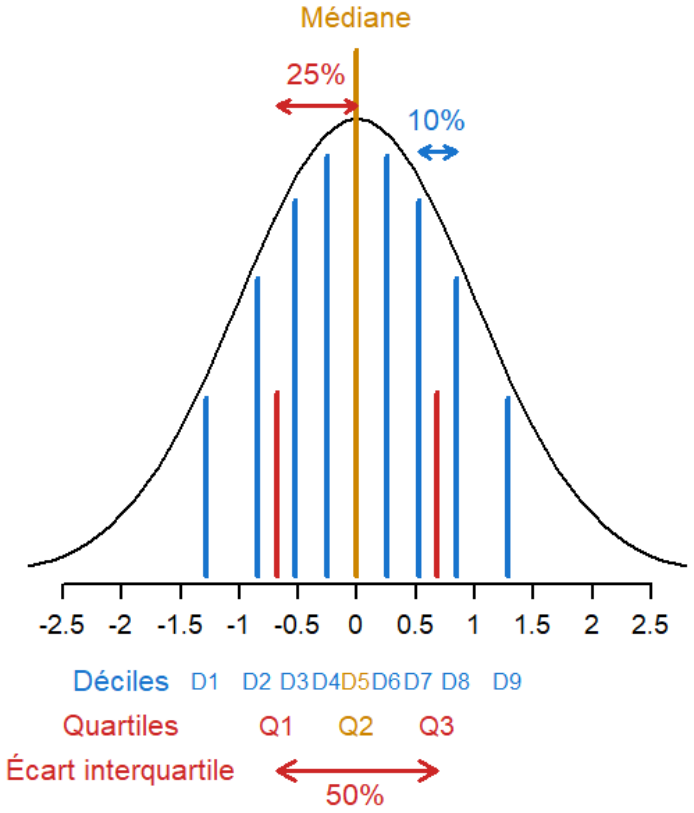
Notion de quantiles

La notion de **quantile empirique** s'applique aux valeurs ordonnées d'une variable quantitative.

Le quantile d'ordre α ($0 \leq \alpha \leq 1$) désigne alors la valeur q_α de la variable telle qu'une proportion α des valeurs de la population soit inférieure ou égale à q_α.

Parmi les quantiles usuels, on peut distinguer :

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



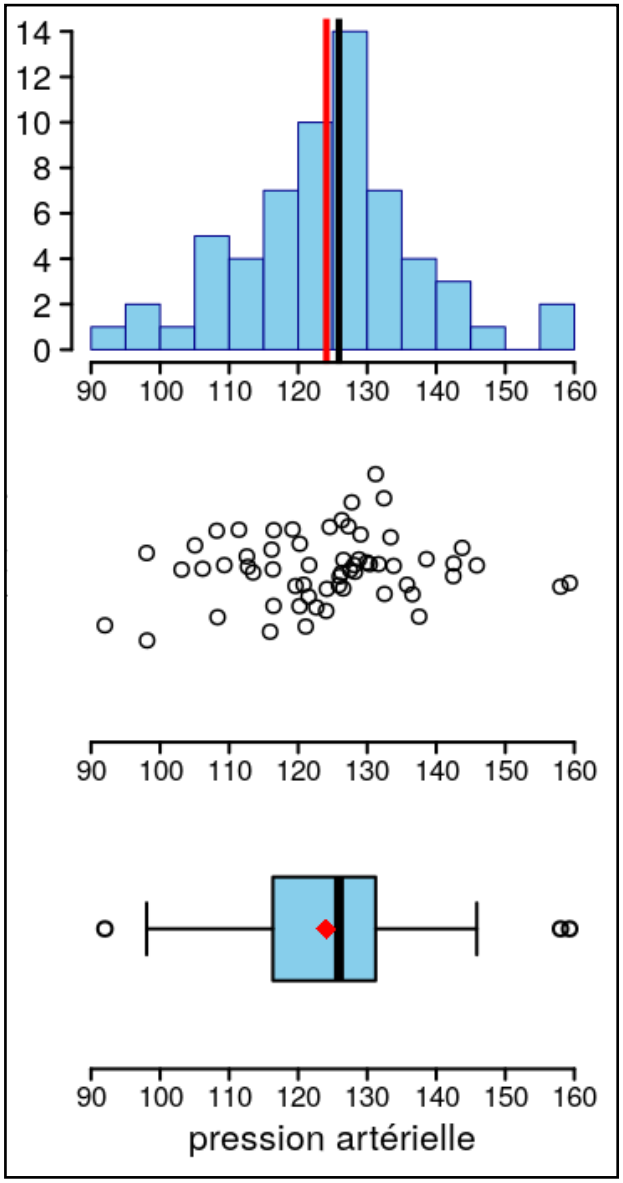
Résumés numériques

Tendance centrale = *valeur autour de laquelle se groupent les observations*

- Moyenne empirique : $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
- Mode
- Médiane Q_2

Dispersion = *étalement des observations autour de la tendance centrale*

- Variance estimée : $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
- Déviation standard (DS) ou écart-type : $s = \sqrt{s^2}$
- Étendue : Max – Min
- Écart interquartile : EIQ = Q3 – Q1
- Coefficient de variation : $\mathrm{CV} = \frac{s}{\bar{x}}$



Écart-type vs Erreur standard de la moyenne

Parmi les valeurs de dispersions, l'**erreur standard de la moyenne (SEM)** est un critère souvent utilisé (parfois incorrectement) à la place de l'écart-type :

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

Il convient donc de bien faire la distinction suivante :

La **DS** d'un échantillon sert à indiquer la variabilité des valeurs au sein de l'échantillon ou de la population (ex. l'âge d'une population de patients).

Par contre, la **SEM** ne reflète PAS la variabilité de l'échantillon mais la variabilité de la moyenne estimée si l'étude était répétée plusieurs fois (écart-type de la moyenne…).

☞ **Quelle que soit la statistique utilisée SEM ou écart-type, celle-ci doit TOUJOURS être spécifiée dans une étude !!!**

Modèles probabilistes

On distingue généralement la **théorie des probabilités** et la **statistique** :

Probabilités : visent à définir des modèles mathématiques (ou lois théoriques) du hasard et à l'étude de leurs propriétés

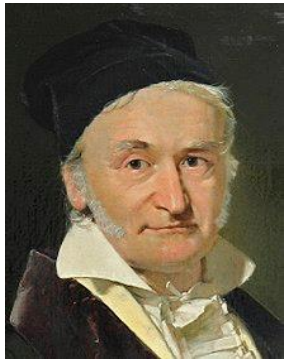
Statistique : vise à confronter ces modèles théoriques aux données réelles

Nombre d'approches dites « paramétriques » reposent ainsi sur le postulat que les données observées sont des réalisations de variables aléatoires issues de lois de distribution connues. Il s'agit alors de choisir, d'ajuster et de valider les modèles probabilistes pouvant servir à tester des hypothèses, prédire ou guider la prise de décisions.

Pour ce faire, nous disposons de plusieurs lois de distributions continues ou discrètes couramment utilisées en pratique : **uniforme**, **normale**, **exponentielle**, **binomiale**, **Poisson**, *etc.*

Loi normale (Laplace-Gauss)

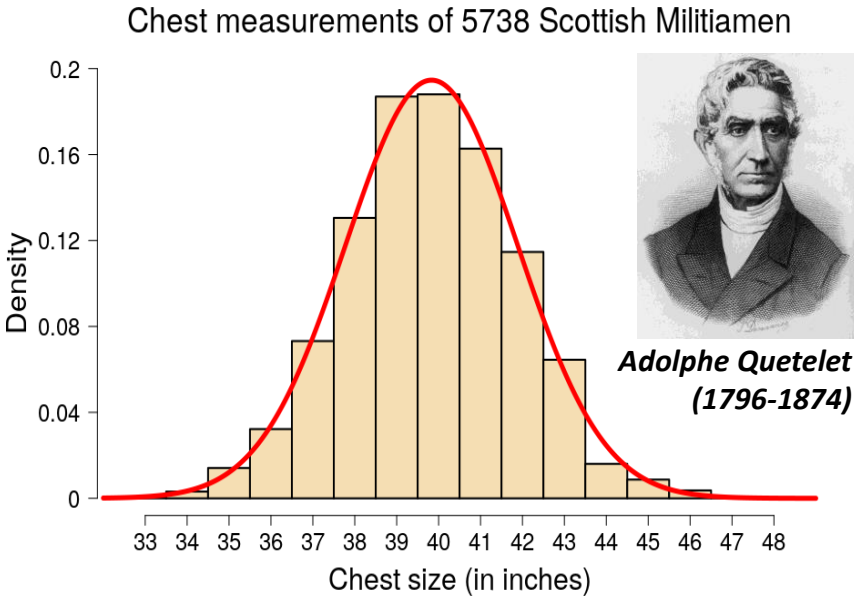
Parmi les lois de probabilité, la **distribution gaussienne** caractérisée par sa courbe « en cloche » est particulièrement importante en pratique car elle permet de représenter la variabilité de nombreux phénomènes naturels (glycémie à jeun, taux de division bactérienne, *etc.*) et la distribution des erreurs de mesures.



Carl Friedrich Gauss (1777-1855)

Ex 1. Planche de Galton (écoulement de billes au travers d'une pyramide de clous)

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



Adolphe Quetelet (1796-1874)

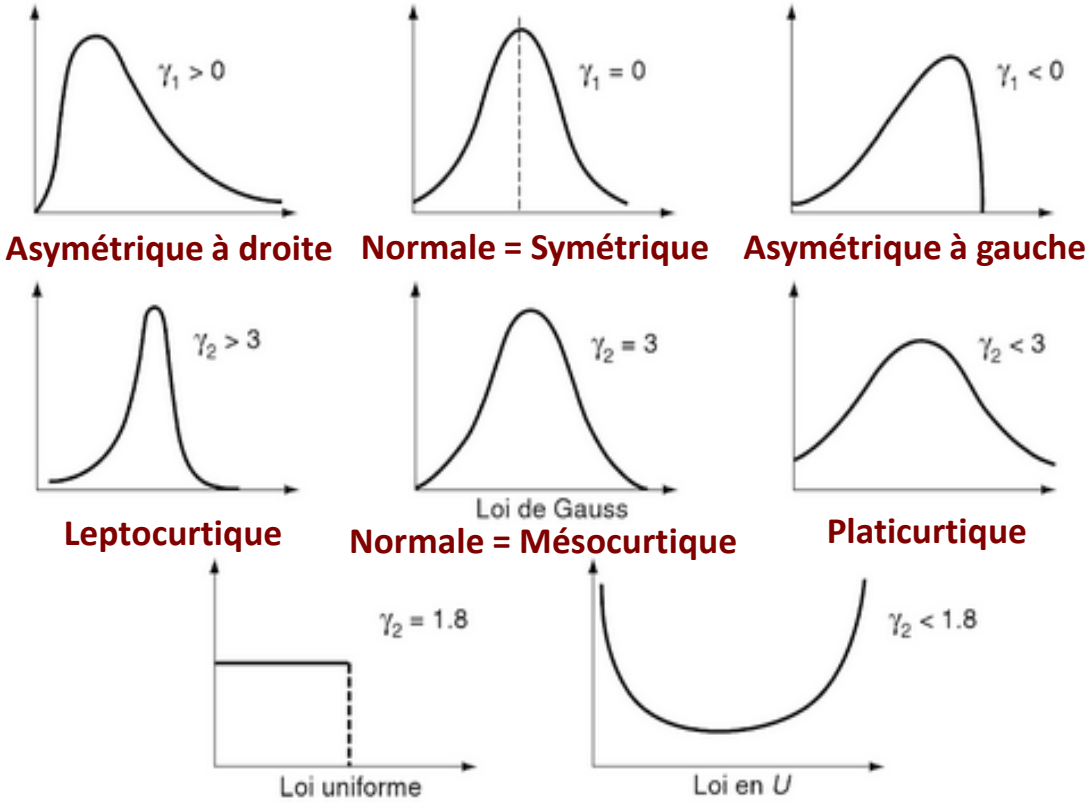
Pierre-Simon de Laplace (1749-1827)

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

Décrire la forme d'une distribution

La densité en cloche de la loi gaussienne sert aussi de référence pour caractériser la forme des autres lois de distribution continues sur les 2 critères suivants :

- **Coefficient de symétrie (skewness)**
- **Coefficient d'aplatissement (kurtosis)**



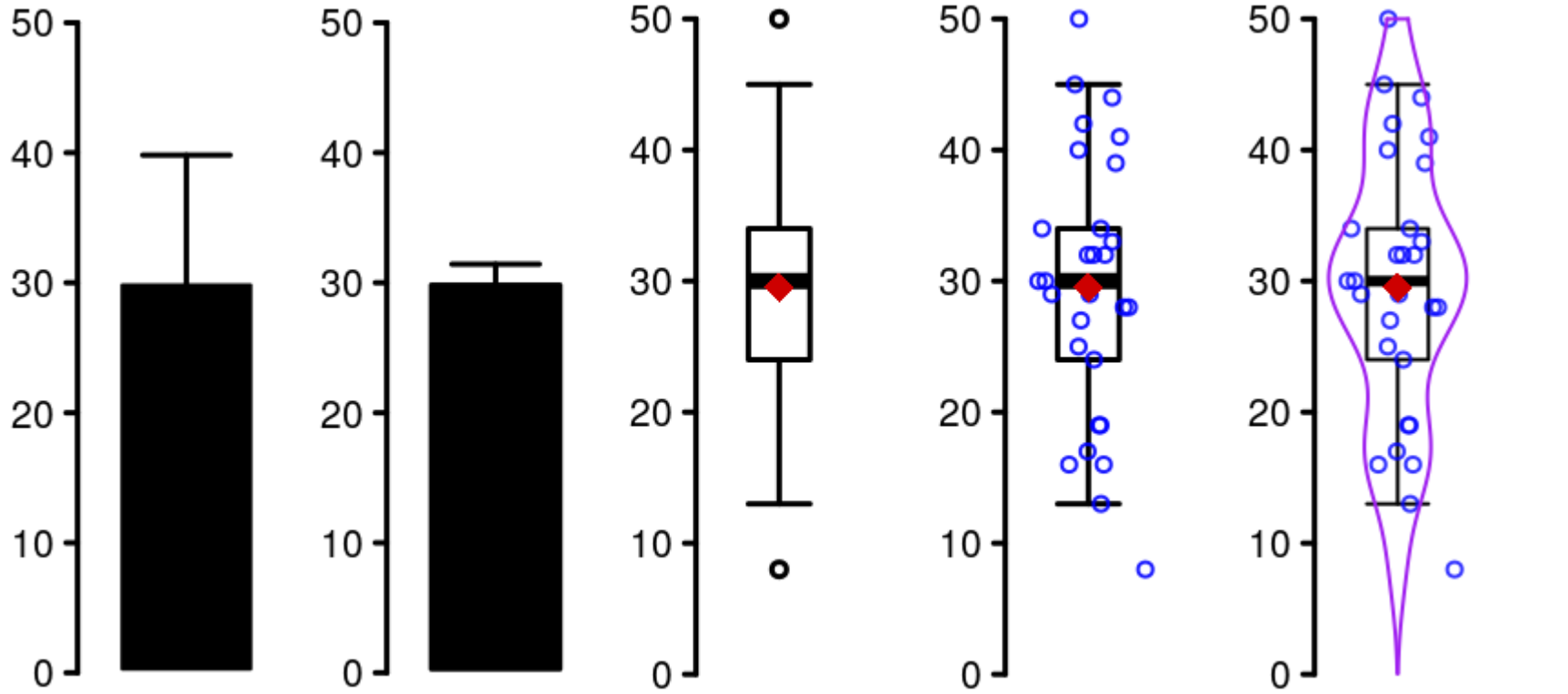
$$\gamma_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^{3/2}}$$

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

Figure extraite de G. Saporta, Probabilités, analyse des données et statistique, Technip

Représentations graphiques usuelles

5 représentations possibles de la distribution des valeurs de scores UPDRS-OFF de 29 patients parkinsoniens : **laquelle est la plus informative ?**



Plot moyenne + DS

Plot moyenne + erreur standard de la moyenne (SEM)

Boxplot + point moyen

Boxplot + point moyen + valeurs de l'échantillon

Violin plot + Boxplot + point moyen + valeurs de l'échantillon

Source : projet Nucleipark

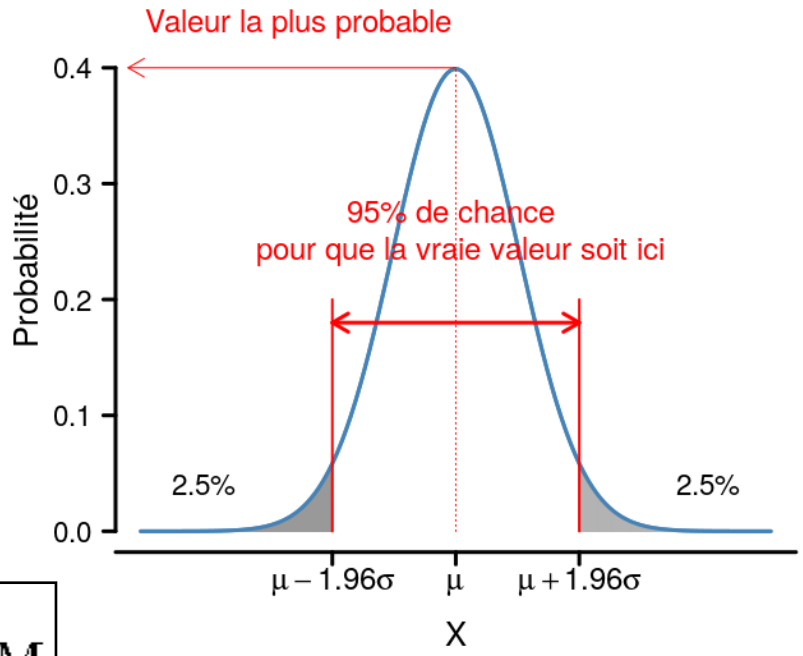
UPDRS = Unified Parkinson's Disease Rating *Scale : échelle d'évaluation de la progression de la maladie de Parkinson ; OFF = hors traitement*

Intervalles de confiance à 95%

Intervalle ayant 95% de chance de contenir le paramètre d'intérêt :

- **Paramètre numérique**
  si X suit une loi normale :

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



- **Moyenne**
  quel que soit la loi de X, si n > 30 :

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

- **Fréquence**
  si n×p, n×(1-p) > 10 :

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

Analyse descriptive bidimensionnelle

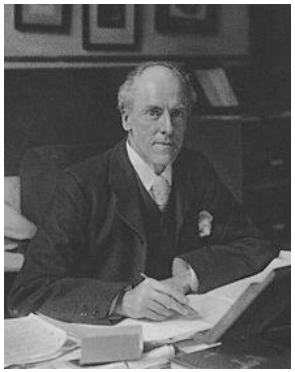
L'analyse bidimensionnelle porte sur l'***étude de la liaison entre 2 variables X et Y observées sur le même échantillon d'individus***.

Méthodes usuelles :

- **2 variables quantitatives** : corrélation + nuage de points
- **2 variables qualitatives** : table de contingence + bubble plot, mosaic plot ou barplot bivarié
- **1 variable quantitative et 1 variable qualitative** : rapport de corrélation + boxplot

⚠ **La liaison permet de mettre en évidence la variation simultanée de 2 variables, mais elle n'entraîne pas nécessairement une relation de causalité !**

Corrélation et nuages de points



Karl Pearson (1857-1936) à l'origine des statistiques appliquées à la biomédecine

Contexte : Le **coefficient r_{XY} de corrélation de (Bravais-) Pearson** mesure la liaison linéaire existant entre **2 variables quantitatives**.

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

☞ ***$|r_{XY}|$ = 1 avec des pentes de régression linéaire + ou moins importantes***

☞ ***r_{XY} = 0 n'implique pas nécessairement l'absence de liaison (non-linéaire) entre X et Y***

2 variables X et Y quantitatives

http://guessthecorrelation.com/

Coefficient ρ de corrélation des rangs de Spearman

Charles Spearman (1863-1945)

Contexte : Le coefficient ρ de corrélation des rangs de Spearman s'applique pour établir :

- la liaison entre **2 variables qualitatives ordinales**
- une relation non-linéaire monotone entre **2 variables quantitatives**

Principe :

1. les valeurs (ou niveaux) ordonnées des variables X et Y sont remplacées par les rangs notés x_i et y_i
2. le coefficient de Spearman est alors donné par la formule suivante :

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

- ✓ si *ρ* est proche de 0 : pas de relation entre X et Y
- ✓ si *ρ* est proche de -1 : relation négative forte entre X et Y
- ✓ si *ρ* est proche de 1 : relation positive forte entre X et Y

Pearson ou Spearman ?

L'exemple ci-dessous illustre les valeurs des coefficients de corrélation de Pearson et Spearman pour 3 types de relations monotones :

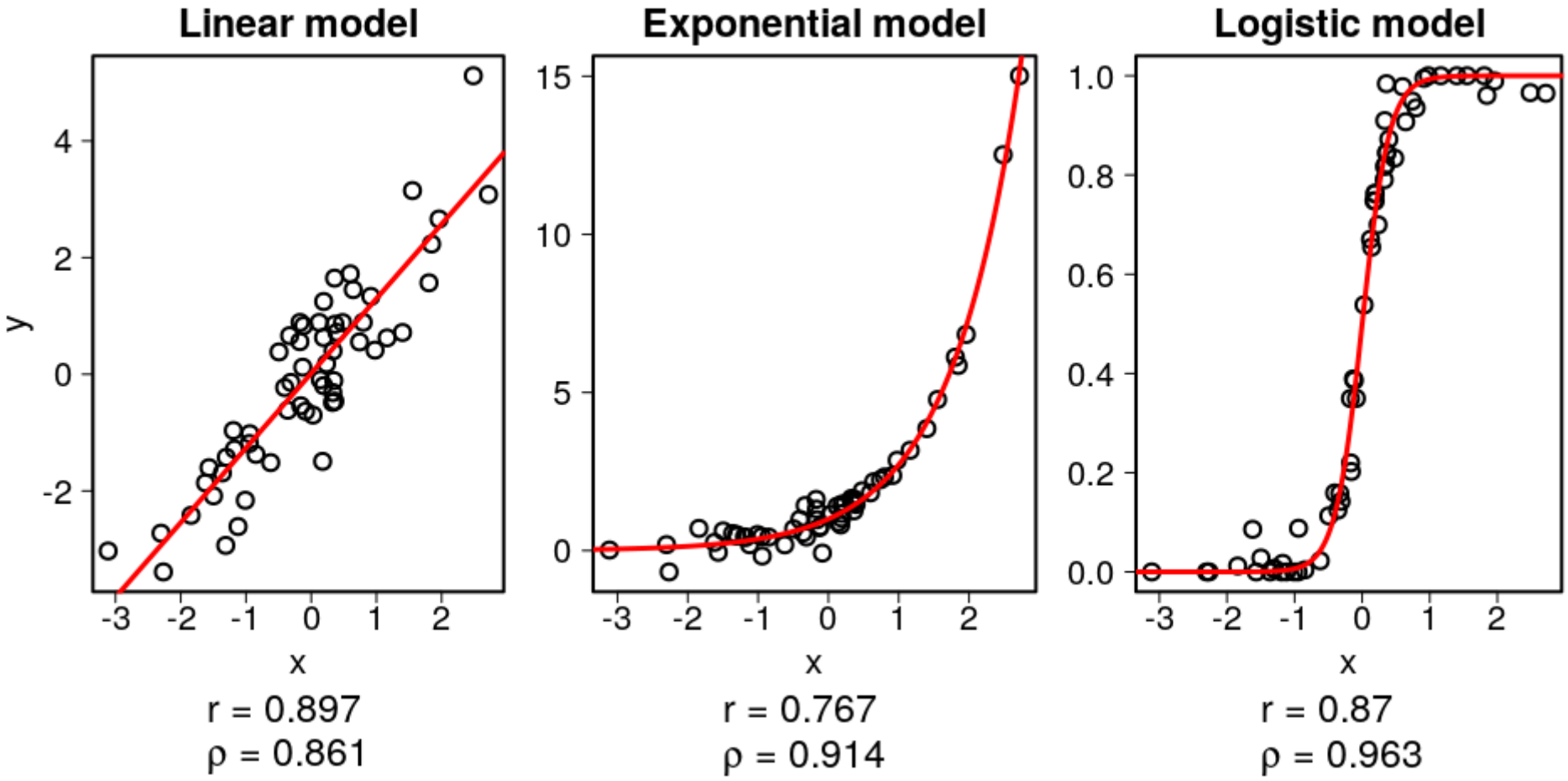


Table de contingence et coefficient χ^2

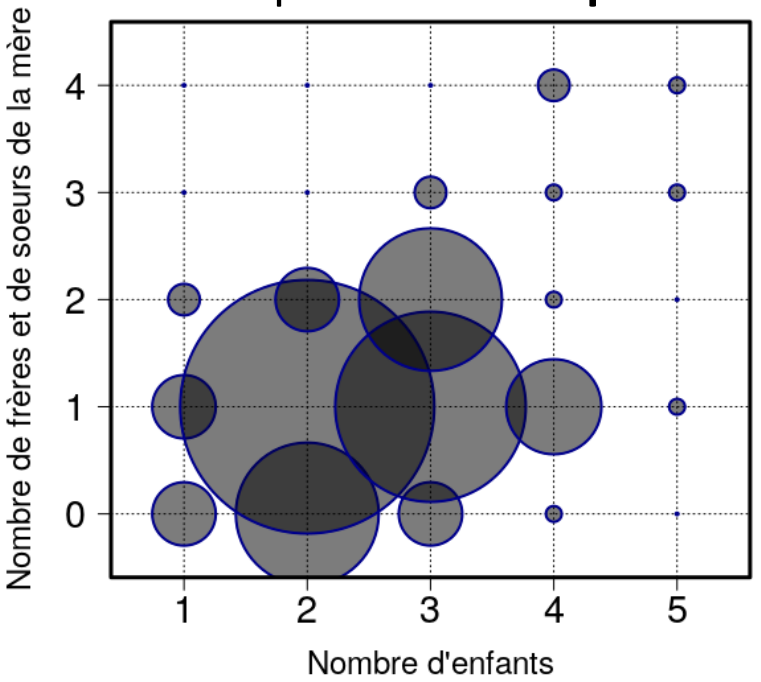
Table de contingence

X \ Y	y_1	$\cdots$	y_h	$\cdots$	y_c	sommes
x_1	n_{11}	$\cdots$	n_{1h}	$\cdots$	n_{1c}	n_{1+}
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_ℓ	$n_{\ell 1}$	$\cdots$	$n_{\ell h}$	$\cdots$	$n_{\ell c}$	$n_{\ell +}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_r	n_{r1}	$\cdots$	n_{rh}	$\cdots$	n_{rc}	n_{r+}
sommes	n_{+1}	$\cdots$	n_{+h}	$\cdots$	n_{+c}	n

Distributions marginales de X et Y

Distributions conditionnelles de X / Y = y_h et Y / X = x_l

Exemple de **Bubble plot**



Le **coefficient χ^2** mesure l'écart entre les effectifs « observés » n_{lh} et les effectifs « théoriques » $n_{l+} * n_{+h}$ attendus en cas d'indépendance de X et Y :

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

« + χ^2 est grand
+ la liaison est forte entre X et Y »

2 variables X et Y qualitatives

Mesures d'association liées au χ^2

Basées sur le coefficient χ^2, **3 mesures de liaison** sont utiles pour **évaluer l'intensité d'association entre 2 variables qualitatives** :

- **V de Cramér**

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

- **Coefficient de contingence CC**

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

- **Coefficient Phi de Pearson**

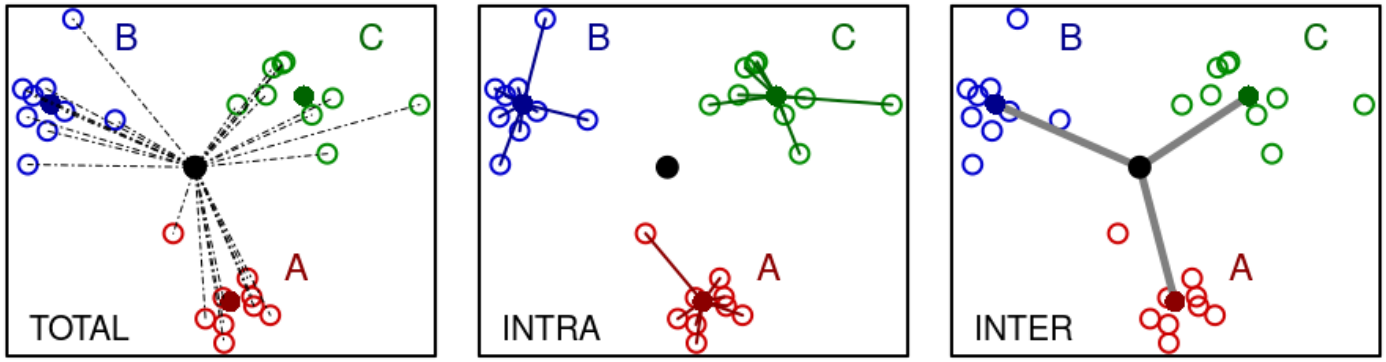
$$\phi = \sqrt{\frac{\chi^2}{n}}$$

2 variables X et Y qualitatives

Rapport de corrélation

Soient Y une variable « dépendante » quantitative observée auprès de n individus et G un facteur indiquant le regroupement des individus en K classes distinctes.

Exemple pour K = 3 classes



Les observations de Y peuvent différer sous l'effet de deux sources de variations :

- Variation à l'intérieur des classes : **Variation intra-classe**
- Variation entre les classes : **Variation inter-classe** (effet du facteur)

On utilise alors le **rapport de corrélation** :

$$\eta^2_{Y/G} = \frac{\text{Variation inter-classe}}{\text{Variation totale}}$$

pour mesurer l'intensité de l'effet du facteur G sur la variable Y (valeur comprise entre **0 = absence de liaison** et **1 = liaison parfaite** entre Y et G).

1 variable Y quantitative et 1 variable G qualitative

Analyse descriptive multidimensionnelle

Dans le cas d'un tableau de données comportant p variables (**p > 3**), il devient impossible de « voir » les individus dans un **espace à p dimensions**.

Pour réduire la dimension, l'objectif des **méthodes factorielles** est de rechercher un nombre restreint de variables composites, appelées **facteurs** ou **composantes principales**, qui sont fabriquées à partir des variables d'origine de telle sorte à résumer le mieux possible les données. Ces méthodes permettent d'obtenir des représentations graphiques des données (individus et variables) avec ces facteurs utilisés comme des axes.

Principales méthodes factorielles :

- Analyse en composantes principales (ACP, variables quantitatives)
- Analyse des correspondances multiples (ACM, variables qualitatives)
- Analyse factorielle multiple (AFM, groupes de variables quantitatives et/ou qualitatives)

Objectifs : résumer et visualiser les données

Comprendre la réduction de dimension

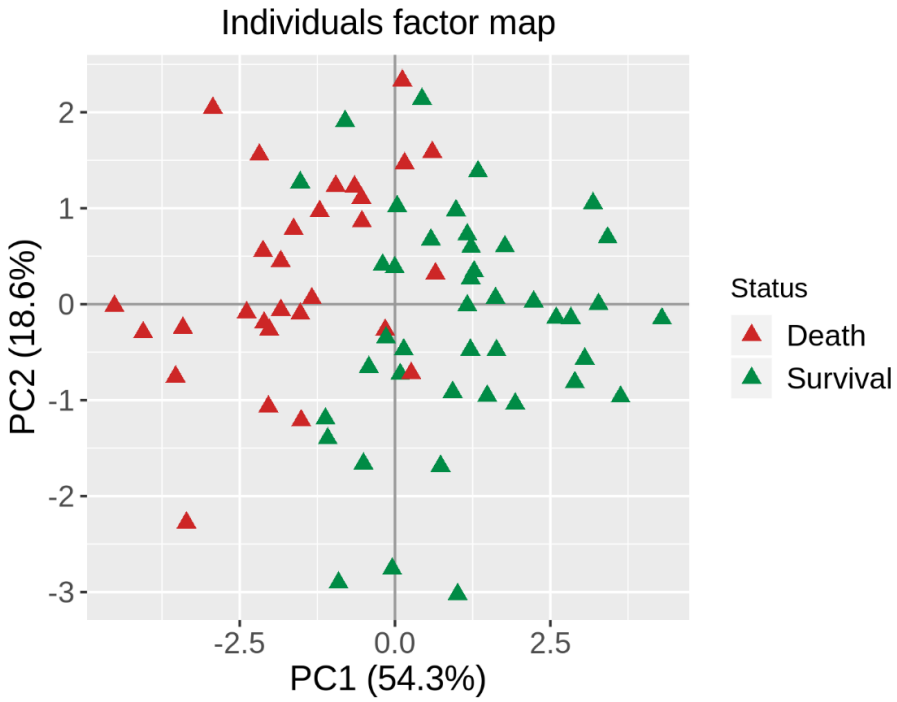
Un exemple commun de **réduction de dimension** est la prise de photographies qui fait passer d'un espace à 3 dimensions (celui où nous vivons) à un espace à 2 dimensions (notre photo).

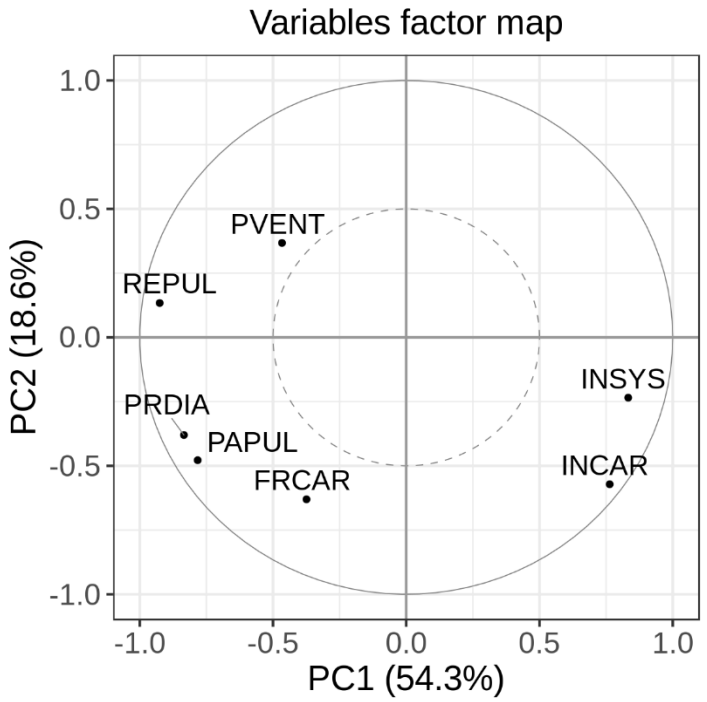
Par ailleurs, selon l'angle sous lequel nous prenons notre sujet, toutes nos photos n'apporteront pas le même niveau d'information.

Figure empruntée à J.-P. Fénelon : Chameau ou dromadaire ?

ACP : Exemple 1

Problème : 71 sujets victimes d'un infarctus du myocarde (29 décès, 42 survivants) pour lesquels on a mesuré 7 variables à leur admission dans un service de cardiologie.





L'ACP permet d'avoir une **projection des individus** dans un plan construit à partir des 7 données cardiovasculaires.

Le **cercle des corrélations** permet de visualiser quelles sont les variables expliquant le mieux la variation des sujets sur les 2 axes.

FRCAR = Fréquence cardiaque, INCAR = Index cardiaque, INSYS = Index systolique, PRDIA = Pression diastolique, PAPUL = Pression artérielle pulmonaire, PVENT = Pression ventriculaire, REPUL = Résistance pulmonaire

Source : J.-P. Nakache

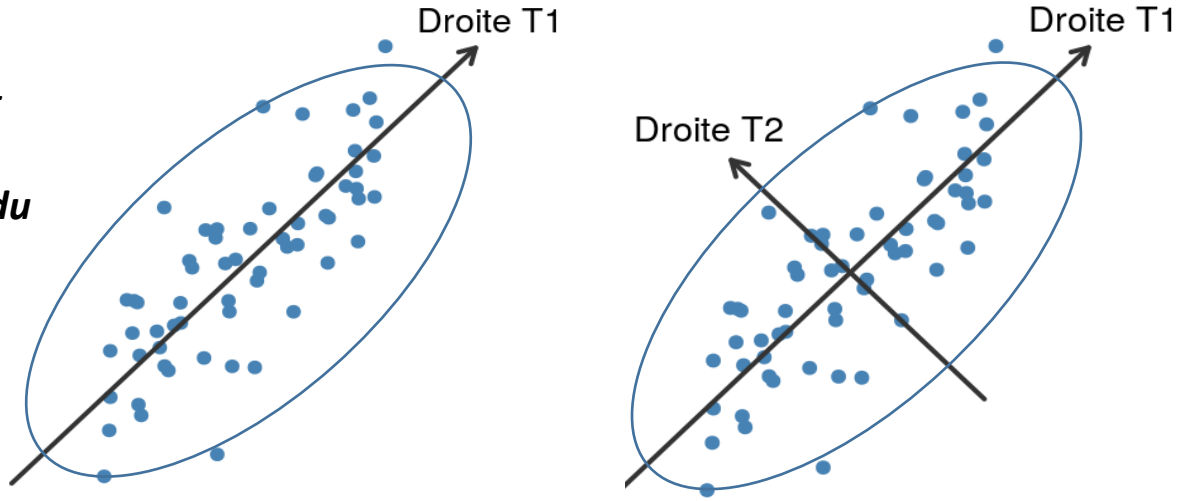
ACP : Principe de la méthode

La projection des données multidimensionnelles sur un plan (2D) nous donne une vision déformée de la réalité. Le but de l'ACP est de **déterminer des espaces de dimension réduite minimisant ces déformations**. On peut alors visualiser les données dans un espace « optimal », appelé **plan principal**, engendré par 2 droites perpendiculaires appelées **composantes principales**.

Calcul des composantes :

- Construction d'une 1ère composante T1 de façon 1/ à **minimiser** les carrés des distances des points à T1 et 2/ à **maximiser** la dispersion du nuage projeté sur T1
- Construction de T2 orthogonale à T1 et maximisant la dispersion
- Et ainsi de suite de telle sorte à capturer autant de variance que souhaitée…

Fig. La droite T1 doit « capturer » le maximum d'inertie du tableau de données



ACP : un coup de projecteur sur les données

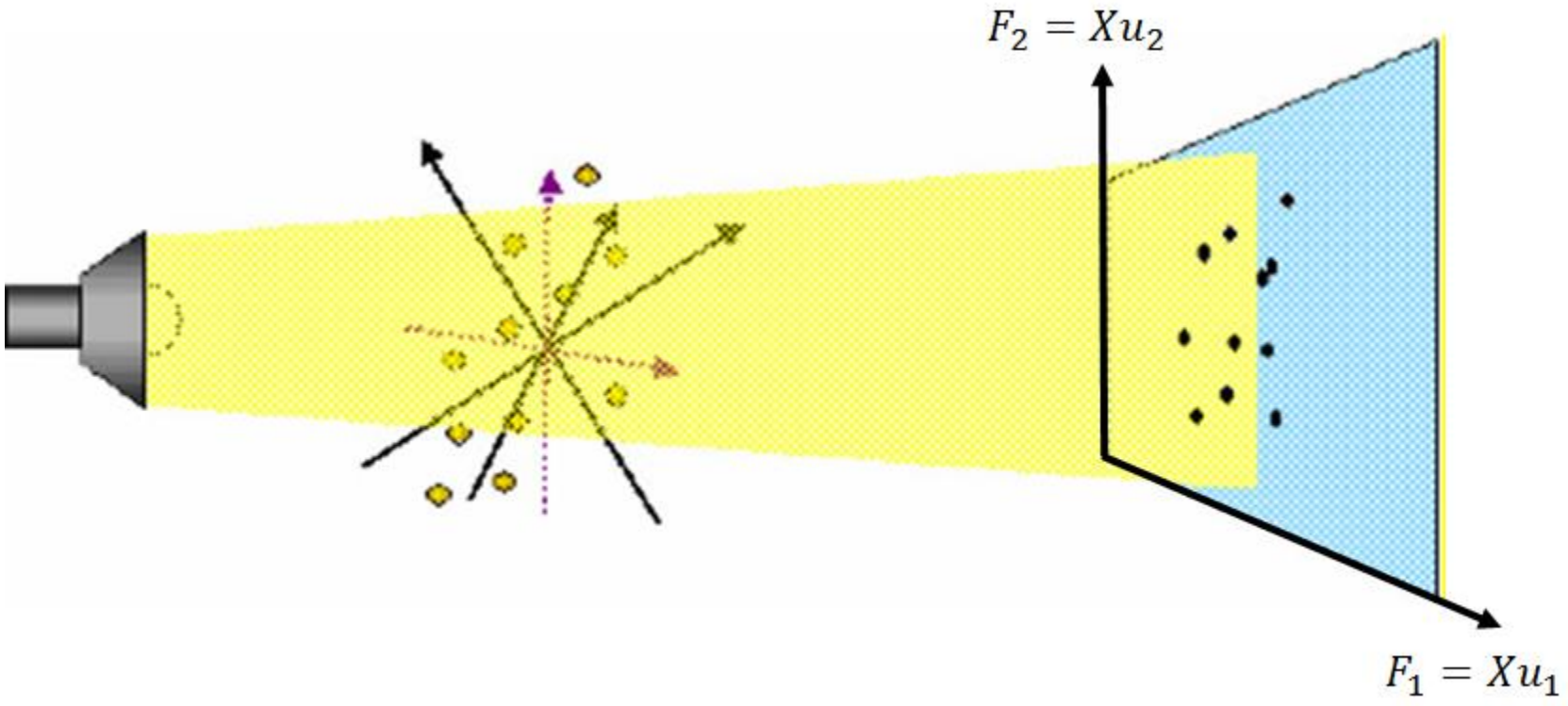


Image extraite de Umetrics AB, Umeå, Suède,
Et reproduite dans M. Tenenhaus,
Statistiques : Méthodes pour décrire, expliquer et prévoir, Dunod, 2007

ACP : la méthode en bref

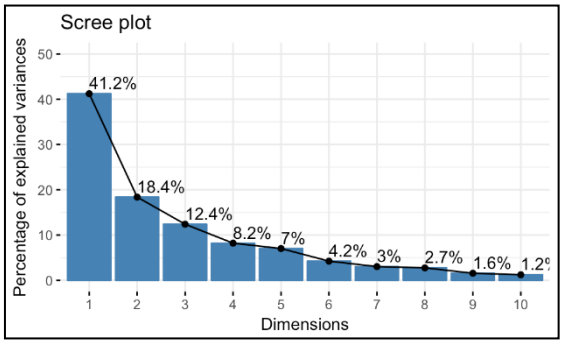
(1) Standardisation des données

(2) Étape de diagonalisation

Tableau n x p des données « centrées-réduites »

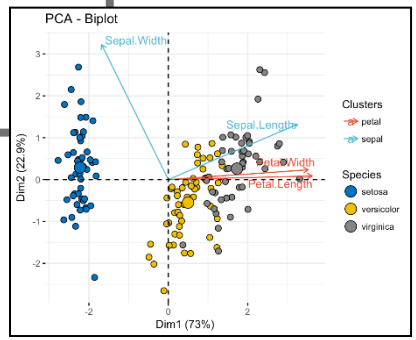
Matrice p x p de corrélation

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



Matrice p x p des vecteurs propres

Matrice n x p' (p' << p) des composantes principales



(3) Projection des données sur les vecteurs propres (directions orthogonales de variance max)

ACP : Exemple 2

Les données représentent les valeurs d'expression RNA-Seq de 8 tissus, avec pour chacun des tissus plusieurs réplicats biologiques.

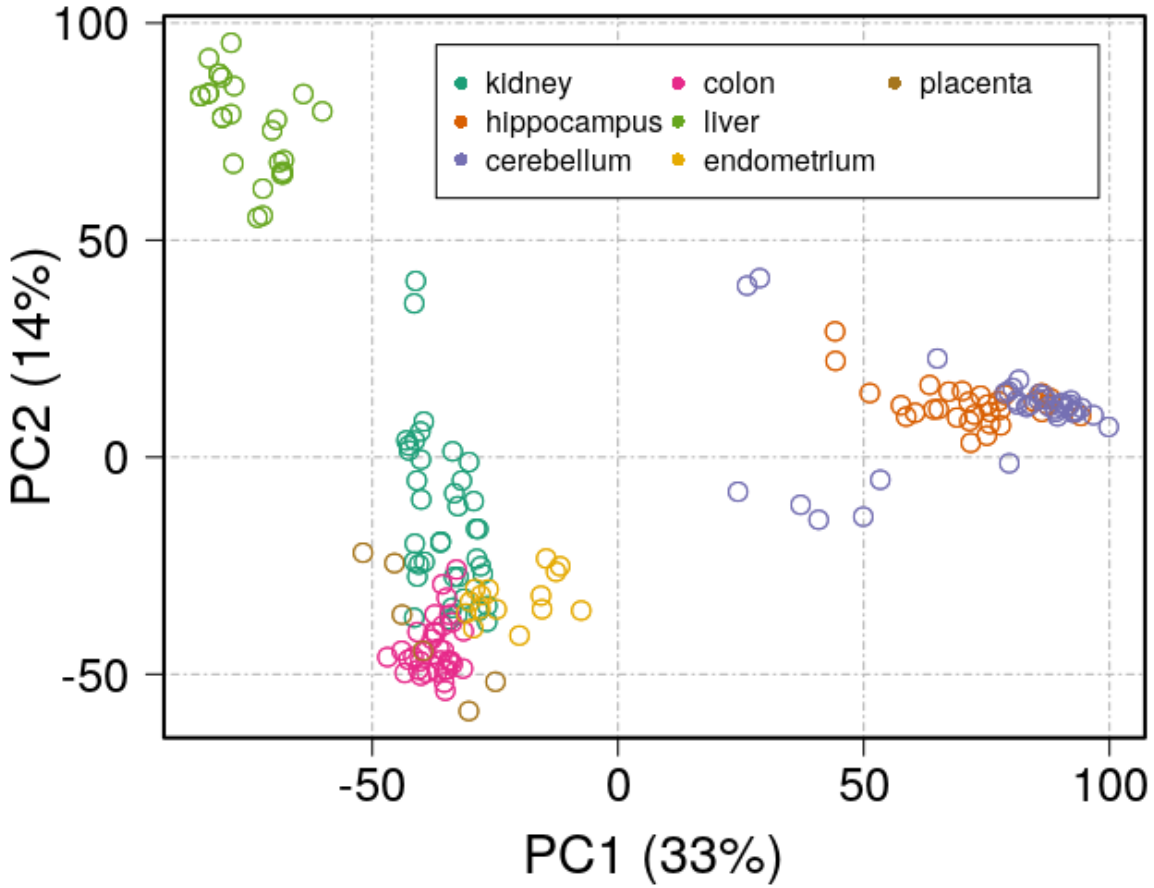


Fig. En coloriant les points par tissu, la projection des échantillons indique l'homogénéité des profils d'expression dans chaque tissu, ainsi que les différences d'expression entre les 8 tissus.

Pour aller plus loin sur l'ACP…

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

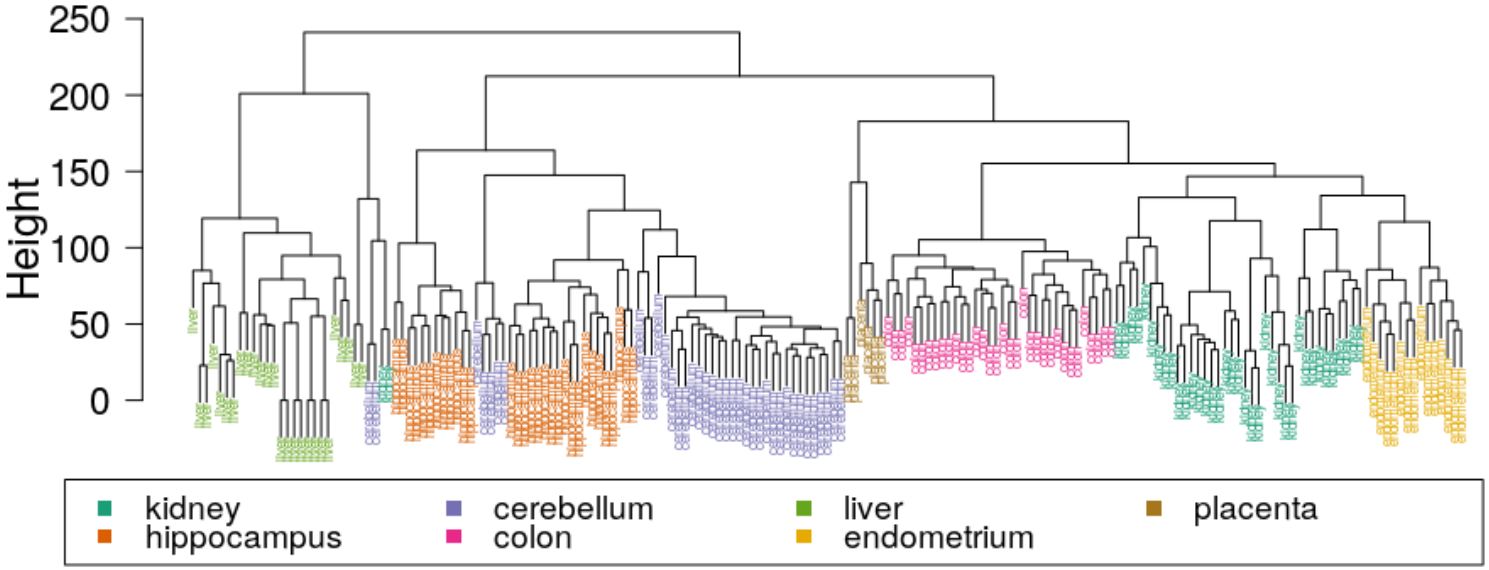
Classification ascendante hiérarchique

La **méthode de CAH** vise à rassembler les individus selon un critère de ressemblance au sein de groupes homogènes et bien séparés entre eux.

L'algorithme repose principalement sur 2 critères :

1. **Choix d'une distance** : euclidienne, max, manhattan…
2. **Stratégie d'agrégation** : diamètre (complete), moyenne (average), ward…

Représentation des échantillons par un dendrogramme



Données d'expression des 8 tissus : distance euclidienne et méthode du diamètre

Données manquantes

Les **données manquantes** (DM) sont fréquentes voire inévitables dans les bases de données, celles-ci peuvent provenir de causes diverses :

- Oubli de mesure ;
- Donnée mesurée mais a été perdue ou pas notée ;
- Donnée mesurée mais la valeur est considérée inutilisable (erreur manifeste de mesure, la valeur semble aberrante) ;
- Donnée non disponible : réponse du type « Ne sait pas » ;
- Cas de censure : la valeur est en dehors des limites de détection de l'appareil ;
- Censure dans une étude de survie :
  - C. à gauche : le sujet a déjà subi l'événement avant le début de l'étude,
  - C. à droite : l'événement n'a pas été observé à la fin de l'étude ;
- Génétique : absence ponctuelle de génotype (SNPs de certains individus)…

Méthodes de gestion des DM

Dans certains cas, l'analyse est possible sans imputer les données manquantes. En particulier, lorsque le retrait des individus à DM n'occasionne pas une perte trop importante de l'information disponible.

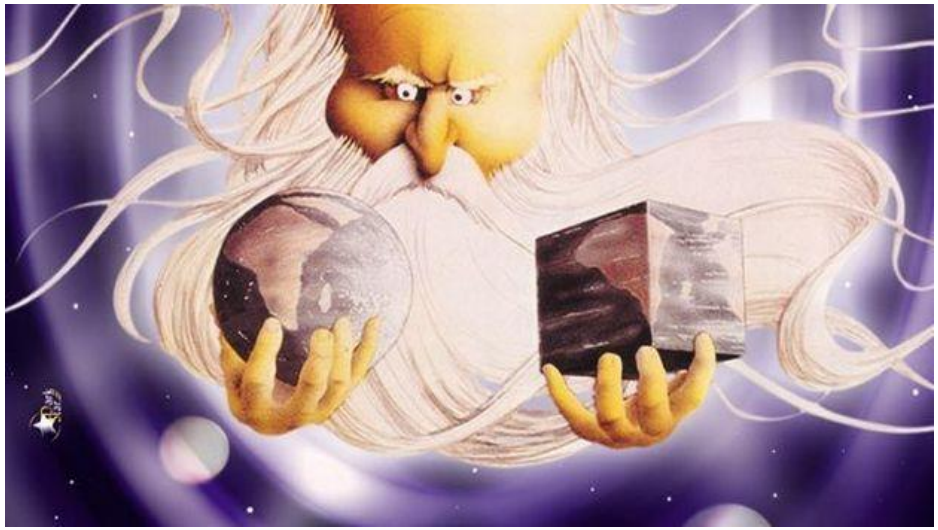
Sinon il existe différentes stratégies d'**imputation des DM** :

- **Imputation simple** : La DM est remplacée par une valeur unique : moyenne des k observations les plus proches (k-NN), régression locale, algorithme NIPALS, SVD, utilisation des forêts aléatoires…
- **Imputation multiple** : Une DM est remplacée par plusieurs valeurs candidates permettant de prendre en compte dans l'analyse l'incertitude supplémentaire liée au remplacement de la DM
- **Approche bayésienne** : On suppose que les DM sont issues d'une distribution *a priori*

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

- **Génétique** : reconstruction des SNPs manquants par haplotype à partir d'une population de référence (=> valeurs « probables » des génotypes)

Partie 2 : TESTER



Choisir entre 2 hypothèses

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

Principe des tests statistiques

Un **test statistique** est une procédure de décision entre 2 hypothèses concernant un ou plusieurs échantillons.

Ex. Dans une étude comportant un groupe « médicament » et un groupe « placebo », on mesure les tensions artérielles dans les 2 groupes pour déterminer si le médicament a un effet sur la tension.

Formalisation : Si µ1 et µ2 sont les moyennes de tension des 2 groupes « médicament » et « placebo », une manière de démontrer que le médicament modifie la tension est de montrer que µ2 est différent de µ1 à partir des observations effectuées.

But du test : Déterminer si la ***différence observée des 2 moyennes*** est simplement ***due au hasard***, c'est-à-dire aux fluctuations d'échantillonnage, ou si au contraire la différence observée est ***bien réelle***.

Hypothèses statistiques

Les **hypothèses statistiques** traduisent la question biologique (par *ex. l'effet d'un médicament*) sous la forme de 2 énoncés complémentaires utilisant les caractéristiques de distribution de la population étudiée (valeurs de paramètres, forme de la distribution…), appelés :

- **Hypothèse nulle notée H0** : celle que l'on considère vraie *a priori* ;
- **Hypothèse alternative notée H1** : hypothèse complémentaire de H0.

Le but du test est alors de décider si le modèle décrit par H0 est « plausible ».

Ex.

H0 : le médicament n'a pas d'influence contre H1 : il en a une

ou

H0 : $\mu 1 = \mu 2$ vs H1 : $\mu 1 \neq \mu 2$

H1 est dite **bilatérale** lorsque que l'on ne cherche pas à connaitre le sens de la différence (*i.e.* $\mu 1 \neq \mu 2$), ou **unilatérale** si l'on s'intéresse à un sens particulier (*i.e.* $\mu 1 < \mu 2$ ou $\mu 1 > \mu 2$).

Exemples de formulation pour H0 et H1

1. **Comparaison de 2 traitements nouveaux A et B**

   H0 *: les 2 traitements sont équivalents*
   H1 bilatérale *: les 2 traitements ont une efficacité différente*

2. **Comparaison d'un traitement A à un placebo (produit inactif)**

   H0 *: le traitement A et le placebo sont équivalents*
   H1 unilatérale *: le traitement A a une efficacité supérieure au placebo*

3. **Effet d'un médicament anti-tumoral lié à la présence d'un variant V**

   H0 *: indépendance entre l'effet obs. (+ ou -) et la prés. du variant (V-/V+)*
   H1 *: existence d'une liaison entre les 2 facteurs*

4. **Comparaison de 4 traitements A, B, C et D**

   H0 *: les 4 traitements sont équivalents*
   H1 bilatérale *: au moins un traitement est différent des autres*

Dans un test statistique, « choisir entre H0 et H1 » se fait de sorte à **éviter 2 types d'erreurs**, appelées **erreurs de 1ère et 2ème espèce**.

Statistique de test

Puisqu'il faut choisir entre « entre H0 et H1 », l'issue d'un test va reposer sur une **variable de décision** appelée **statistique de test S**. S est une variable aléatoire qui a pour but de résumer l'information contenue dans l'échantillon et dont la valeur observée $\mathbf{s_{obs}}$ peut être calculée à partir des observations.

Ex. de statistiques S usuelles :

Tests paramétriques : **T** de Student, **F** de Fisher dans l'ANOVA…

Tests non paramétriques : **U** de Mann-Whitney, **K** de Kruskal-Wallis…

 Pour décider…

☺ La loi de distribution de S est connue sous H0, ce qui permet de contrôler l'erreur de 1ère espèce de « rejeter H0 à tort » [FAUX POSITIF] :

Ex. ***Conclure à l'efficacité d'un traitement qui est en fait inefficace…***

☹ Par contre, la loi de S est inconnue sous H1, ce qui rend difficile de contrôler l'erreur de 2ème espèce d'« accepter H0 à tort » [FAUX NÉGATIF] :

Ex. ***Risque de ne pas mettre en évidence l'efficacité d'un traitement…***

Risque α de 1ère espèce

Si H0 est vraie…

l'erreur α de type I est la Probabilité de « rejeter H0 à tort » :
erreur α = P_{H0}(rejeter H0), où α et P_{H0} sont connues

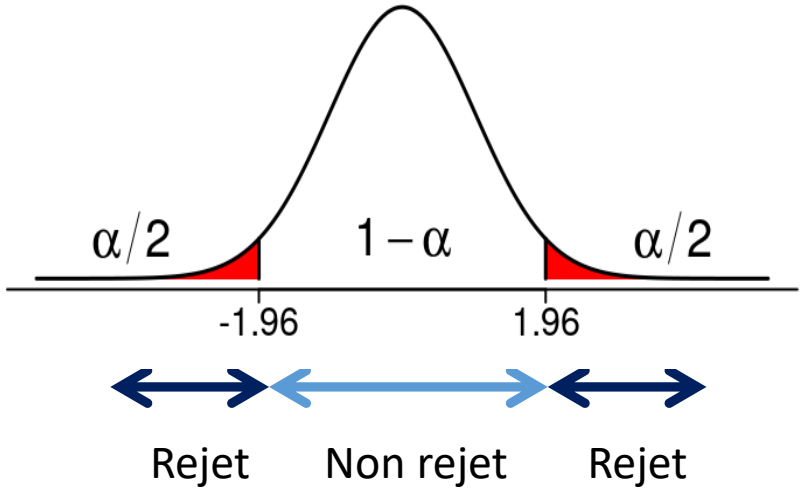
Le **seuil α** est fixé *a priori* par l'expérimentateur (le plus souvent à 5% ou 1%) pour définir le risque maximum acceptable pour l'**erreur α**.

Soit $\mathbf{s_{obs}}$ la valeur observée de la statistique **S** : au regard de P_{H0} (distribution théorique de S sous H0), s_{obs} peut être « suffisamment probable » ou « peu probable » :

Ex. si **S** ~ N(0,1) :

-1.96 et 1.96 sont les quantiles d'ordre 0,025 et 0,975 de la loi Normale délimitant la **région de rejet**

Règle de décision selon les valeurs de $\mathbf{s_{obs}}$ **(H1 bilatérale,** α = 5%) :



Risque β de 2ème espèce

Si H1 est vraie…

l'erreur β de type II est la Probabilité d'« accepter H0 à tort » :

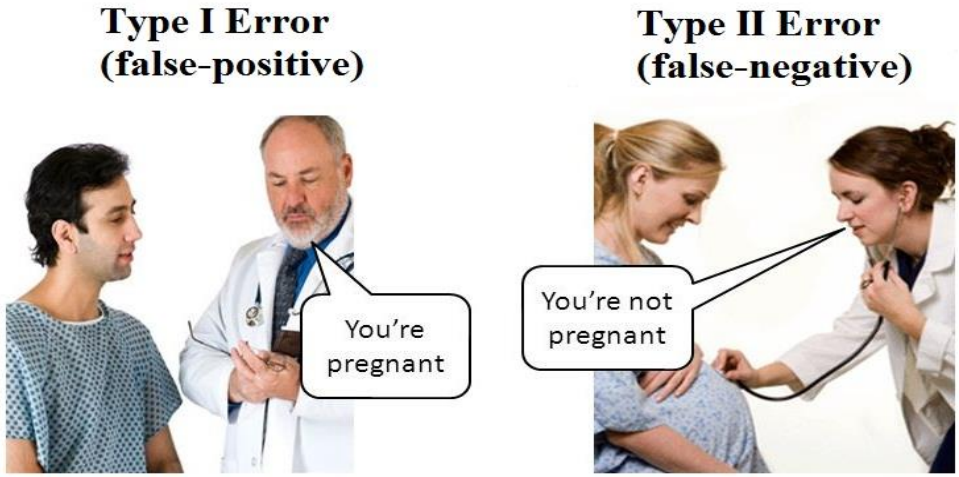
$\beta = P_{H1}$(rejeter H1), avec P_{H1} inconnue et l'erreur β indéterminée en général

La quantité 1−β est la **puissance** du test.

	Réalité	
Décision	H0 est vraie	H1 est vraie
H0 est acceptée	Bonne décision **VN** Niveau de confiance 1 – α	Mauvaise décision **FN** Erreur β
H0 est rejetée	Mauvaise décision **FP** Erreur α	Bonne décision **VP** Puissance du test 1 - β

Erreurs α et β

Idéalement, un « bon test » se doit de minimiser les 2 types d'erreur :



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

Cette minimisation est en fait un compromis à faire car les 2 types d'erreurs sont étroitement liés :

- ***moins le seuil α est sévère (par ex. 10% plutôt que 5%), détection plus facile d'un effet mais risque accru de commettre l'erreur α***
- ***plus le seuil α est sévère (par ex. 1% plutôt que 5%), risque de manquer un effet soit un risque accru de commettre l'erreur β***

Si le contrôle de l'erreur α ne pose pas de problème, nous verrons dans la suite comment il est possible de contrôler l'erreur β (dépendante des tailles d'effet et d'échantillons).

p-valeur (1/2)

Pour **reporter le résultat d'un test**, il est d'usage d'indiquer sa **p-valeur** qui

Si H0 est vraie…

représente la probabilité ***p*** d'obtenir la valeur observée de la statistique de test (ou une valeur encore plus extrême) :

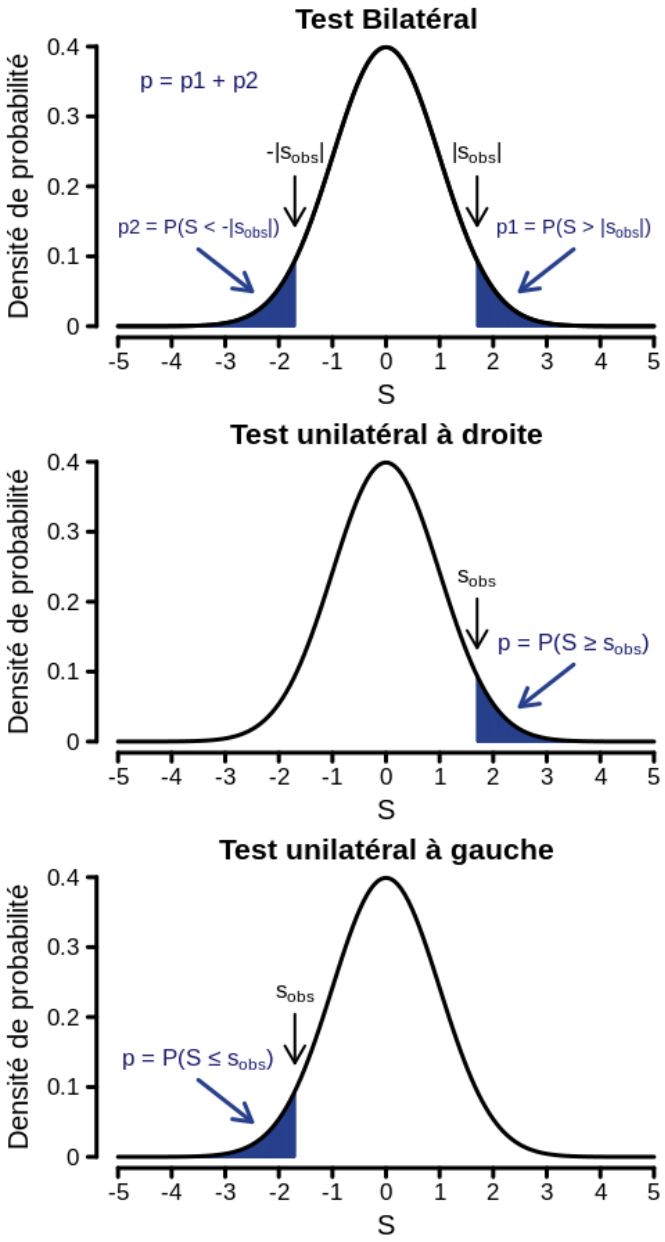
- Si le test est bilatéral : $p = P_{H0}(|\mathbf{S}| > \mathbf{s_{obs}})$
- Si le test est unilatéral à droite : $p = P_{H0}(\mathbf{S} \geq \mathbf{s_{obs}})$
- Si le test est unilatéral à gauche : $p = P_{H0}(\mathbf{S} \leq \mathbf{s_{obs}})$

☞ *Plus cette valeur est petite, plus H0 a de chances d'être rejetée, la valeur obtenue de s_{obs} étant « trop peu probable » au sens de la distribution sous H0.*

On utilise généralement les seuils :

- $p < 0{,}001$: *très forte significativité* *** ☺
- $p < 0{,}01$: *forte significativité* **
- $p < 0{,}05$: *significativité* *
- $p > 0{,}05$: n.s. ☹

Distribution théorique de S sous H0



p-valeur (2/2)

L'interprétation du résultat d'un test statistique basée uniquement sur la p-valeur peut s'avérer délicate en pratique, en ce sens que **la notion de « significativité statistique » n'a pas systématiquement valeur de « significativité biologique »**, faisant référence à l'intensité ou l'importance réelle de l'effet biologique observé.

Situations pratiques fréquentes :

1. Différence entre 2 groupes de sujets, considérée « importante » par l'expérimentateur, ne passant pas le seuil de la significativité statistique du simple fait des variations d'échantillonnage mesurées sur de trop petits effectifs engendrant un « manque de puissance » du test.
2. Effets minimes, de faible ampleur ou de faible importance d'un point de vue biologique, considérés comme statistiquement significatifs du fait de trop gros effectifs pour lesquels « tout devient significatif ! ».

Dans ces 2 cas, l'indication complémentaire d'une mesure quantitative décrivant l'ampleur réel de l'effet observé (**taille d'effet**), de préférence indépendante des tailles d'effectifs (ex. différence de moyennes), peut s'avérer utile pour accorder les 2 notions de significativités statistique et scientifique.

Taille d'effet ou *lorsque la p-valeur ne suffit pas* !

En complément de la p-valeur, on peut s'intéresser à l'éloignement de l'échantillon à la **norme indiquée par H0**. Cet éloignement est appelé un **effet** et l'importance de cet effet peut être évaluée grâce à une statistique appelée **taille d'effet**.

Taille d'effet absolu :

- $M_{obs} - M_{H_0}$ (si par ex. M est une différence de moyennes)
- $p_{obs} - p_{H_0}$ (si par ex. p est une proportion)

Taille d'effet relative sans unités :

$$d = \frac{M_{obs} - M_{H_0}}{DS_{échantillon}} \quad \text{ou} \quad \text{rapport des proportions :} \quad \frac{p_{obs}}{p_{H_0}}$$

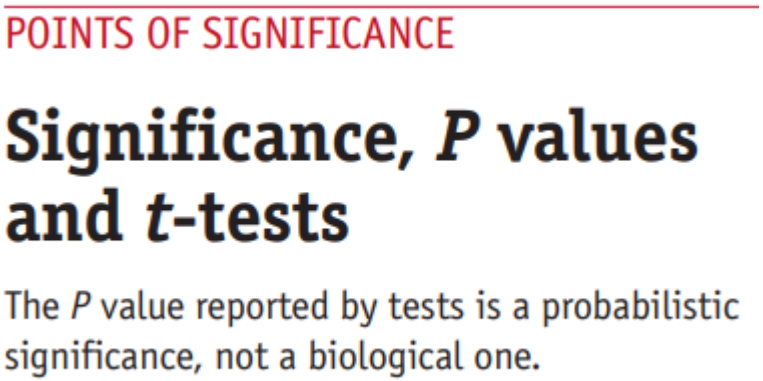
Pour d, on peut décrire la taille d'effet en se référant à l'**échelle de Cohen** : effet « faible » autour de 0.2, « moyen » autour de 0.5 et « fort » autour de 0.8.

La notion de taille d'effet peut également servir à réaliser des **méta-analyses** combinant dans une étude globale les tailles d'effet issues de différentes études.

Autres types de taille d'effet : Corrélation, rapports de cote (odds ratio)…

(Sullivan & Feinn, J Grad Med Educ 2012)

Pour aller plus loin avec les *p*-valeurs, la significativité et les tailles d'effet…

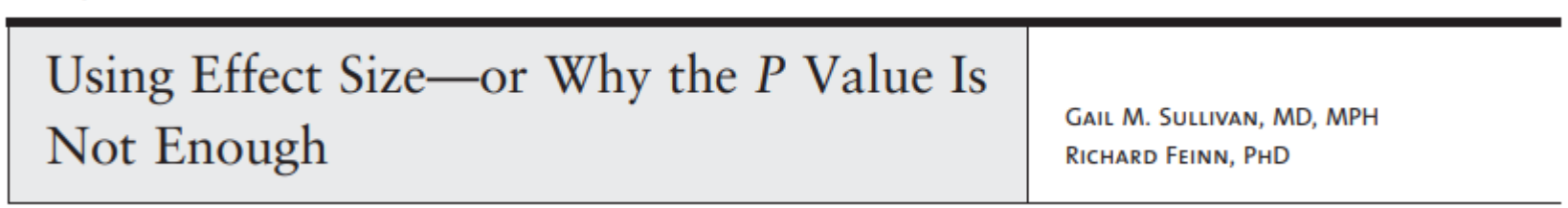


POINTS OF SIGNIFICANCE

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

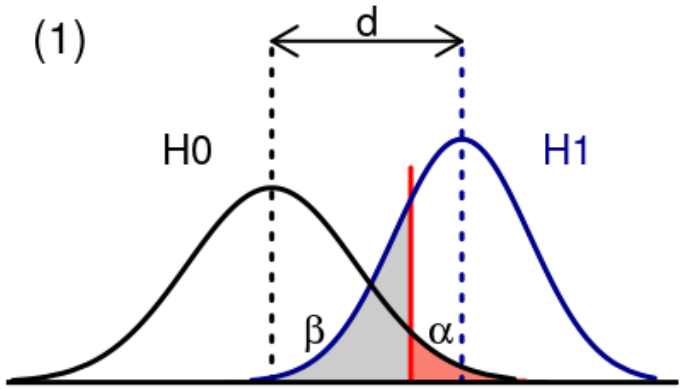
Using Effect Size—or Why the *P* Value Is Not Enough

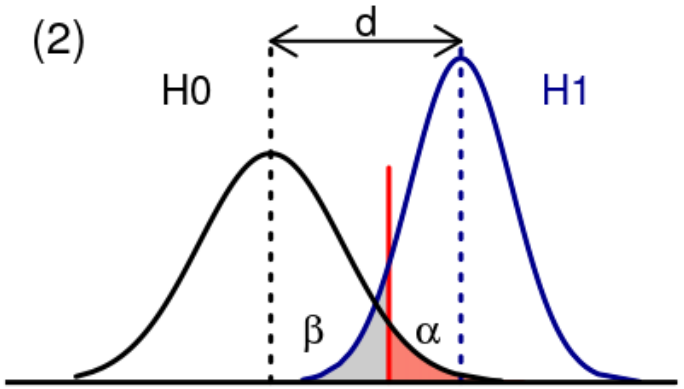
GAIL M. SULLIVAN, MD, MPH
RICHARD FEINN, PHD

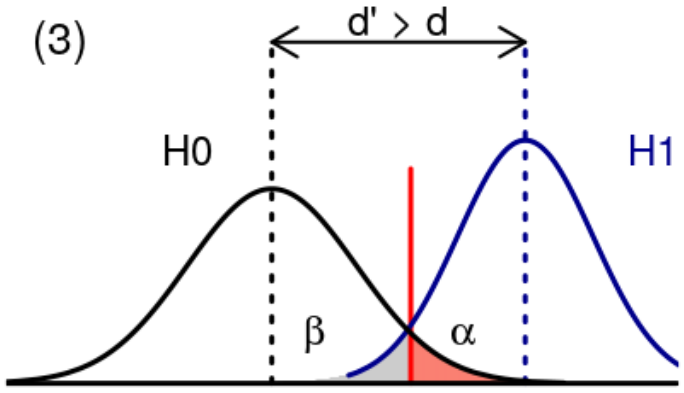
Sullivan GM, Feinn RS. Using effect size—or why the P. value is not enough. J Grad Med Educ. 2012;4(3):279–282. doi:10.4300/JGME-D-12-00156.1.

Puissance d'un test

La **puissance d'un test** est sa capacité à détecter les écarts par rapport à l'hypothèse nulle. Tandis que le α est fixé, la puissance va dépendre de la taille d'échantillon, de la dispersion et de la taille de la différence (ou taille d'effet d) « que l'on suppose exister ».







(1) vs (2) même taille d'effet d : la puissance augmente avec le nombre de sujets

(1) vs (3) mêmes effectifs : la puissance augmente avec la taille d'effet

Nombre de sujets

Pour mener un test à bien, la **taille d'échantillon** joue un rôle important sur la puissance de l'expérience :

☹ Une **taille trop petite** aura tendance à donner des écart-types plus grands pouvant conduire à manquer des effets lorsque ceux-ci existent réellement (manque de puissance).

☹ *A contrario*, une **taille trop grande** aura tendance à détecter des effets infimes « statistiquement significatifs » (rejet systématique de H0), y compris des effets n'ayant pas de réalité biologique.

Il convient donc d'avoir le nombre de sujets le mieux adapté à l'expérience, ***« quelque part entre trop et pas assez… »***. Pour cela, des règles de calcul existent pour déterminer la taille d'échantillon selon :

- la taille d'effet,
- la dispersion,
- le niveau de puissance souhaité

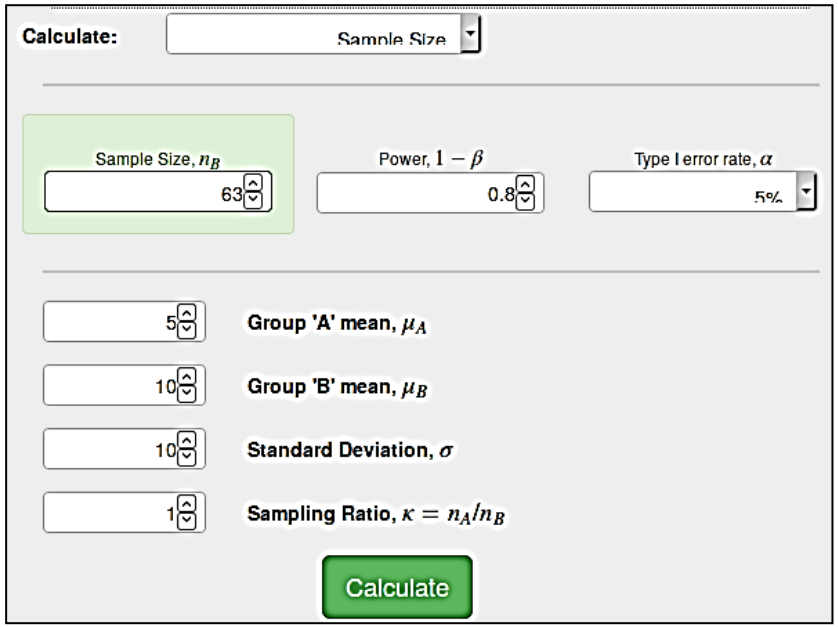
Puissance et taille d'échantillon

La puissance est importante car elle indique la probabilité que le test identifie une différence ou un effet significatif lorsque celui-ci existe réellement.

Pour mener un test avec un niveau de puissance élevé (souvent 1 – β = 0.8), une connaissance ou une idée *a priori* de la taille d'effet et de la dispersion est donc nécessaire afin d'**évaluer la taille d'échantillon requise**. Dans ce but, une **étude pilote réalisée sur un échantillon réduit** peut indiquer ces valeurs, en vue d'une étude à plus large échelle.

Des formules explicites existent pour la plupart des tests paramétriques courants permettant ces calculs de puissance ou de tailles d'échantillon.

Ces calculs peuvent se faire simplement à l'aide de **calculateurs en ligne** où il suffit de sélectionner le test choisi et les valeurs anticipées des paramètres.



http://powerandsamplesize.com/

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

Conduire un test statistique

Principales étapes d'un test :

1. Choix du **type de test** selon la question posée et le type des données
2. Établir **H0 et H1** (les 2 hypothèses doivent être mutuellement exclusives et inclure toutes les possibilités de l'expérience)
3. Choix du **risque α**
4. Calcul de la **statistique de test** (variable de décision dont on connait la distribution théorique sous H0) ***-> LOGICIEL STAT***
5. Calcul de la **p-valeur** ***-> LOGICIEL STAT***
6. Conclusion

Les choix (étapes 1, 2 et 3) ainsi que l'interprétation finale (étape 6) restent à l'appréciation de l'expérimentateur. D'autre part, il faut toujours garder en tête que le résultat d'un test comprend toujours une dose d'incertitude.

☞ **ON NE SAURA JAMAIS SI ON A BIEN PRIS LA BONNE DÉCISION !**

Problème des comparaisons multiples (1/2)

Le problème des **comparaisons multiples** survient lorsqu'une analyse statistique implique de tester plusieurs hypothèses à la fois.

La multiplication des tests sur un même jeu de données entraîne alors une augmentation du risque de se tromper en mettant en évidence des différences significatives qui ne sont dues qu'au hasard (cas de faux positifs).

D'un point de vue statistique, on dit que le **risque alpha global** augmente. En général, le risque alpha, pour un seul test, est fixé à 5%. Lorsque k tests sont réalisés, le risque alpha global devient : $\alpha_{global} = 1 - (1 - \alpha)^k$

Pour $\alpha = 5\%$, on a ainsi : $\alpha_{global} = 0{,}487$ (k=13) et $\alpha_{global} = 0{,}512$ (k=14)

Ce qui implique qu'au delà de 13 comparaisons, on a plus d'1 chance sur 2 d'avoir une comparaison ou plus, significative purement par hasard !!! ☹

☞ **Des corrections existent pour arriver à un taux d'erreur global de 5% pour l'ensemble des tests effectués.**

Problème des comparaisons multiples (2/2)

Corrections de tests multiples

Bonferroni : *les p-valeurs n'augmentent qu'en fonction de leur nombre (nombre de tests effectués)*

Carlo Bonferroni (1892-1960)

- Correction de type FWER (Familywise Error Rate)
- Procédure **très conservatrice** (= sévère)
- Calcul :

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

nbp : nombre de tests

Benjamini-Hochberg : *les p-valeurs augmentent en fonction de leur nombre et du taux de p-valeurs non-significatives*

Yoav Benjamini (1949-)

- Correction de type FDR (False Discovery Rate)
- **Peu conservatrice** (mieux adaptée pour sélectionner des caractères « potentiellement intéressants »)
- Souvent utilisée dans les analyses d'expression différentielle
- Calcul :

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

i : rang de *p* dans les p-valeurs ordonnées en ordre croissant

1/ Tests paramétriques

Pour des observations supposées suivre une distribution normale ou un échantillon de suffisamment grande taille pour accepter la normalité asymptotique par le **théorème de la limite centrale** (en pratique quand n > 30) :

1 échantillon

- Comparaison de la moy. d'éch. à une valeur théorique (variance supposée connue, **Gauss**)
- Comparaison de la moy. d'éch. à une valeur théorique (variance inconnue et estimée, **Student**)
- Comparer une proportion à une valeur théorique (**test binomial à un échantillon**)

2 échantillons indépendants

- Comparaison de 2 moyennes (variances égales ou échantillons suffisamment grands, **Student**)
- Comparaison de 2 variances (**Fisher**)
- Comparaison de 2 proportions (**test du chi-deux pour l'égalité des proportions avec correction de continuité de Yates**)

2 échantillons appariés : même échantillon observé à 2 instants différents ou dans 2 conditions différentes (**Student apparié**)

Plusieurs échantillons : **ANOVA à un facteur**, tests de **Bartlett** et de **Levene** pour comparer plusieurs variances

Extrait de https://www.math.univ-toulouse.fr/~besse/Wikistat/pdf/st-l-inf-tests.pdf

Test T de Student (n < 30)

William Sealy Gosset « Student » (1876-1937)

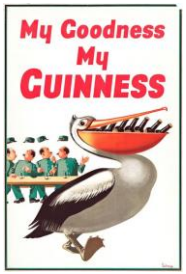
But :

- Comparer la moyenne d'un échantillon à une moyenne théorique
- Comparer les moyennes de 2 échantillons
- Comparer les moyennes de 2 séries « appariées »
- Tester un coefficient de corrélation

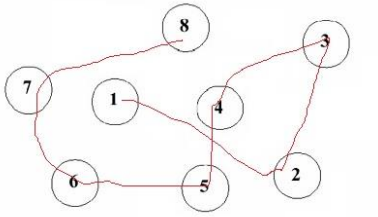
⚠ Test T de Welch pour 2 échantillons de variances inégales

Principe du test (variances égales) :

1. Estimer la moyenne et l'écart-type de chaque échantillon
2. Calculer la valeur de la statistique $$t_0 = \frac{\bar{x}_1 - \bar{x}_2}{s_p \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$
3. Déterminer le nombre de ddl = $n_1 + n_2 - 2$ pour extraire la valeur critique de la distribution de Student correspondant au niveau de risque α
4. Comparer t_0 à la valeur critique et conclure



L'entreprise Guiness ne l'ayant pas laissé signer de son vrai nom, William Gosset, maître brasseur et statisticien, publiera ses travaux en statistiques sous le nom de Student.



Exemple : Test de Student

Le Trail Making Test (TMT) est un test papier-crayon en deux parties largement utilisé pour évaluer la vitesse motrice et l'attention visuelle. Dans l'exemple suivant, les performances dans la partie A du test (temps d'exécution en secondes) sont comparées entre deux groupes de 16 patients Alzheimer et 14 témoins appariés selon l'âge.

Two Sample t-test

***alternative hypothesis*: "true difference in means between group Control and group Alzheimer is not equal to 0"**

```
t.test(TMT.A~Group, var.equal=TRUE)
```

t = -3.5105, df = 28, p-value = 0.001534 (**)

mean in group Control	mean in group Alzheimer
32.44673	52.38770

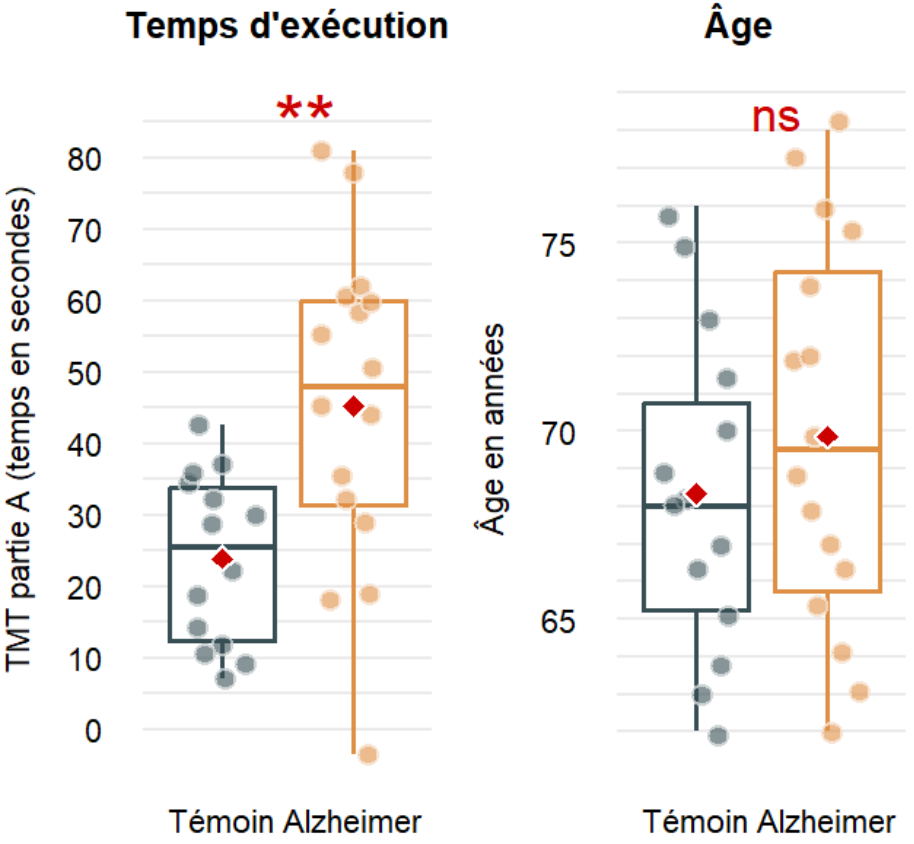
95% confidence interval: -31.576698 -8.305244

```
t.test(Age~Group, var.equal=TRUE)
```

t = -0.86553, df = 28, p-value = 0.3941 (ns)

mean in group Control	mean in group Alzheimer
68.35714	69.87500

95% confidence interval: -5.110113 2.074399



ANOVA à 1 facteur (1/2)

Ronald Fisher (1890-1962)

Pour une variable X mesurée sur n individus regroupés en *p* groupes, l'ANOVA consiste à construire le test d'hypothèse suivant :

$$\begin{cases} H_0 : \mu_1 = \mu_2 = \ldots = \mu_p = \mu \\ H_1 : \exists \mu_j \neq \mu \end{cases}$$

« Tous les groupes égaux » vs « Au moins 1 groupe différent des autres »

où $\mu_1, \mu_2, \ldots, \mu_p$ désignent les moyennes des p groupes et μ la moyenne globale.

Conditions d'application du test :

Les observations doivent vérifier les 3 hypothèses suivantes :

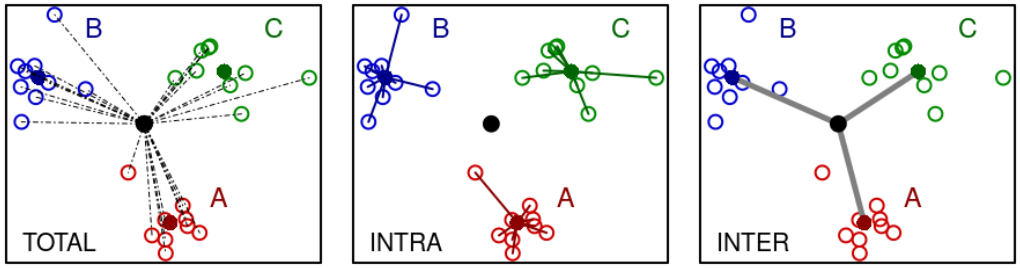
- **Indépendance** des données
- **Normalité** des distributions dans les groupes
- **Homoscédasticité** : variance identique dans les différents groupes (test de Bartlett)

ANOVA à 1 facteur (2/2)

Suivant le **principe de décomposition de la variance** : Variance totale = Variance intra-classe + Variance inter-classe, la **statistique *F*** de test évalue le rapport entre la variance « expliquée » (INTER) et la variance résiduelle (INTRA) :

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

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



Sous H_0 :

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

Reste à comparer *F* à la valeur critique de la loi de Fisher à p-1 et n-p ddl correspondant au niveau de risque α pour conclure.

Remarque : Le test ANOVA permet de détecter une différence des moyennes mais il n'indique pas quels sont les groupes différents des autres !

☞ ***Pour cela il faut effectuer des tests post-hoc (comparaisons multiples 2 à 2) !***

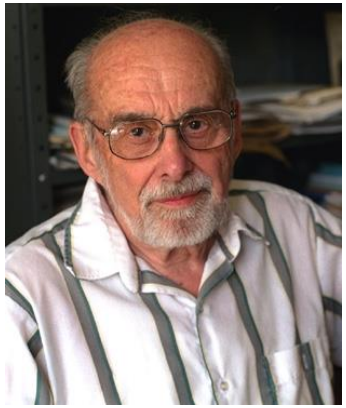
Tests post-hoc de l'ANOVA

Après un test ANOVA positif (*i.e.* « au moins un groupe différent des autres »), un **test post-hoc** est un test de comparaisons multiples qui permet de déterminer les différences significatives entre les groupes 2 à 2.

2 tests post-hoc sont notamment fréquemment utilisés :

- **Tukey HSD (honestly significant difference)** : Test t de comparaisons multiples appariées comparant toutes les moyennes entre elles (soit k×(k-1)/2 comparaisons possibles avec k groupes).
- **Dunnett** : Test T de comparaisons multiples appariées comparant toutes les moyennes des groupes expérimentaux à la moyenne d'un groupe témoin unique.

John Tukey (1915-2000)



Charles Dunnett (1921-2007)

Exemple : ANOVA (1/4)

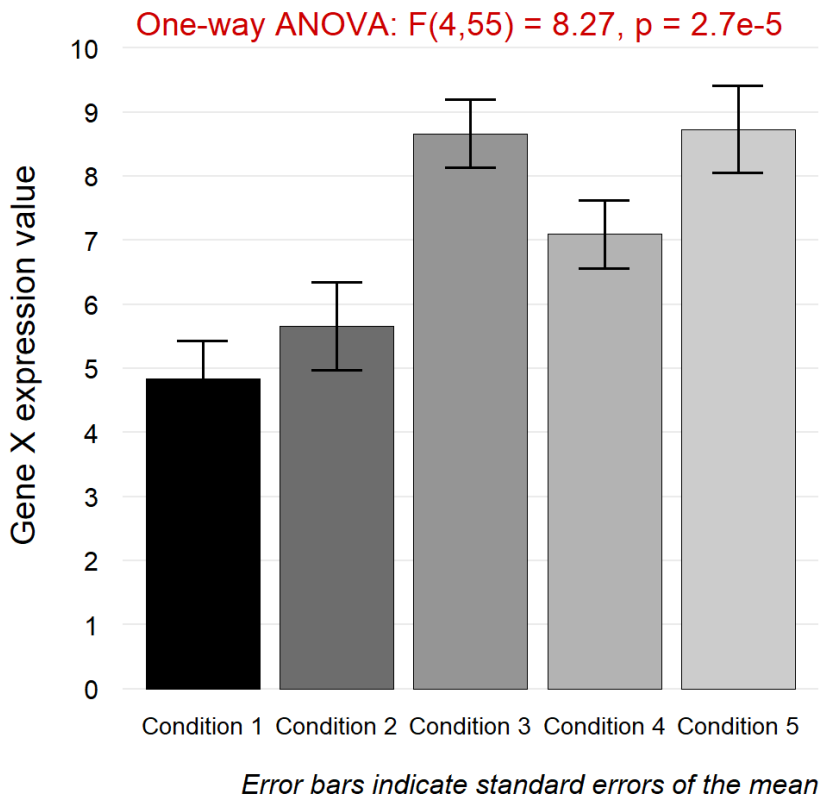
Exemple sur des données simulées comparant les niveaux d'expression d'un gène X mesurés dans 5 conditions différentes.

```
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



Exemple : ANOVA (2/4)

Comme le test ANOVA est concluant, on applique le test post-hoc de Tukey HSD pour comparer tous les traitements 2 à 2 :

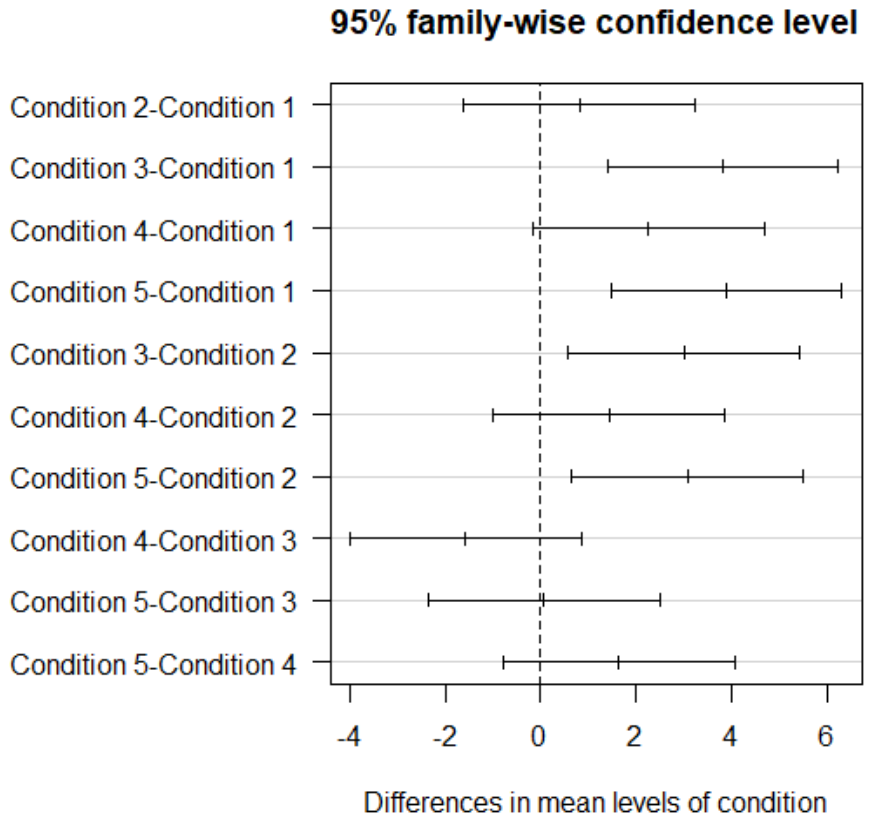
```
TukeyHSD(aov(Expression~Condition, data=dataset))
```

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



Exemple : ANOVA (3/4)

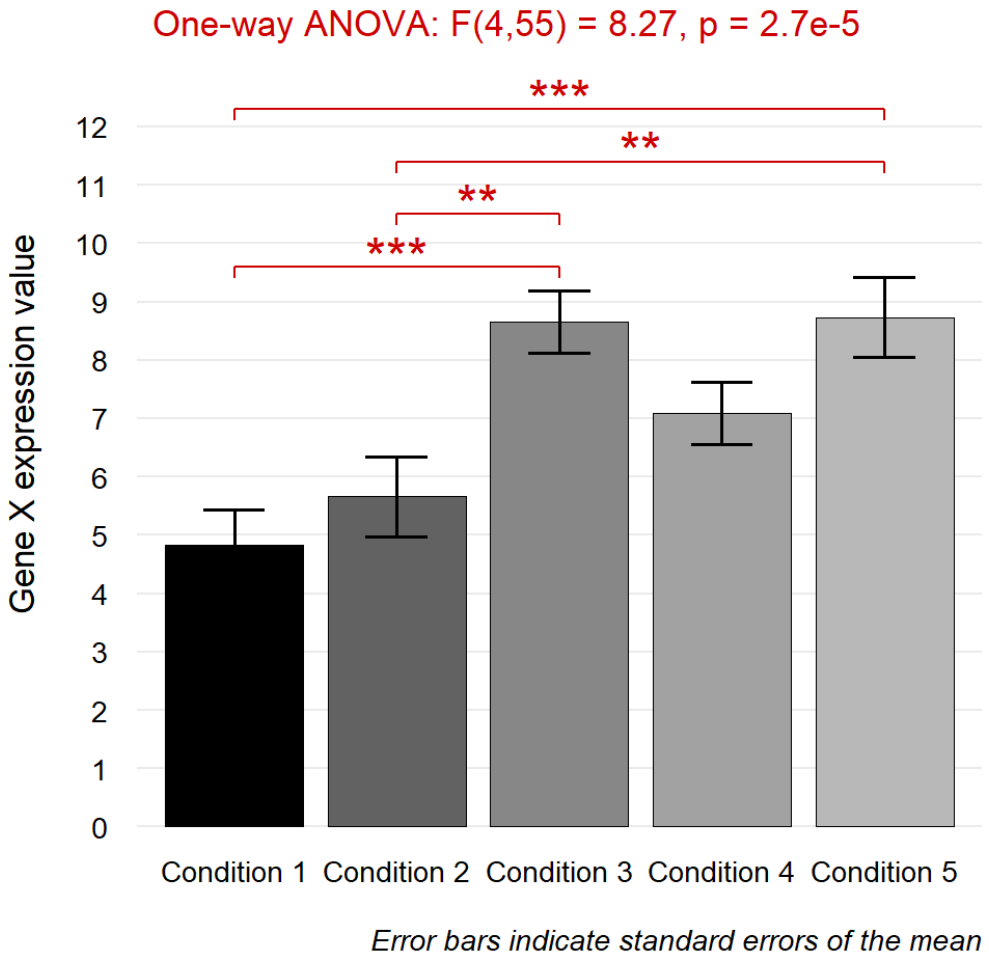
Comme le test ANOVA est concluant, on applique le test post-hoc de Tukey HSD pour comparer tous les traitements 2 à 2 :

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



Exemple : ANOVA (4/4)

Ce qu'il aurait fallu faire avant le test ANOVA : vérifier les conditions d'application !

1. **Indépendance des groupes** : on suppose que les mesures sont indépendantes au sein et entre les conditions. ☑

2. **Normalité des groupes :** on applique le test de **Shapiro-Wilk** aux 5 conditions

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

La + petite des 5 p-valeurs obtenues vaut **0,1426** ☑ ***Non rejet de l'hypothèse de normalité***

3. **Homogénéité des variances** : le test de **Bartlett** n'indique pas de différence significative de variance entre les groupes.

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

Bartlett test of homogeneity of variances

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

☑ ***Non rejet de l'hypothèse d'homogénéité***

2/ Tests d'adéquation et d'indépendance

Tests d'adéquation à une loi donnée

- Adéquation de la distribution des observations à une loi de distribution connue (**Kolmogorov-Smirnov**)

- Normalité d'une distribution (**Kolmogorov-Smirnov**, **Shapiro-Wilk**)

Test d'homogénéité

- Comparaison de la distribution d'une variable qualitative dans plusieurs échantillons (**Chi-deux** ou **test exact de Fisher**)

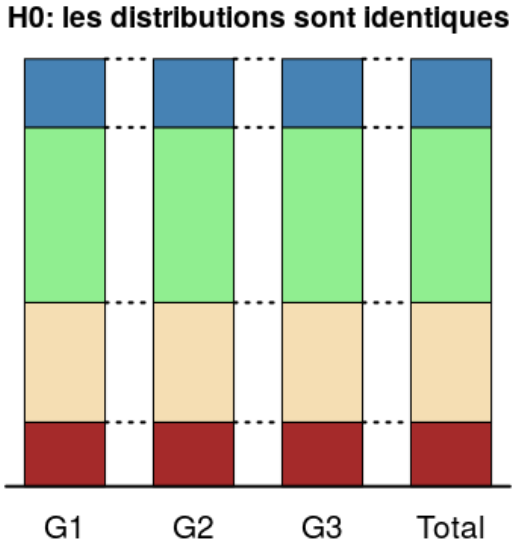
Test d'indépendance

- Étude de la distribution conjointe de 2 variables qualitatives (**Chi-deux** ou **test exact de Fisher**)

Test du χ^2 d'homogénéité

2 variables qualitatives : la notion de moyenne et de variance n'existent plus. On cherche alors à comparer 2 ou plusieurs distributions observées sur les échantillons.

Ex. Distributions de **4 catégories (l = 4)** comparées sur **3 groupes d'individus (m = 3)**. Sous H_0, les distributions sont *« toutes identiques et identiques à la distribution observée sur le total des échantillons »*.



Sous H_0, χ^2 suit approximativement une loi de $\chi^2((l-1)*(m-1))$ dès que ***n ≥ 30*** *et les* ***effectifs théoriques sont ≥ 5****.*

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

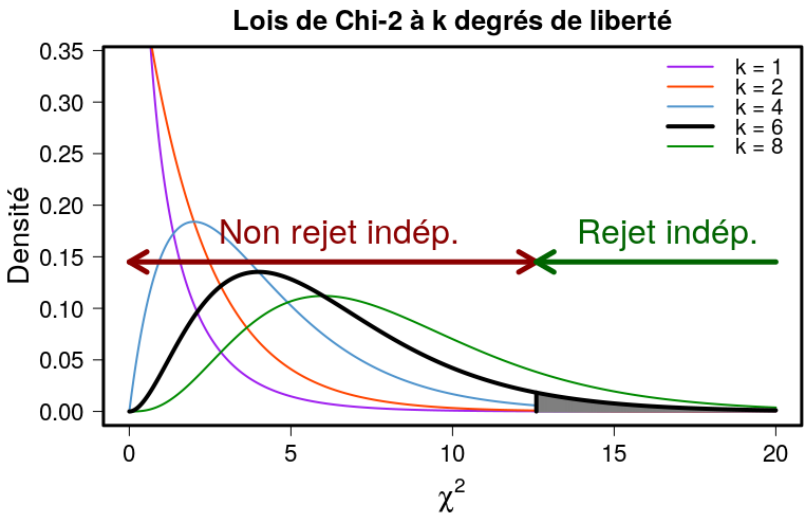


Fig. Région de rejet de l'exemple avec α = 5% et 6 ddl (courbe noire)

Exemple : Test du χ^2 d'indépendance

Étude de Cooper* sur l'efficacité de la Zidovudine (AZT) dans une cohorte de 936 sujets séropositifs au VIH, asymptomatiques avec un nombre de lymphocytes T CD4+ > 400 mm^3. La progression de la maladie sur 3 ans est définie par l'apparition des symptômes du SIDA ou une diminution importante des cellules CD4+.

TRAITEMENT	PROGRESSION	PAS DE PROGRESSION	TOTAL
AZT	76	399	475
Placebo	129	332	461
Total	205	731	936

2 variables qualitatives à 2 niveaux :

- ❑ TRAITEMENT (AZT ou Placebo)
- ❑ PROGRESSION (OUI ou NON)

H0 : Les variables TRAITEMENT et PROGRESSION sont indépendantes

```
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 :
Rejet de l'hypothèse nulle d'indépendance = effet probable du traitement

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

Application en génomique : Test d'enrichissement (1/2)

Question : Dans un ensemble de gènes donné, une fonction biologique « f » est-elle plus représentée que dans n'importe quel autre ensemble de gènes de même taille obtenu « au hasard » d'un tirage aléatoire dans l'ensemble du génome ?

Exemples de fonction : **terme Gene Ontology** (GO), **voie métabolique** (KEGG ou Reactome), ou tout autre liste de gènes associés à une fonction biologique…

Sous H_0 : Il n'existe aucun lien entre « f » et le mode de sélection de l'ensemble de gènes, autrement dit *« La fonction biologique ne caractérise pas particulièrement la liste de gènes »*.

Le test d'enrichissement fonctionnel repose sur un **test exact de Fisher** basé sur la distribution de la **loi hypergéométrique** (discrète) décrite par les 3 paramètres :

- Taille **N** de la population (génome de référence ou « background »)
- Taille **n** de l'ensemble de gènes étudié
- Probabilité ***p*** d'événement favorable dans la population (c'est-à-dire de tirer au hasard un gène associé à « f »)

Application en génomique : Test d'enrichissement (2/2)

Soient **N** la taille (connue) du génome, **E** le nombre de gènes (connu) du génome appartenant à « f » tel que ***p*** = E/N, et **n** la taille de l'échantillon étudié, le test d'enrichissement donne *a posteriori* la probabilité sous H_0 d'avoir obtenu dans l'échantillon un nombre égal ou supérieur au nombre effectivement observé **e** de gènes associés à « f » :

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

Si $P(X \geq e) < 5\%$: « f » est sur-représentée dans l'échantillon, on dit que l'échantillon étudié est « enrichi » pour la fonction « f ».

Si le test d'enrichissement est effectué pour plusieurs fonctions représentées dans l'échantillon, cela représente plusieurs hypothèses testées et les p-valeurs doivent être corrigées pour contrôler le taux de faux-positifs (Bonferroni ou 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/ Tests de normalité

Les **hypothèses de normalité** sont souvent requises dans les analyses statistiques :

- Intervalles de confiances
- Test T de Student, ANOVA
- Régression linéaire (normalité des résidus)
- *Etc.*

Il convient alors de **vérifier ces hypothèses** soit par une **approche graphique** :

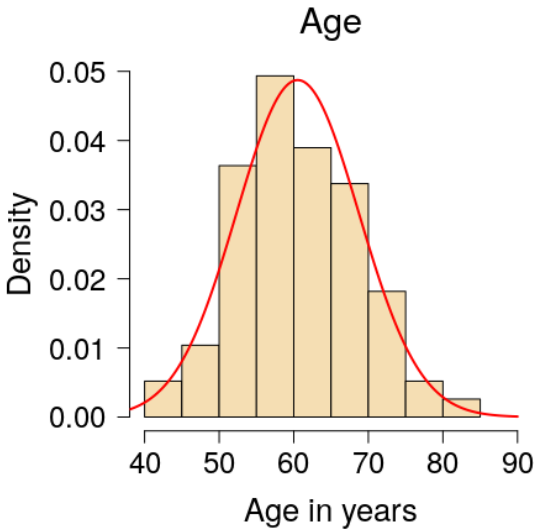
- Superposition de la densité gaussienne sur l'histogramme des observations
- Diagramme Quantile-Quantile (droite de Henry)

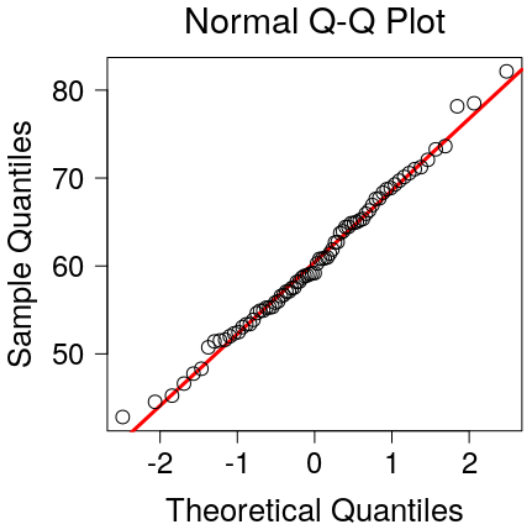
Soit à l'aide d'un **test d'adéquation à la loi normale** :

- Test de Shapiro-Wilk ($n < 50$)
- Test de Kolmogorov-Smirnov ($n > 50$)

Exemple : Tests de normalité

Âges et scores cognitifs (MMSE) de 77 sujets d'une étude clinique (données simulées) :

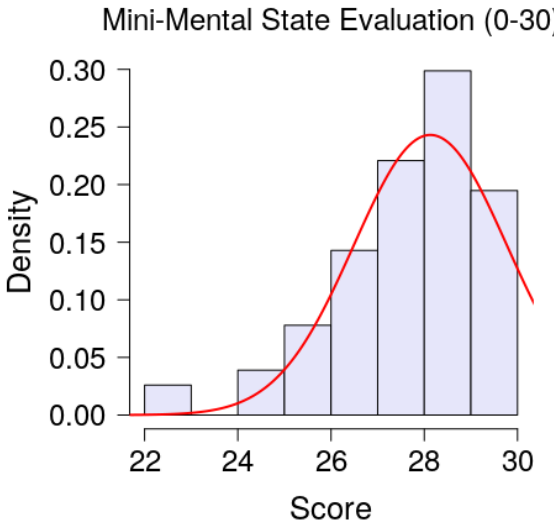


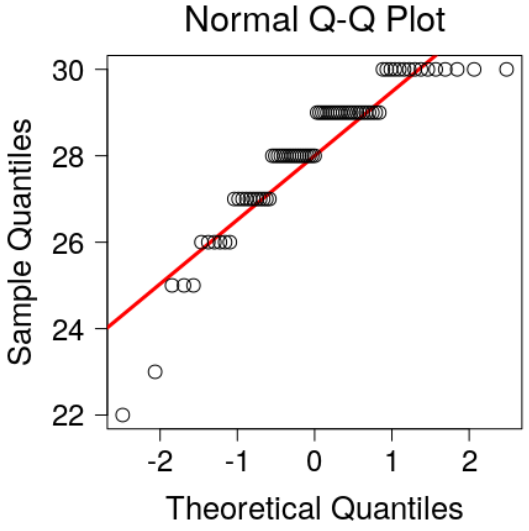



```
ks.test(age,
"pnorm",
mean(age),
sd(age))
```

One-sample Kolmogorov-Smirnov test
data: age
D = 0.060703, **p-value = 0.9226**
alternative hypothesis: two-sided

Non rejet de l'hyp. nulle de normalité






```
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

Rejet de l'hyp. nulle de normalité

4/ Tests non paramétriques

Un test non paramétrique (NP) est un test ne nécessitant pas d'hypothèse sur la distribution de la population étudiée. Il est généralement basé sur l'étude des rangs des observations, sans dépendre des moyennes et des variances des données d'origine.

Avantages :

- Utilisables lorsque certaines conditions de validité d'un test paramétrique ne sont pas vérifiées (ex. normalité, égalité des variances…)
- Tests adaptés aux petits échantillons ($n < 30$)
- Tests adaptés aux variables ordinales (ex. degré de satisfaction)

Inconvénients :

- Lorsque les conditions de validité sont vérifiées : les tests NP sont moins puissants que les tests paramétriques
- Difficultés d'interprétation car on ne compare plus des paramètres (moyenne, proportion, variance…)

☞ ***La plupart des tests paramétriques ont des tests non-paramétriques équivalents.***

Tests non paramétriques sur les rangs

Dans le cas de petits échantillons et distributions non gaussiennes (*cf.* tests d'adéquation), on peut utiliser une **stratégie de test remplaçant les valeurs des observations par leurs rangs** :

2 échantillons indépendants :

- **Test de Wilcoxon-Mann-Whitney**
  (également appelé **Test U de Mann-Whitney**
  ou **Test de la somme des rangs de Wilcoxon**)

2 échantillons appariés :

- **Test des rangs signés de Wilcoxon**

Plusieurs échantillons :

- **Test de Kruskal-Wallis (+ test post-hoc de Dunn)**

Extrait de
https://www.math.univ-toulouse.fr/~besse/Wikistat/pdf/st-l-inf-tests.pdf

Exemple : Test de Wilcoxon-Mann-Whitney

But : **comparer 2 groupes A et B de patients** dont on a étudié la cytorachie (présence de cellules dans le liquide cérébrospinal mesurée en nombre de cellules par µL).

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

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

Principe du test : *sous H_0 les valeurs de A (rouges) et B (bleues) ordonnées sur les rangs sont mélangées de façon homogène (les groupes A et B sont équivalents).*

```
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

Exemple extrait de T. Ancelle, Statistique épidémiologie 3ème édition, Maloine, 2012

Tests non paramétriques de permutation

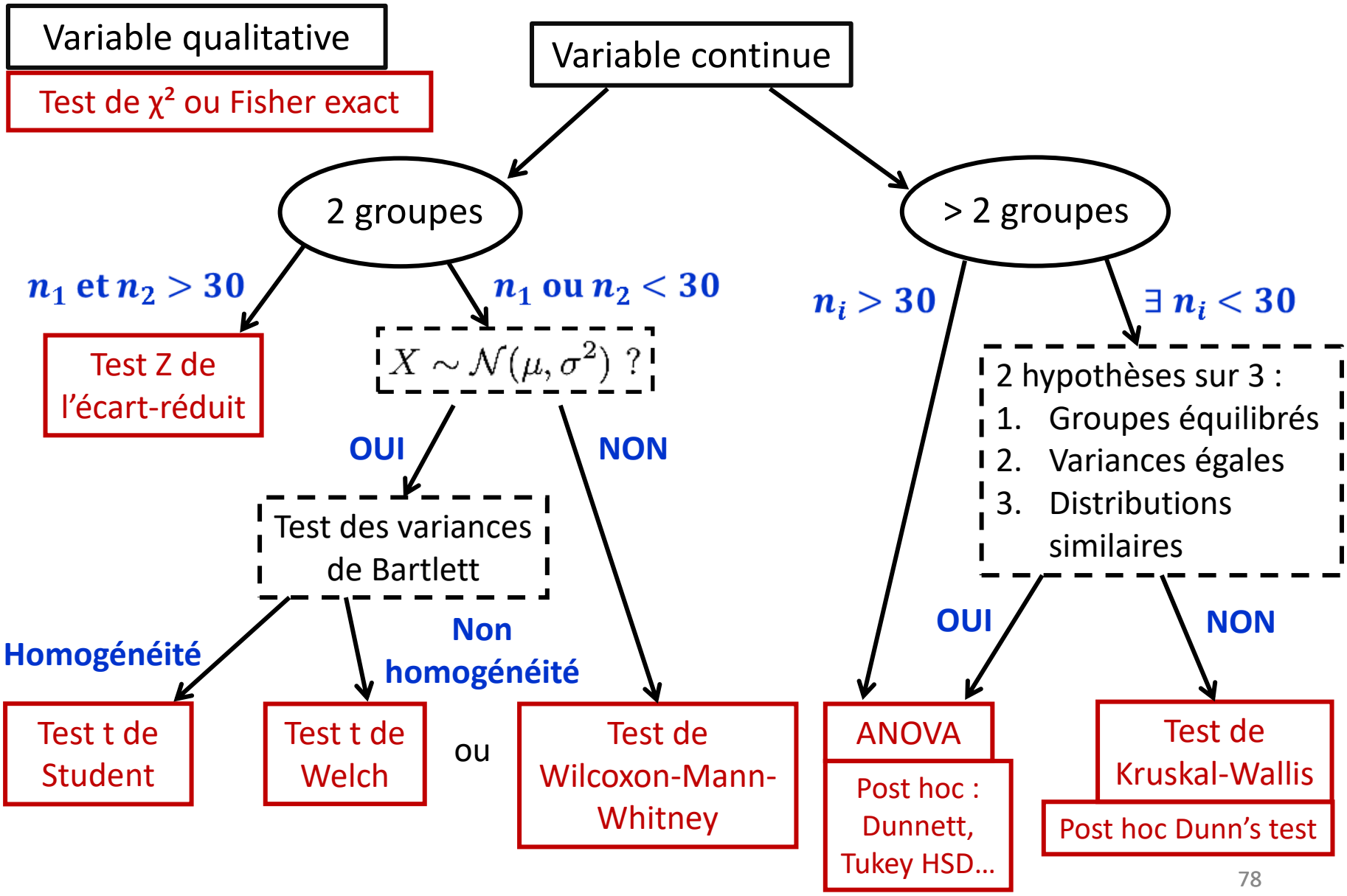
Dans le cas de distributions non gaussiennes, les **tests de permutation** sont des approches robustes basées sur le **rééchantillonnage des observations**. Ainsi, la significativité ne repose plus sur la distribution théorique de la statistique sous H_0, mais sur une **distribution empirique** calculée à partir d'un grand nombre de permutations (100, 1000 ou +).

Exemple : pour comparer les moyennes de 2 groupes par un test bilatéral de permutation, l'hypothèse H_0 d'*« égalité des 2 moyennes »* peut se traduire par *« les observations sont interchangeables entre les 2 groupes »*.

La procédure de test est alors la suivante :

1. Calculer la vraie valeur (non permutée) T_0 de la statistique de test
2. Générer un grand nombre de permutations aléatoires des individus entre les 2 groupes
3. Calculer la valeur de la statistique T de test pour chaque permutation
4. Déduire des T « permutées » la distrib. et les quantiles emp. 2,5% et 97,5% sous H_0
5. Comparer T_0 aux quantiles emp. et conclure

Comparaison de groupes



Partie 3 : Modéliser

Régression et modélisation

Dans cette partie consacrée aux modèles de régression, on s'intéresse plus particulièrement au problème d'une variable « à expliquer » Y qui a été conjointement observée avec une ou plusieurs variables « explicatives » X sur les mêmes individus.

La problème de modélisation peut alors se définir par la recherche d'une représentation simplifiée de Y à l'aide des variables X en vue de la décrire, de l'expliquer ou de prédire ses valeurs.

Du point de vue mathématique, il s'agit de déterminer une fonction *f* selon un critère prédéfini, capable d'approcher les valeurs de Y à partir des valeurs observées de X :

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

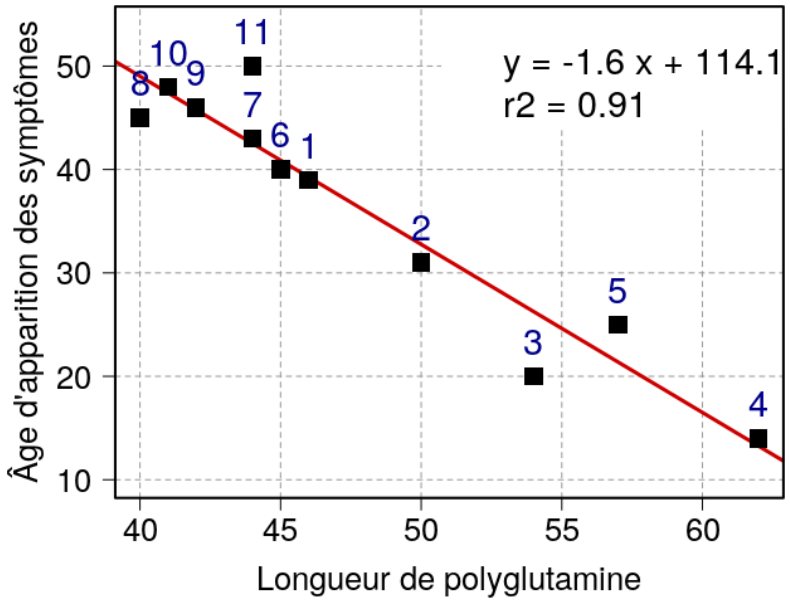
où ε représente le bruit ou l'erreur de mesure.

Régression linéaire simple

Exemple de régression simple : on cherche à expliquer l'âge d'apparition des premiers symptômes de 11 patients atteints d'ataxie spinocérébelleuse SCA1 à partir de la longueur de polyglutamine (répétitions de CAG).

Variable explicative — Variable à expliquer

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



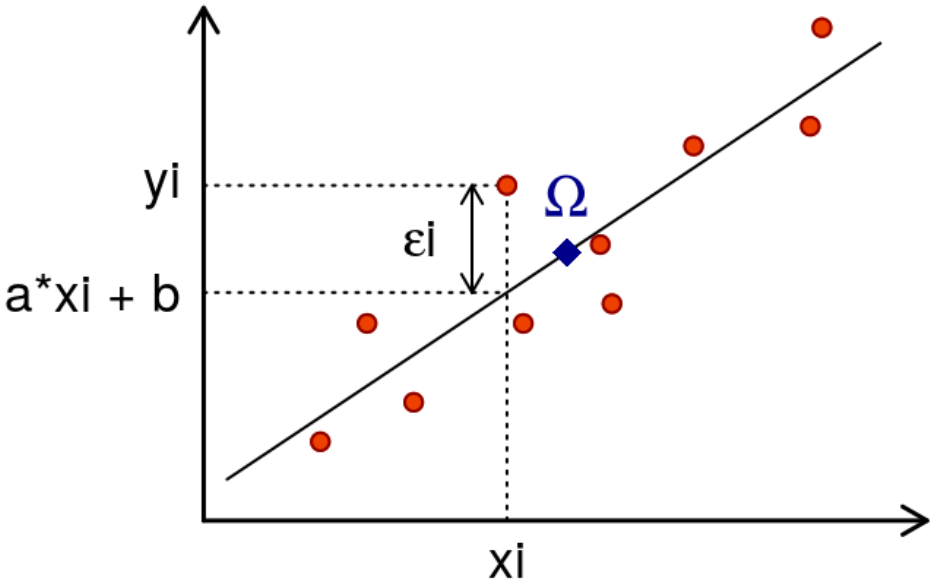
Modèle de régression simple : $y_i = a \times x_i + b + \varepsilon_i$

ε_i : termes d'erreur indépendants et identiquement distribués de loi $\mathcal{N}(0, \sigma_\varepsilon^2)$

Problème : **Estimation de a, b et σ_ε ?**

Source : projet BIOSCA

Estimateurs des MCO (moindres carrés ordinaires)



Le **critère des MCO** consiste à trouver les valeurs de a et b qui minimisent la somme des erreurs au carré :

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

Estimateurs MCO :

$$\begin{cases} \hat{a} = \dfrac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2} = \dfrac{\widehat{\mathrm{Cov}}(X,Y)}{\widehat{\mathrm{Var}}(X)} = r_{XY}\dfrac{\hat{\sigma}_Y}{\hat{\sigma}_X} & \text{(pente de la droite)} \\ \hat{b} = \bar{y} - \hat{a} \times \bar{x} & \text{(la droite passe par le centre de gravité } \Omega \text{ du nuage de points)} \end{cases}$$

Valeurs prédites des y_i : $\hat{y}_i = \hat{a} \times x_i + \hat{b}$ **Résidus :** $\hat{\varepsilon}_i = y_i - \hat{y}_i$

Analyse de variance et coefficient de détermination

Sachant que le critère MCO $0 \leq S \leq +\infty$; il s'agit de déterminer un critère de qualité de la régression.

Pour cela, on se base sur la **formule de décomposition de la variance** :

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

$$\mathrm{SST} = \mathrm{SSE} + \mathrm{SSM}$$

***SST** : Somme des Carrés Totaux*

***SSE** : Somme* des Carrés Résiduels

SSM : Somme des Carrés expliqués par le Modèle

On définit alors le **coefficient R^2 de détermination** :

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

R^2 proche de 1, le modèle est excellent
R^2 proche de 0, le modèle ne sert à rien !

et le **coefficient de corrélation linéaire multiple** : $R = \sqrt{R^2}$

Dans le cas d'un modèle simple (1 seul régresseur), on a : $r_{XY} = sign(\hat{a}) \times R$

Validation du modèle

Pour **valider le modèle de régression**, il reste à vérifier les hypothèses sur les aléas (termes i.i.d. et gaussiens). Cette vérification se fait généralement à l'aide de tests graphiques :

- Graphe des résidus versus les valeurs prédites : ne doit pas présenter de structure (indépendance, homoscédasticité/variance constante, normalité) ;
- Normalité : Histogramme et QQ-plot (points alignés sur la 1ère bissectrice) ;
- Indépendance : autocorrélation des résidus (acf), test de Durbin Watson…

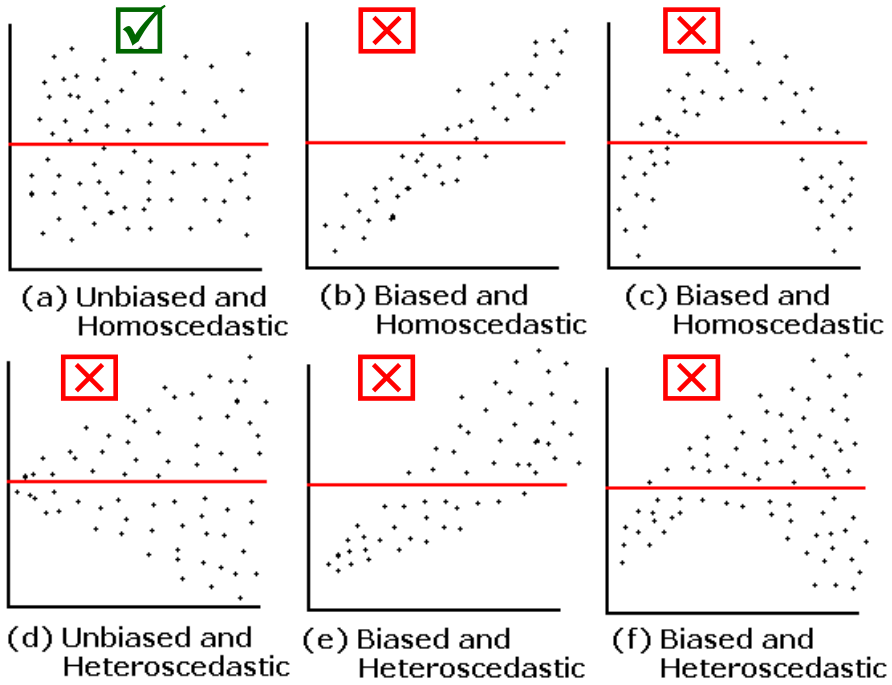


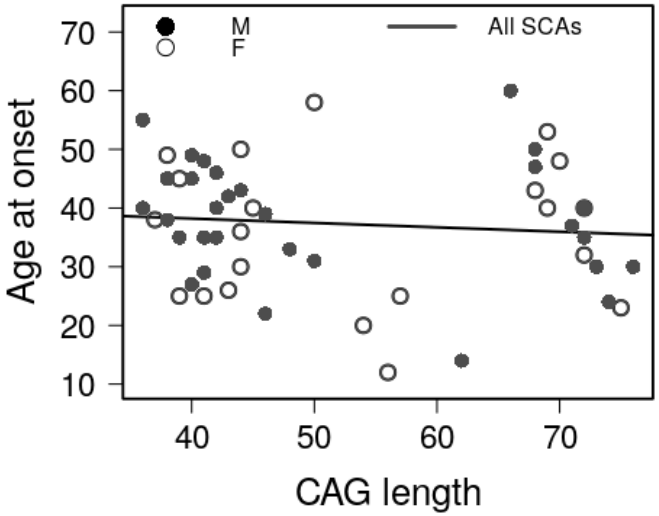
Fig. Graphes résidus (axe y) vs valeurs prédites (axe x).
Cas (a) valide la régression linéaire.
Cas (b-f) indiquent soit un problème d'hétéroscédasticité (d-e-f), soit une structure qui n'est pas prise en compte par le modèle (b-c-e-f).

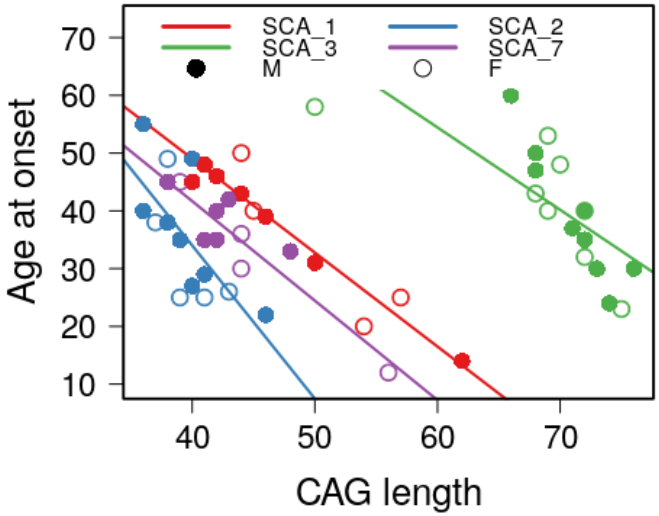
Solutions : utilisation d'estimateurs « robustes » des MCO, transformation des données, ajustement des données par d'autres types de modèles non linéaires.

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

Se méfier des apparences…

Ne pas prendre en considération les différentes formes génétiques de SCA conduirait à conclure (un peu hâtivement…) que la longueur de polyglutamine n'influe pas sur l'âge d'apparition de la maladie !





	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 (***)

Tableau : Estimation et significativité des pentes de régression

Source : projet BIOSCA

Modèle linéaire

Pour **1 variable dépendante quantitative** associée à **1 ou plusieurs variables explicatives quantitatives et/ou qualitatives**, on peut remarquer que régression linéaire et ANOVA sont les cas particuliers du même modèle statistique appelé **modèle linéaire** :

a) 1 var. explicative quantitative : **régression linéaire simple**

b) ≥ 2 var. explicatives quantitatives : **régression linéaire multiple**

c) 1 var. explicative qualitative : **t-test (2 niveaux)** ou **ANOVA à 1 facteur**

d) ≥ 2 var. explicatives qualitatives : **ANOVA à plusieurs facteurs**

e) Combinaison de var. explicatives quantitatives et qualitatives : **ANCOVA**

De manière générale, l'utilisation du modèle linéaire permet aussi d'évaluer la liaison en présence de *covariables* (quantitatives) ou *cofacteurs* (qualitatifs). Un modèle « ajusté » permet ainsi de séparer les effets des covariables ou cofacteurs de celui de la variable d'intérêt déterminé à l'aide d'une **p-valeur ajustée**.

Cas simples de régressions non linéaires

Les modèles basés sur les fonctions « exponentielle » et « puissance » sont 2 cas simples de **relations non linéaires**, parce qu'ils sont **linéarisables** à l'aide de la fonction logarithme :

Modèle exponentiel du type : $\boxed{y = B \cdot \mathrm{e}^{\alpha x}}$, $(B, \alpha) \in \mathbb{R}^2$.

Passage aux logarithmes : $\ln y = \ln B + \alpha x$

$\Leftrightarrow Y = \alpha X + \beta$ en posant $Y = \ln y$, $X = x$ et $\beta = \ln B$.

Modèle puissance du type : $\boxed{y = B \cdot x^{\alpha}}$, $(B, \alpha) \in \mathbb{R}^2$.

Passage aux logarithmes : $\ln y = \ln B + \alpha \ln x$

$\Leftrightarrow Y = \alpha X + \beta$ en posant $Y = \ln y$, $X = \ln x$ et $\beta = \ln B$.

Modèle linéaire généralisé

Comment prédire une variable Y à valeurs discrètes à partir d'une variable explicative X discrète ou continue ?

Le **modèle linéaire généralisé** (GLM) englobe le modèle de régression linéaire, ainsi que d'autres modèles intéressants permettant, par exemple, de prédire une ***variable dépendante Y à valeurs discrètes*** à partir d'1 ou plusieurs variables explicatives continues ou discrètes.

Exemple 1 – Y variable binaire 0/1 : on souhaite expliquer l'absence (Y = 0) ou la présence (Y = 1) d'une maladie coronarienne en fonction de l'âge chez 100 sujets.

Exemple 2 – Y variable de comptage : on souhaite quantifier l'évolution d'un nombre de bactéries en fonction du temps.

Exemple 3 – Y variable de durée de vie : on souhaite étudier la durée de vie en semaines du diagnostic au décès en fonction du nombre de globules blancs (échelle log10) à t_0 pour 33 patients atteints de leucémie.

Formulation des GLM

Les modèles linéaires généralisés sont formés de **3 composantes** :

- **Composante aléatoire** : Variable dépendante Y associée à une loi de probabilité
- **Composante déterministe** : Prédicteur linéaire ou combinaison linéaire des variables explicatives X_1, …, X_k : $\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k$
- **Fonction de lien** : Fonction *g* décrivant la relation fonctionnelle entre le prédicteur linéaire et l'espérance mathématique de la variable dépendante Y

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

Distribution	Type de données	Type de GLM	Fonction de lien (g)
Normale	Var. ~ Normale	Modèle linéaire	Identité : g(y) = y
Poisson	Comptage	Modèle log-linéaire	Log : g(y) = log(y)
Binomiale	Pourcentage	Régression logistique	Logit : g(y) = log(y/(1-y))
Gamma	Durée	Modèle Gamma avec fonction de lien inverse	Inverse : g(y) = 1/y

Estimation des paramètres

Contrairement au cas de la régression linéaire pour laquelle l'estimation des coefficients est basée sur la méthode des moindres carrés ordinaires ***[soit la minimisation des écarts au carré entre la réponse observée et la réponse prédite]***, les paramètres du GLM sont déterminés par une autre méthode d'estimation dite **méthode du maximum de vraisemblance**.

Principe de la méthode du maximum de vraisemblance :

Trouver les valeurs des paramètres qui maximisent la **« probabilité que les valeurs observées de Y se réalisent conditionnellement aux paramètres supposés connus »** :

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

Les logiciels calculent les estimations à l'aide d'un algorithme itératif (Fisher scoring ou Newton-Raphson) où partant d'une valeur initiale (fixée) des paramètres, la valeur est réactualisée à chaque itération jusqu'à convergence de l'algorithme.

Cas de la régression logistique binaire

L'interprétation du modèle logistique est relativement simple dans la mesure où son écriture fait intervenir une cote (Odds) exprimant le ***rapport des chances entre les événements Y = 1 et Y = 0 sachant la valeur de X*** :

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

Le coefficient $e^{\beta 1}$ s'interprète alors en terme d'**Odds-Ratio** (OR) traduisant l'évolution du rapport des chances d'apparition de l'événement Y = 1 contre Y = 0 lorsque X passe de x à x + 1 :

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

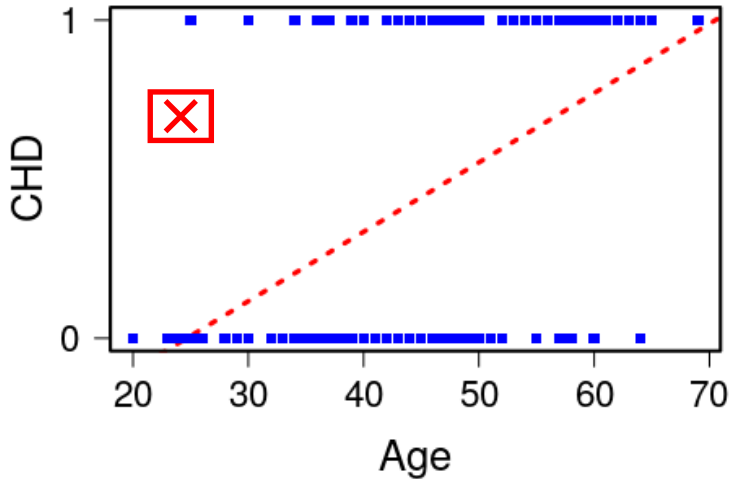
3 tests sont disponibles pour évaluer l'apport de la variable *X* au modèle :

- Test de Wald
- Test du rapport des vraisemblances
- Test du score

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

Exemple 1 : Régression logistique binaire

But : étudier la relation entre l'âge et la présence (1) ou absence (0) d'une maladie coronarienne (CHD) dans une population de 100 individus.



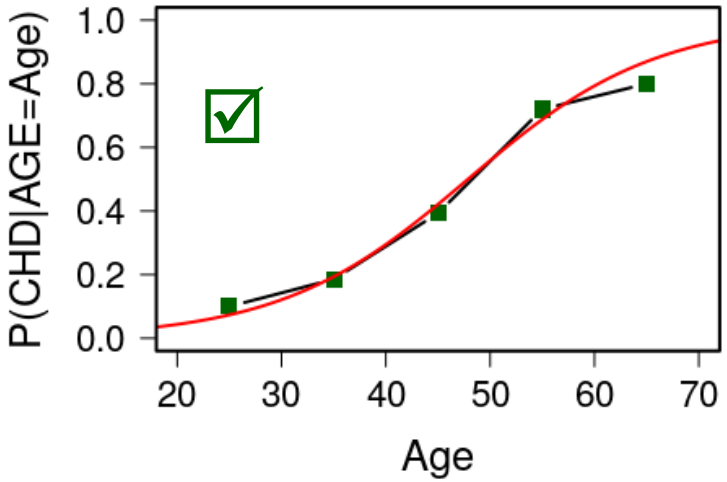


Fig. de gauche : pas d'ajustement linéaire possible entre l'âge et la variable CHD binaire.

Fig. de droite : l'allure sigmoïdale (courbe en S) du lien entre l'âge et l'espérance conditionnelle de CHD convient bien à la fonction logistique.

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

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

Test de Wald

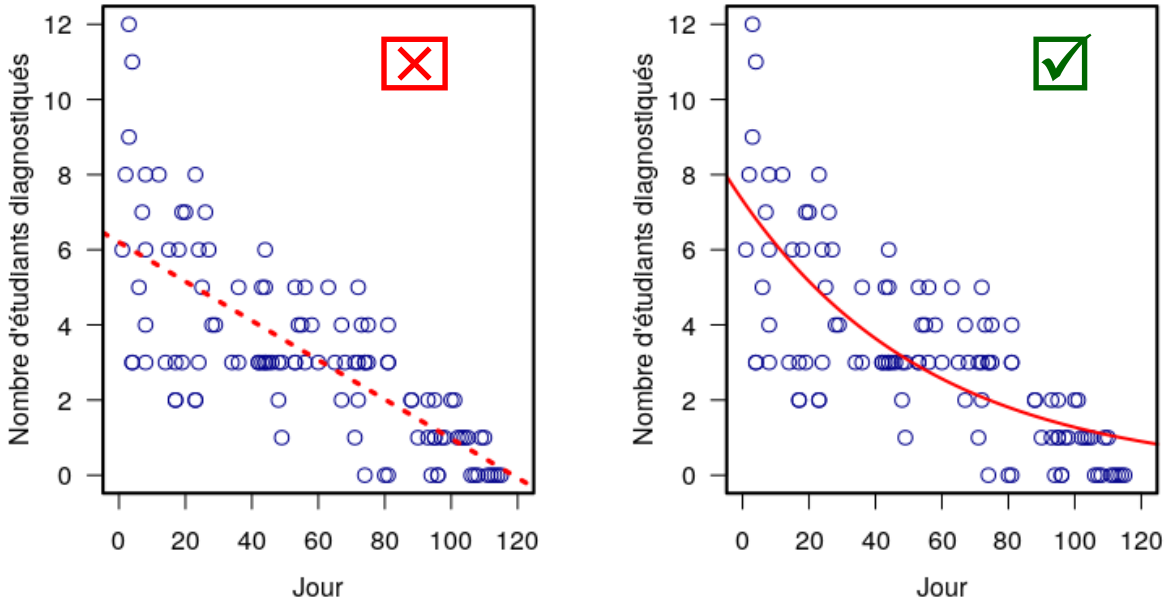
Coefficients:

	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	***

Exemple extrait de Hosmer and Lemeshow, Applied Logistic Regression, Wiley, 2nd edition (2000)

Exemple 2 : Régression de Poisson

But : nombre d'étudiants diagnostiqués pour une maladie infectieuse depuis le premier jour d'épidémie.




```
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 ***

Le comptage moyen de nouveaux cas en fonction du temps suit le modèle :

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

Exemple 3 : Régression « Gamma »

But : étudier le temps de coagulation du sang (sec) lorsque l'on ajoute de la thromboplastine (2 lots testés) au plasma à 9 différentes concentrations (%).

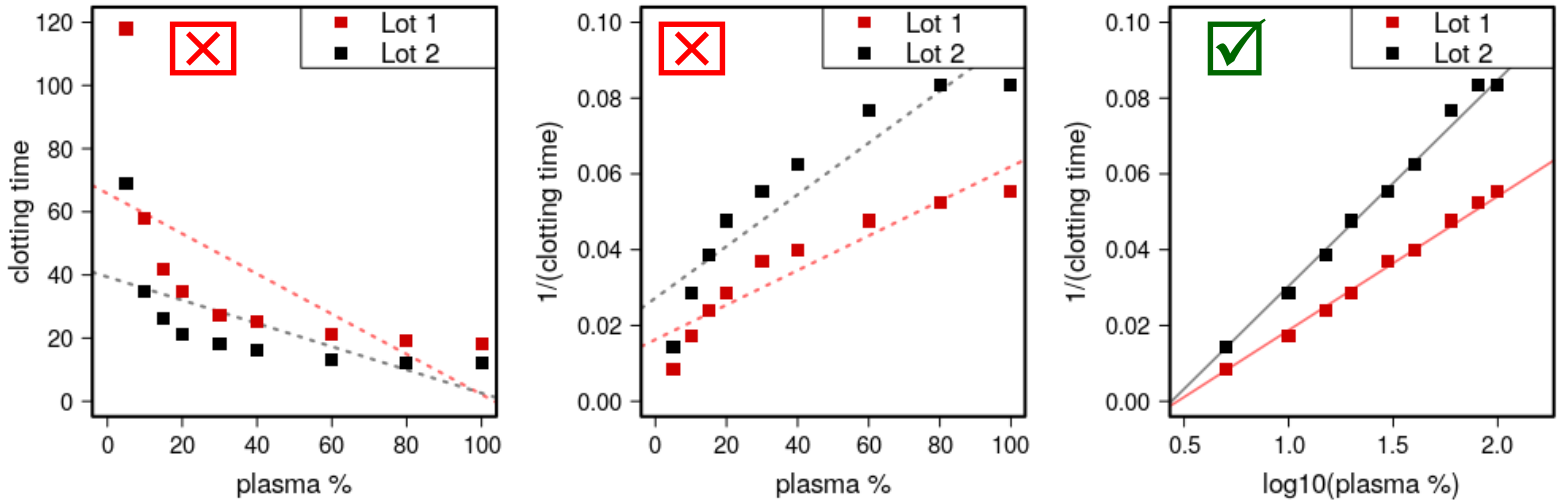


Fig. Etapes de la modélisation par un modèle Gamma utilisant la fonction de lien inverse.

```
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.

Notion de mesures répétées

Répétition des mesures d'une variable effectuées sur les mêmes individus d'un échantillon. Il s'agit bien souvent de données successives collectées au cours du temps (**études longitudinales**).

Les mesures répétées se caractérisent par le fait qu'elles introduisent une corrélation entre les mesures provenant d'un même individu (source de variabilité intra-sujet). En modélisation, l'**autocorrélation entre les résidus** découlant de ces mesures répétées contredit alors l'hypothèse fondamentale d'indépendance des erreurs sur laquelle se fonde le modèle linéaire.

Avantages :

- pour étudier une dynamique d'évolution
- plus d'observations disponibles pour un même nombre d'individus

Inconvénients (perte de l'indépendance des observations) :

- 2 observations issues d'individus différents sont indépendantes
- 2 observations issues d'un même individu ne le sont pas

Modèles mixtes

Outre le cas des mesures répétées, on peut être confronté à des situations où plusieurs observations sont « naturellement groupées » dans une étude.

Ex. Cas d'études familiales impliquant plusieurs membres de la même famille, études effectuées sur des jumeaux, études cliniques multicentriques où les données sont collectées sur les patients de différents hôpitaux, *etc.*

Par rapport au modèle linéaire (généralisé) classique, l'utilisation de modèles plus compliqués, appelés **modèles à effets mixtes**, permet de combiner des effets « fixes » à des effets « aléatoires » pour prendre en considération la corrélation des individus « groupés » ensemble (par ex. patients d'un même centre), la corrélation intra-individuelle de données longitudinales ou encore un mélange des deux (mesures au cours du temps pour les mêmes sujets de différents groupes de sujets).

Effets fixes et effets aléatoires

Difficile de trouver des définitions consensuelles aux **effets fixes et aléatoires** d'un problème de modélisation, on en donne ici une présentation simplifiée pour aider aux applications.

Effet fixe : les données proviennent de tous les niveaux possibles d'une variable qualitative dont on souhaite étudier l'impact spécifique sur la variables réponse.

Exemple : dans l'étude de comparaison de performances à un test cognitif (variable réponse) entre un groupe de sujets malades et un groupe de sujets sains, le facteur groupe est un effet fixe dont on souhaite étudier l'impact sur la variable réponse.

Effet aléatoire : les données proviennent seulement d'un échantillon aléatoire de tous les niveaux possibles d'une variable qualitative (typiquement un facteur de regroupement) dont l'impact spécifique sur la variable réponse n'a pas d'intérêt particulier, mais dont on souhaite néanmoins contrôler l'effet dans le modèle.

Exemple précédent : si les données viennent de 5 centres hospitaliers, le centre peut constituer un effet aléatoire, 1/ parce que l'étude n'implique pas tous les hôpitaux de France, et 2/ parce que le contexte expérimental pourrait varier d'un centre à l'autre.

De la modélisation à la prédiction…

Pour passer de la modélisation à la **prédiction**, il faut d'abord étudier la **propriété de généralisation** du modèle ; c'est-à-dire la capacité du modèle à pouvoir effectuer des prédictions robustes sur de nouvelles données.

Pour évaluer cette capacité, on utilise 2 jeux de données indépendants :

- **1 training set** : jeu de données servant à « entraîner le modèle »
- **1 test set** : jeu de données indépendant à prédire afin d'évaluer l'erreur de prédiction du modèle (selon un critère prédéfini)

L'étape d'apprentissage doit éviter autant que possible les phénomènes de **sur-apprentissage (overfitting)** ou de **sous-apprentissage (underfitting)**.

Le sur-apprentissage traduit une situation où le modèle s'ajuste trop bien aux données d'entraînement, le rendant ainsi moins flexible pour s'appliquer à de nouvelles données. *A contrario*, on parle de sous-apprentissage lorsque le modèle n'est pas assez complexe pour décrire correctement la relation entre les variables à partir des données d'entraînement.

TO BE CONTINUED…

Avec l'équipe Biostatistique de l'ICM :

Sana Rebbah

Baptiste Crinière-Boizet

Gaspard Martet