Option Bio-Info/Bio-Stat

S. Granjeaud

Licence professionnelle Métiers du décisionnel et de la statistique

PARCOURS : INFORMATIQUE DÉCISIONNELLE, STATISTIQUES ET BIG DATA

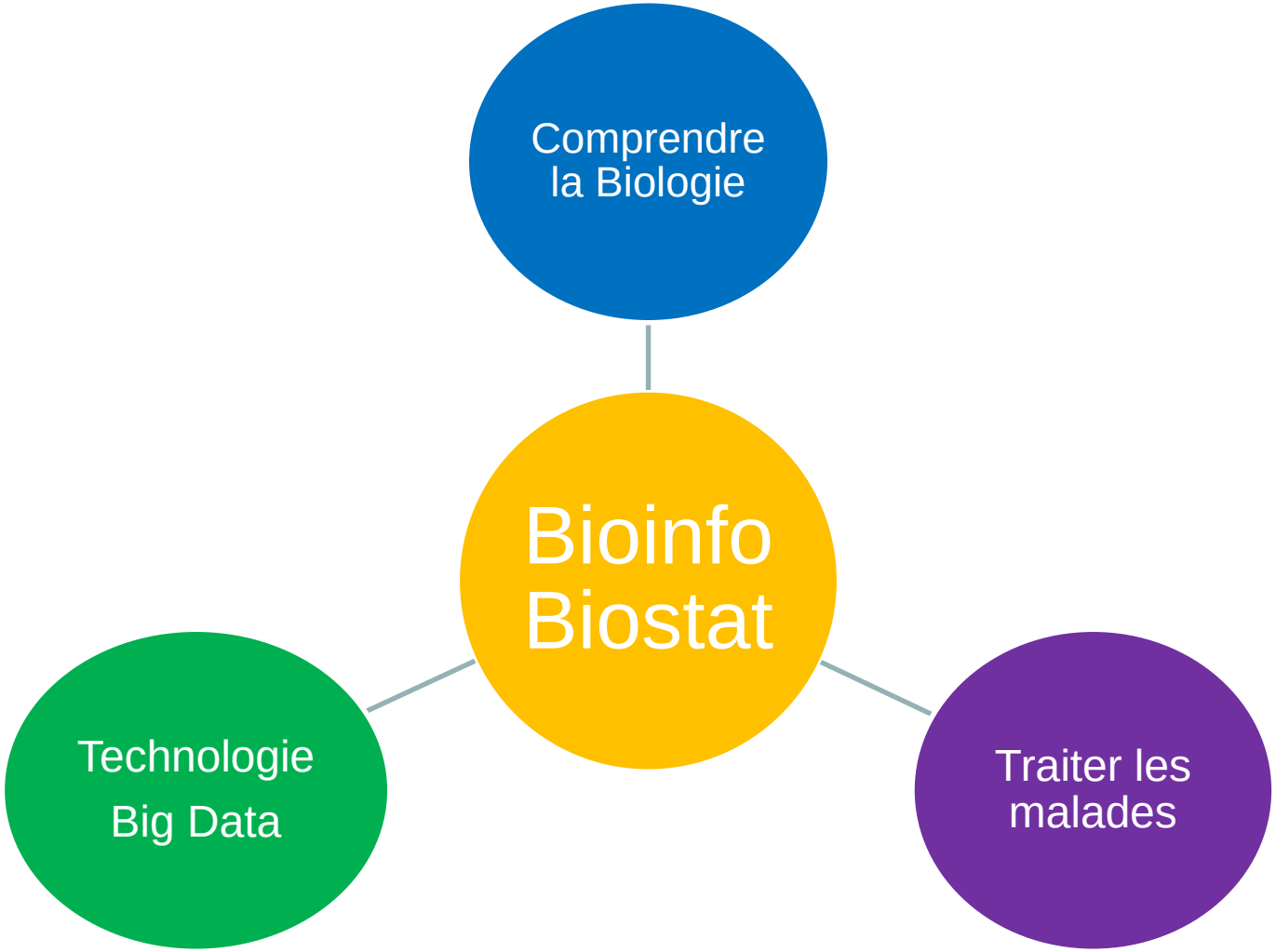
LP MDS

CRCM Cancer Research Center of Marseille 50 years of research and innovation

Aix*Marseille université Socialement engagée

INFORMATIQUE ET STATISTIQUES POUR LA BIOLOGIE ET LA MÉDECINE

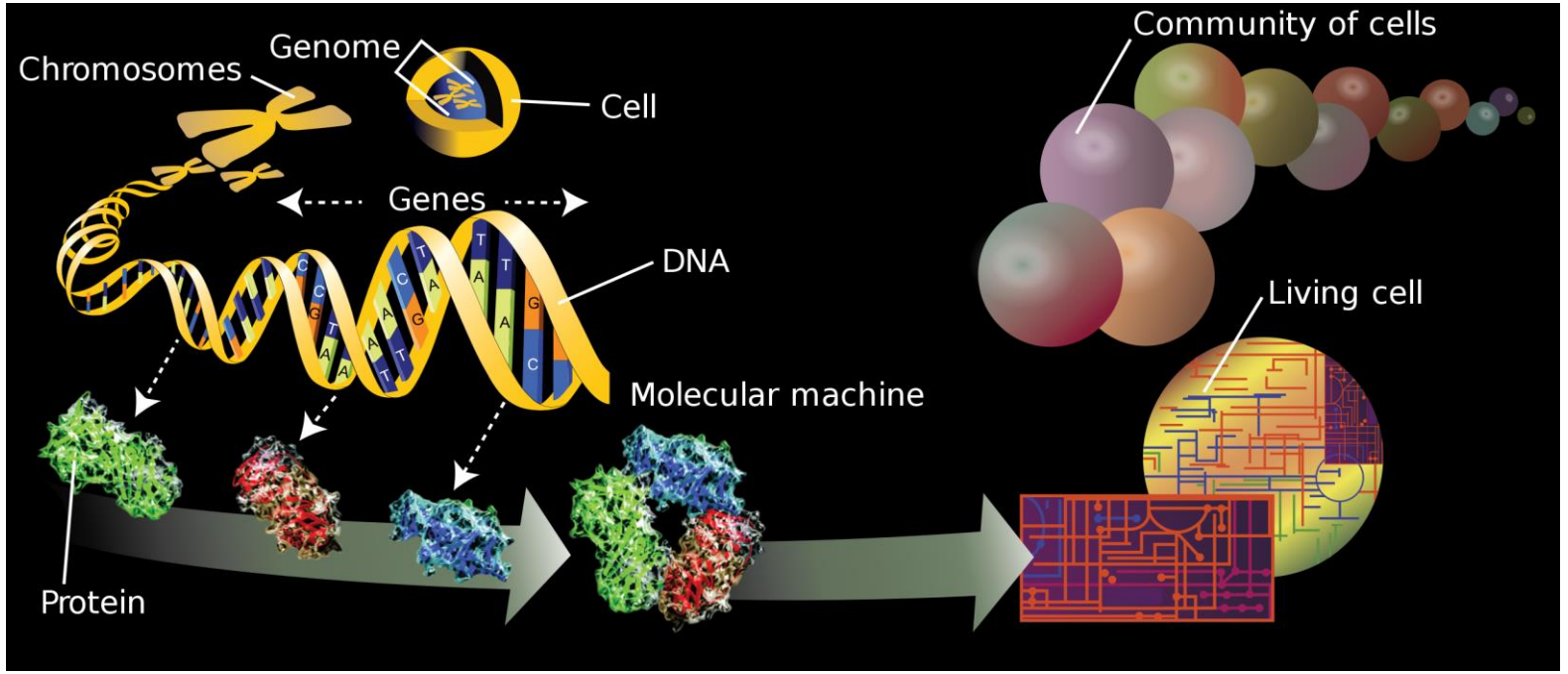
Contexte


Comprendre
la Biologie
Bioinfo
Biostat
Technologie
Big Data
Traiter les
malades

Bioinformatique

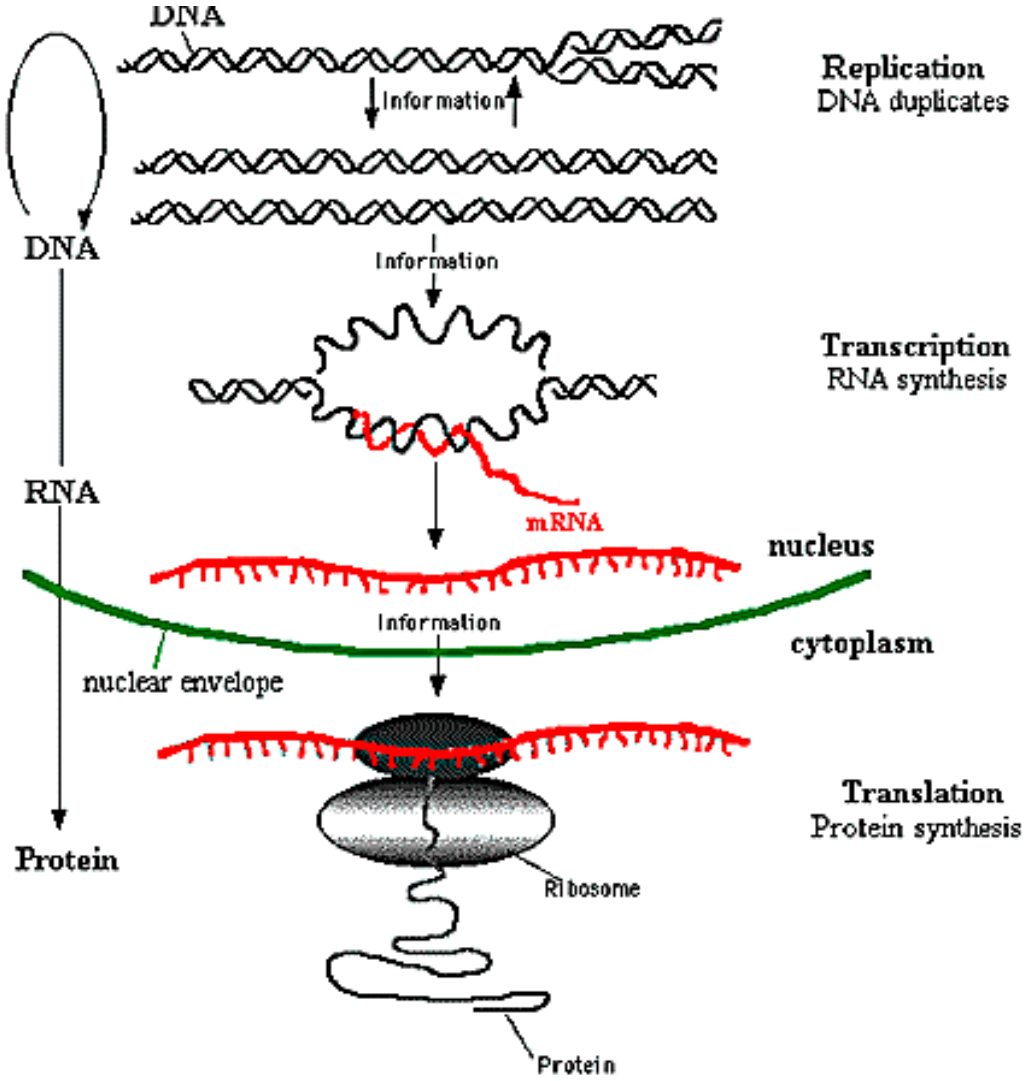
- BIO logie
- INFOR mation
- traitement auto MATIQUE

Biology



Wikipedia omics

Different Kinds of “Omes”

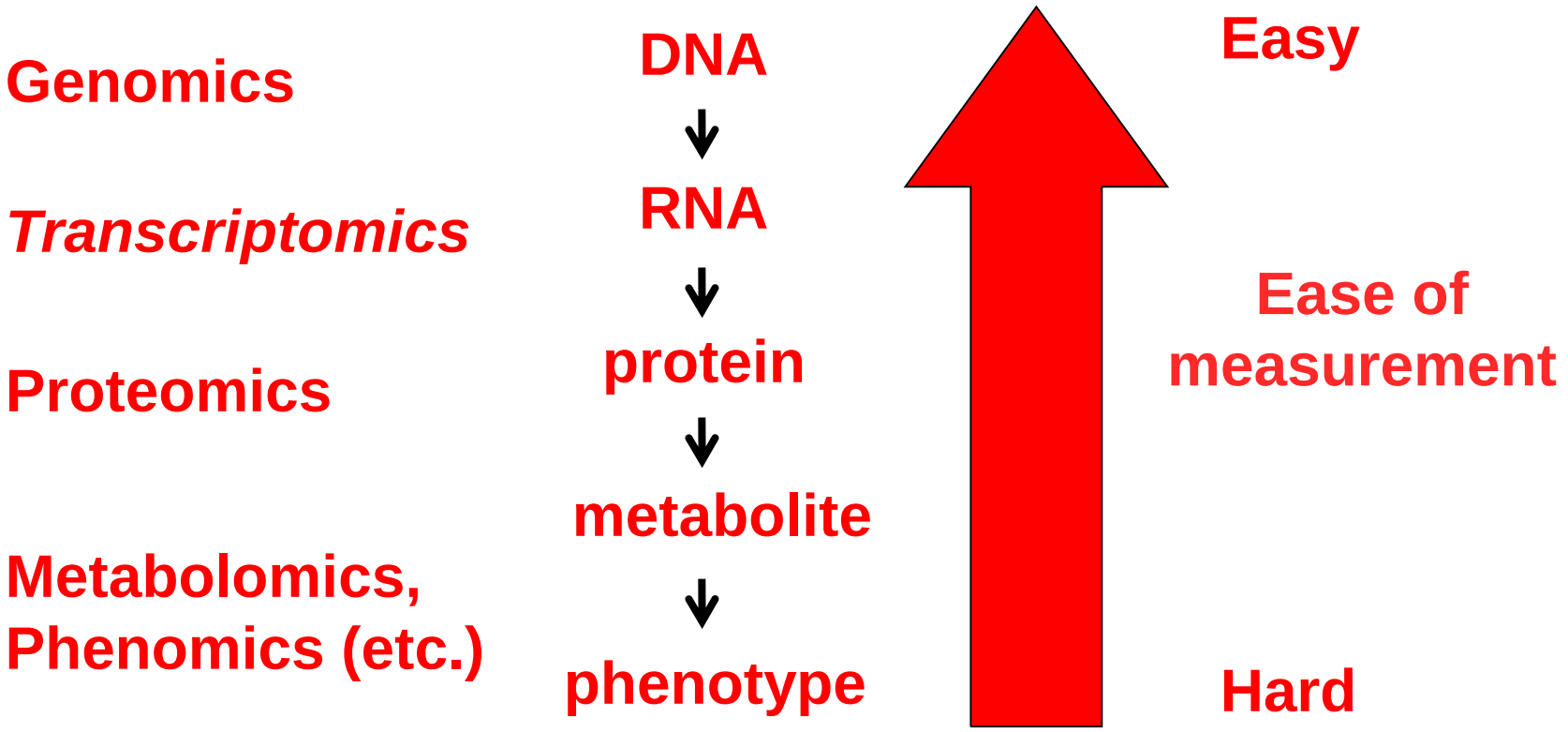


Genome

Transcriptome

Proteome

High Throughput Measurement



Toutes ces analyses sont complémentaires !

Technologie - Instrumentation

- Transcriptome par puces :
    - 1 mesure par sonde, 100k sondes par puce,
      10zaine d'échantillons
- Protéome par spectrométrie de masse :
    - 1 mesure par sonde, 1k protéines par échantillon,
      10zaine échantillon
- Cytométrie :
    - 15-50 paramètres par cellules,
      100k-1000k cellules par tube, 10-100 tubes/échantillons
- Séquençage haut débit :
    - "Bulk" = en masse, des 10 milliers de cellules : RNAseq
    - "SingleCell" = cellule unique, milliers de cellules : scRNAseq
    - Téra Octets de données par semaine

Bioinformatique & Analyse de données

- Emergence des études à l'échelle du génome complet (haut flux).
- Médecine personnalisée :
  - Altérations génomiques
  - Modulation du niveau transcriptomique
  - Dysfonctionnement protéique

- Besoin d'outils informatiques et de méthodes statistiques d'analyse et de décision
- Identifier les acteurs d'une perturbation et de ses conséquences
- Déterminer l'efficacité et les effets d'un traitement

Objectifs

- Diagnostic, détection précoce
- Pronostic de la réponse à un traitement, de survie
- Déterminer la thérapie optimale

- Mutations dans le génome (insertions, deletions)

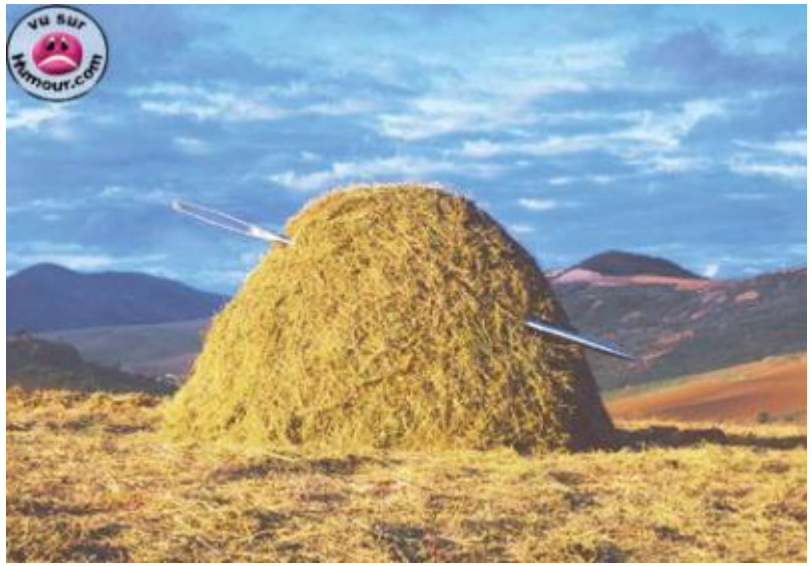
- Profil génomique, transcriptomique, protéique… des tumeurs

- Identifier les acteurs et comprendre les mécanismes

Objectifs

- Objectif : identifier les variables (gènes…) ou une combinaison liées à une caractéristique des individus ou un dysfonctionnement

- Grande échelle
  = variables innombrables
- Biologie/Médecine
  = peu d'échantillons
  car précieux, rares,
  longs à préparer
- Diagnostic/Pronostic
  = besoin de statistiques

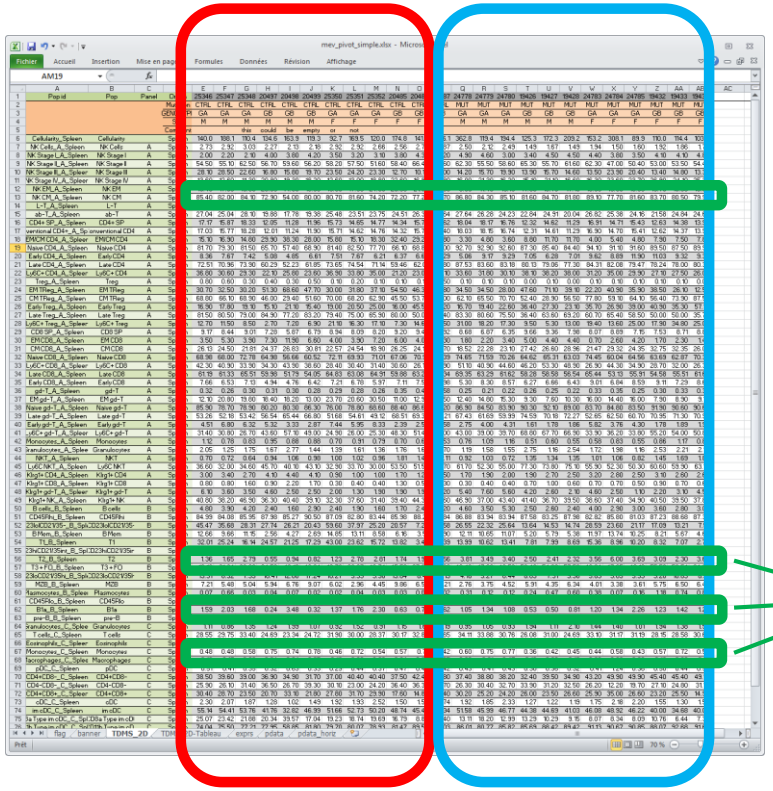


Grandes dimensions

Biologie/Médecine
= peu d'échantillons
car précieux, rares,
longs à préparer

colonne = individu = échantillon
ligne = variable = gène, protéine…

Omics = Grande échelle
= variables innombrables



? Uni-variée

? Multi-variée

Bioinfo/Biostat : où ?

- Extraction de l'information des instruments
  Prétraitement de la mesure
  - Spécifique à l'instrument, à sa technologie

- Analyse numérique pour la sélection et la décision

- Aide à l'interprétation
  - Apport de la connaissance acquise

Analysis process

This is what you learn in school & textbooks

```
# (ideal) data analysis process
raw_data = GET(data)
proc_data = PROCESS(raw_data)
SUMMARY(proc_data)
PLOT(proc_data)
model = FIT_MODEL(proc_data)
prediction = PREDICT(model)
PRINT(prediction)
> "Woo-hoo! validated model =)"
```

This is what you learn in the real world

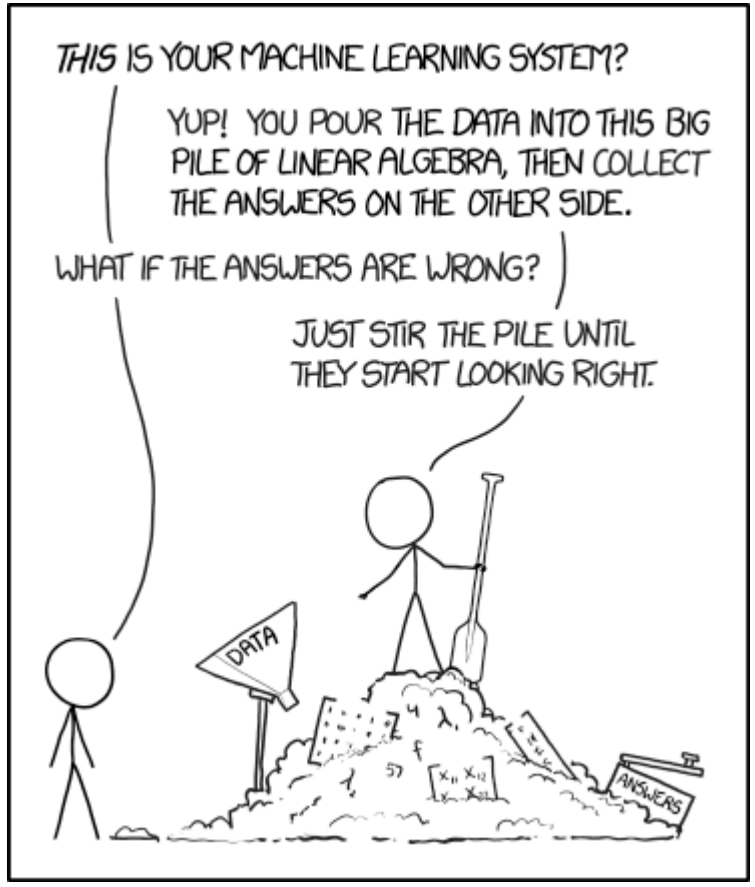
```
# (real) data analysis process
raw_data = GET(data)
clean_data = CLEAN(data)
proc_data = PROCESS(clean_data)
while (QUALITY(proc_data) != "good") {
   clean_data = CLEAN(proc_data)
   proc_data = PROCESS(clean_data)
   # while loop may run indefinitely
}
SUMMARY(proc_data)
PLOT(proc_data)
model = FIT_MODEL(proc_data)
prediction = PREDICT(model)
PRINT(prediction)
> "Ooops! model sucks =("
```

https://www.gastonsanchez.com/Data-analysis-process/

Garbage In = Garbage Out

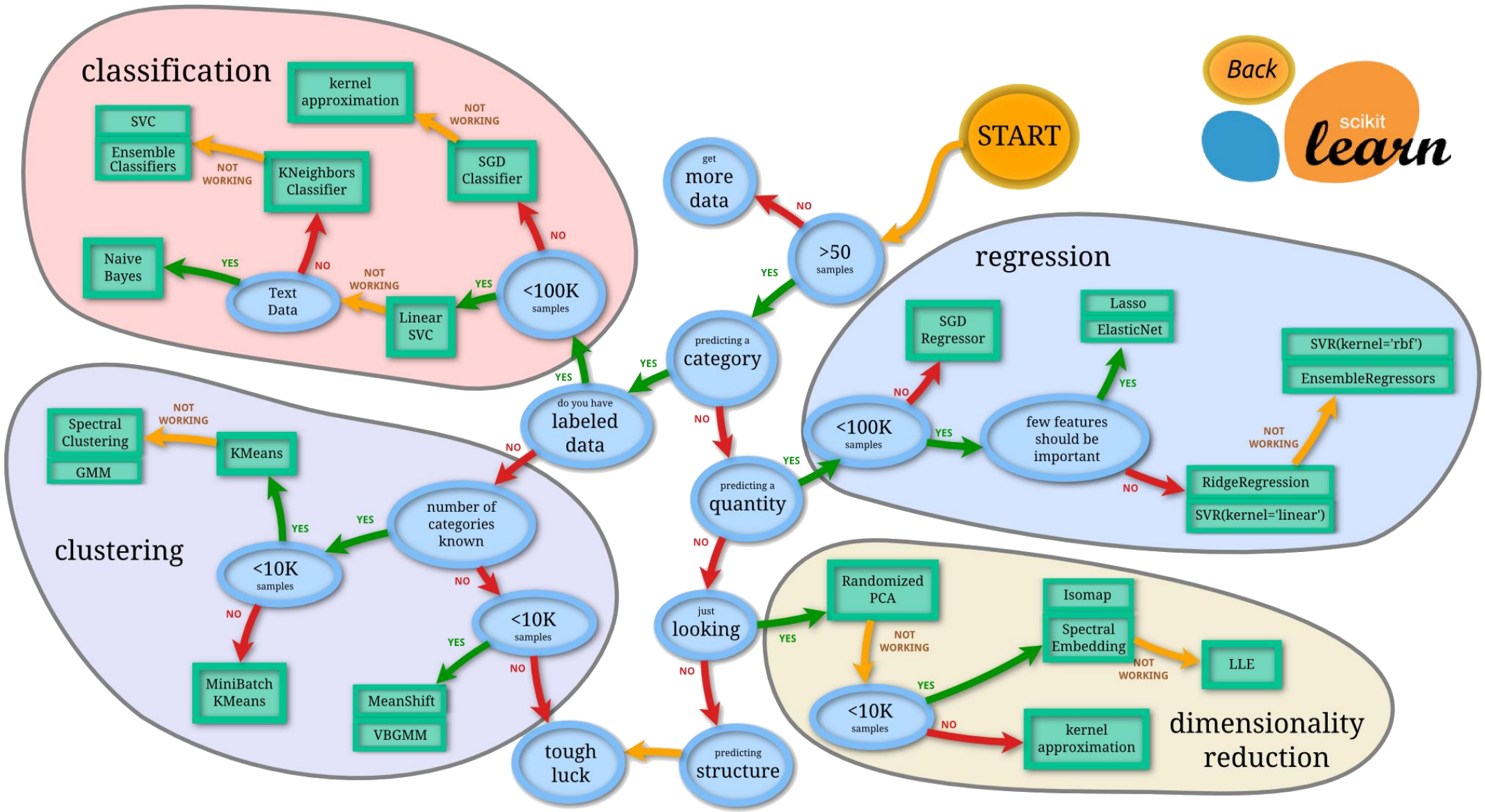
- Quality control is crucial!
- Important for manual analysis, but even more for automated analysis

"No Data Analysis Technique Can Make Good Data out of Bad Data" Howard Shapiro.



https://xkcd.com/1838/

Méthodes d'analyse



Analyse de données

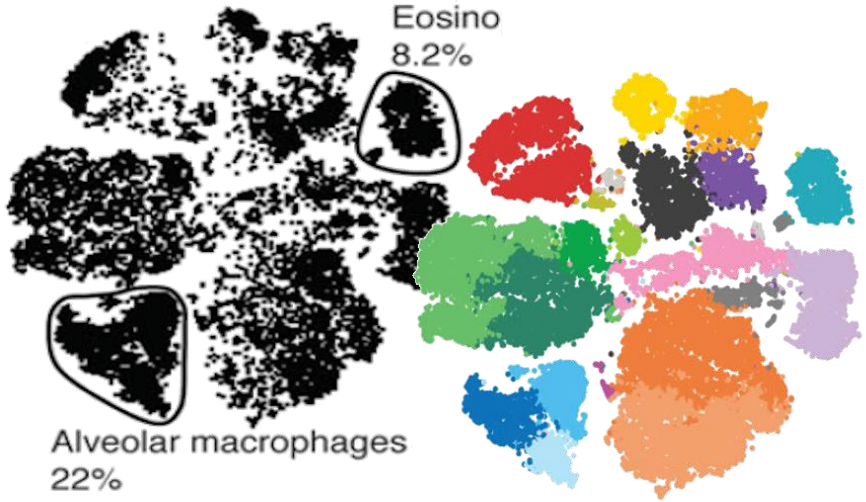
- ***Méthodes Supervisées***
  - Découverte de classes
  - Tests statistiques "classiques"
  - Utilisation des expressions corrélées
  - Score discriminant
  - Réseaux de neurones

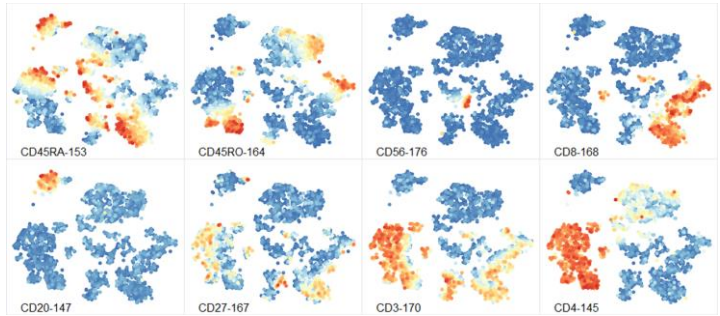
  - Valider les modèles…

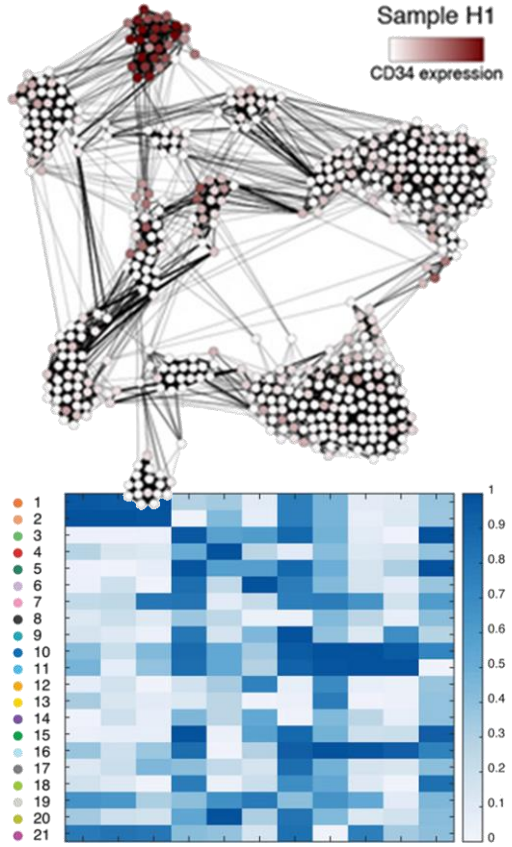
- ***Méthodes Non Supervisées***
  - Exploration des données
  - Grouper
    - Classifications hiérarchiques
    - Nuées dynamiques (K-means)
  - Réduire les dimensions
    - Analyse en composantes principales
    - tSNE, UMAP
  - Grouper et Réduire
    - Cartes de Kohonen (SOM)

VISUALISATION DE DONNÉES

1 Picture vs 1000 Words







Most illustrations from Mair et al 2016 and Kimball et al 2017

Visualisation

- Always make intelligible figures
- Put axes, label the axes, set units
- Set a legend of symbols, colors, line types…
- Define a title

Visualisation, the basics

- Screen
- X, Y coord.
- Color
- Shape
- Size

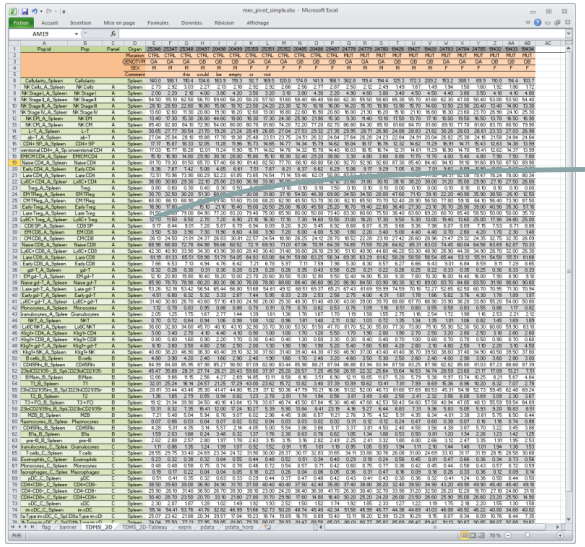
- Link the points => graph, tree

What is a dot plot?

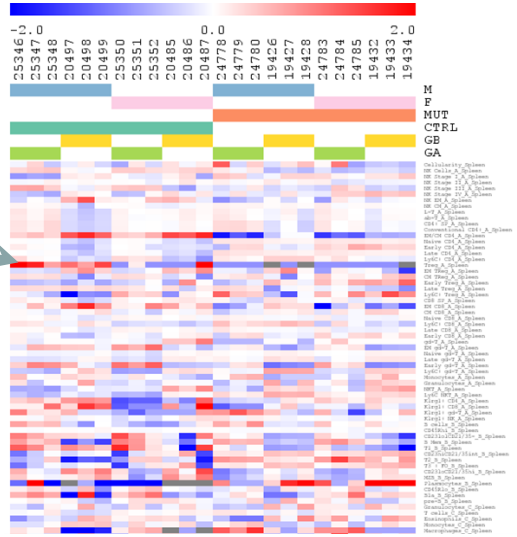
- What are those points on the scatter plot, graph?
  - Single cells or centers of cell group
- What is the meaning of the color
  - Group membership or expression level of a marker
  - Raw or transformed expression level
- Is the distance on the screen meaningful?
- How to evaluate the distance between dots?

From matrix to image

Heatmap



Scale
Convert



What is on a heatmap?

- What are those rectangles?
  - Columns = ? Rows = ?
  - Cell groups x markers or Cell groups x patients
- What is the meaning of the color?
  - Expression level of a marker or percentages
  - Raw or transformed value
  - What is the scaling?
- Criterion for arranging the rows/columns?
  - User defined or hierarchical clustering
  - How to evaluate the distance?

RAPPELS STATISTIQUES

Démarche scientifique

- Poser une question fondée sur des hypothèses
  - science, expertise
- Définir une expérience bien pensée/contrôlée
  - plan d'expérience
- Réaliser l'expérience, annoter son déroulement
  - valeurs aberrantes, manquantes
- Organiser les mesures, les décrire
  - statistique descriptive
- Comparer les mesures inter-groupes
  - statistique inférentielle, test d'hypothèse
- Interpréter les résultats
  - science, expertise

To consult the statistician after an experiment
is finished is often merely to ask him
to conduct a post mortem examination.
He can perhaps say what the experiment died of.
Ronald Fisher 1938

La statistique, c'est PAS compliqué !

- Estimer
  - Modéliser
- Décider
  - Risquer

Populations et échantillons

- En biologie, on souhaite caractériser ou comparer des populations, autrement dit l'ensemble des sujets d'un groupe constitué d'un nombre élevé, parfois infini, de sujets.
- Sur le plan pratique, le biologiste ne peut étudier qu'un échantillon, autrement dit un nombre réduit d'individus de cette population.
- L'expérience mène à une **estimation** de la grandeur d'intérêt.
- L'objectif est de **généraliser à la population le résultat déterminé sur un échantillon**.

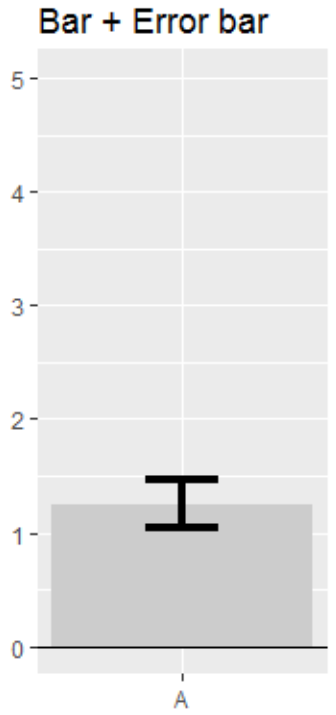
Populations et échantillons

- L’analyse statistique va permettre :
  - de **décrire** d’abord une ou plusieurs populations, connues chacune seulement à partir de l’échantillon étudié,
  - de **comparer** ensuite, et surtout, ces populations entre elles ou avec un modèle représentant une population théorique.

- **Attention** : l’outil statistique ne peut être utilisé que si chaque échantillon est extrait de la population étudiée par une méthode **aléatoire** : chaque individu de la population a la même probabilité d’être choisi.

Quizz

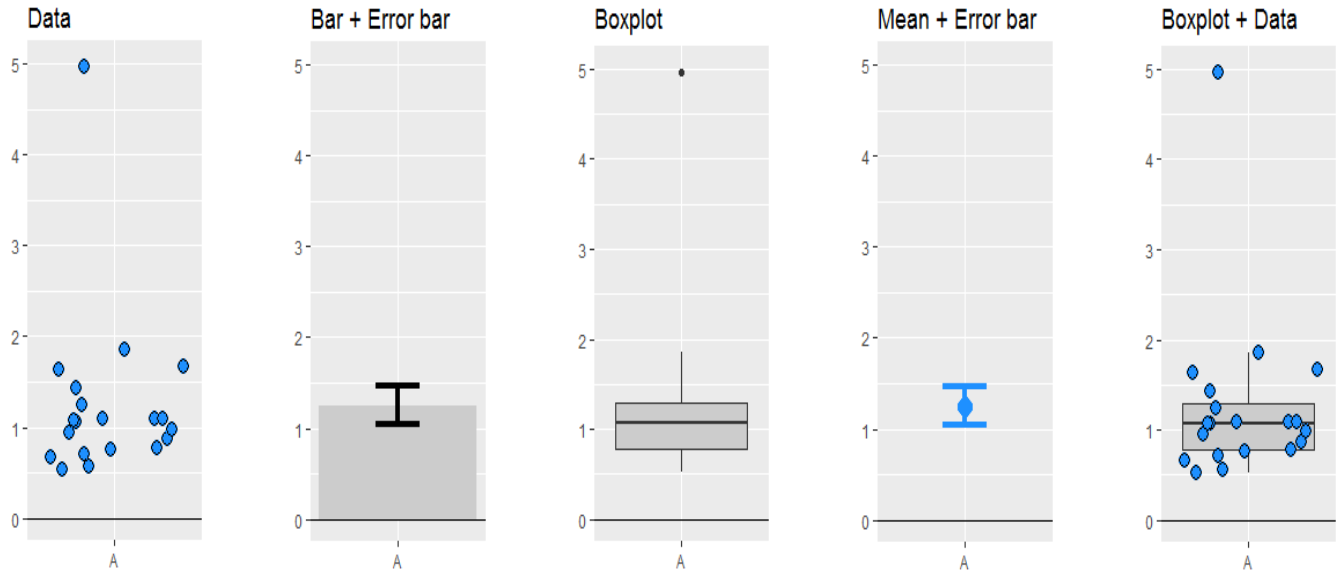
- Que représente la barre d'erreur sur vos graphiques ?
- Quel est le nom de la grandeur représentée ?



Error bars

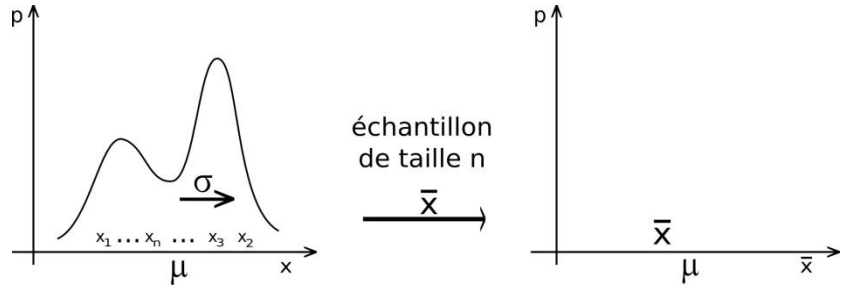
Représenter les mesures

- barplot vs dotplot vs boxplot
- un graphique permet d'apprécier les valeurs extrêmes



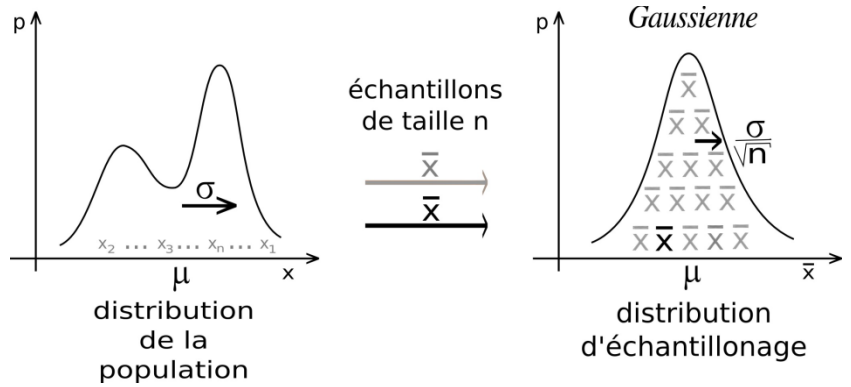
Théorème centrale limite

- Intuitivement, ce résultat affirme que toute somme de variables aléatoires indépendantes et identiquement distribuées tend vers une variable aléatoire gaussienne.
- Quelque soit la forme de la distribution de la population, la distribution de la moyenne des échantillons d'effectif n tend vers une gaussienne, dont la moyenne est la moyenne de la population.
- Simplifier les calculs avant l'ordinateur.



1 expérience

1 échantillon n>>1



multitude d'expériences

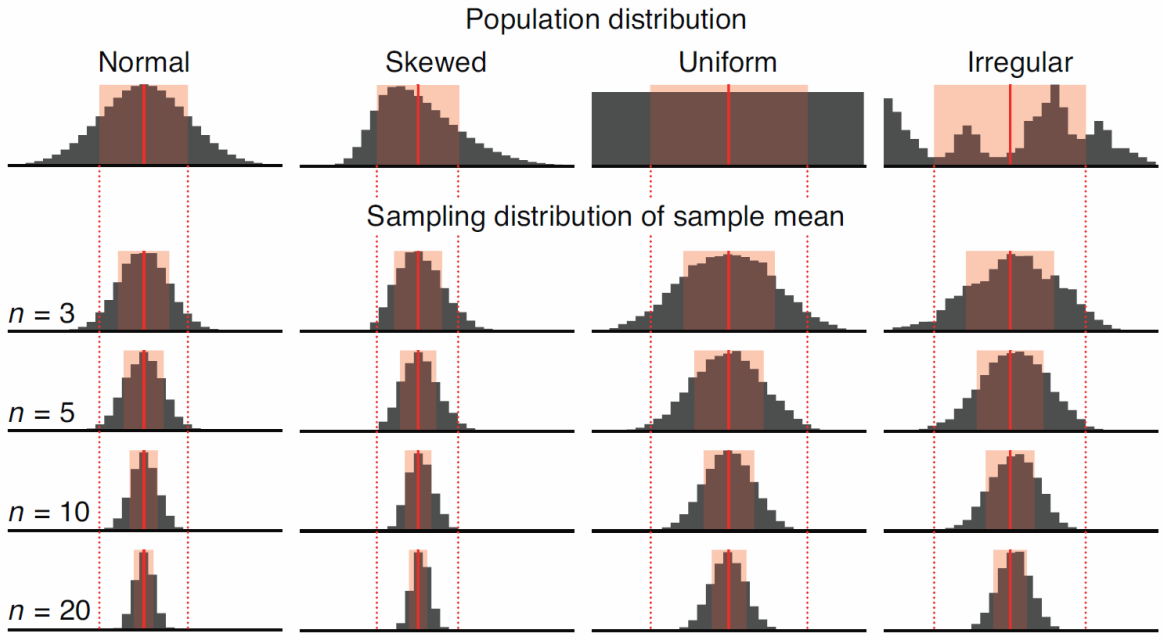
plein d'échantillons n>>1

http://www.incertitudes.fr/livre.pdf

Pierre-Simon de Laplace 1809, De Moivre 1733

Tendance centrale

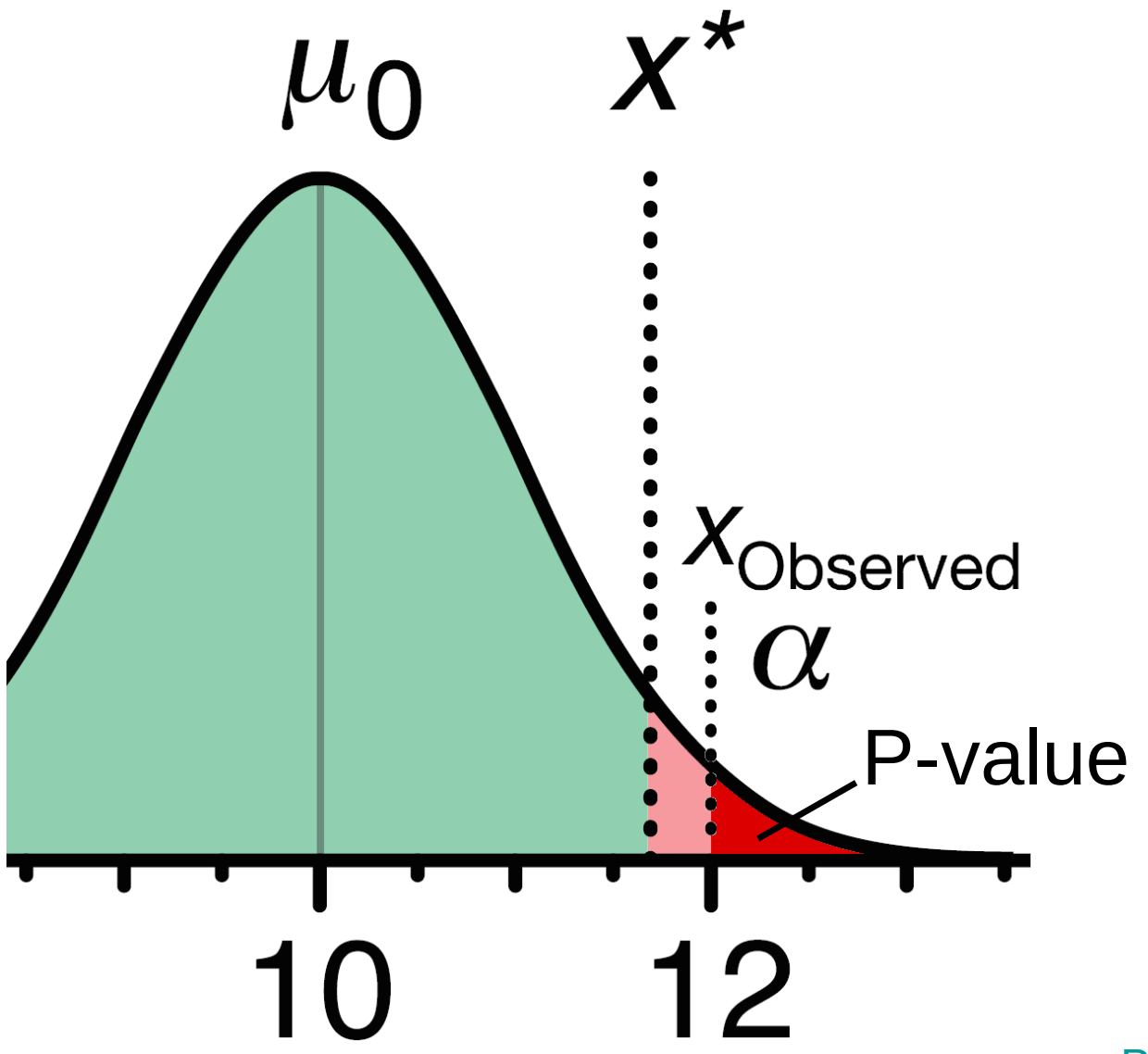
Evolution de la distribution de la moyenne d'un échantillon
en fonction de la taille de l'échantillon



La distribution de la moyenne d'un échantillon de n individus
tend vers une loi "normale" pour toutes les distributions,
c'est le théorème centrale limite

Importance of being uncertain

Principe du test d'hypothèse



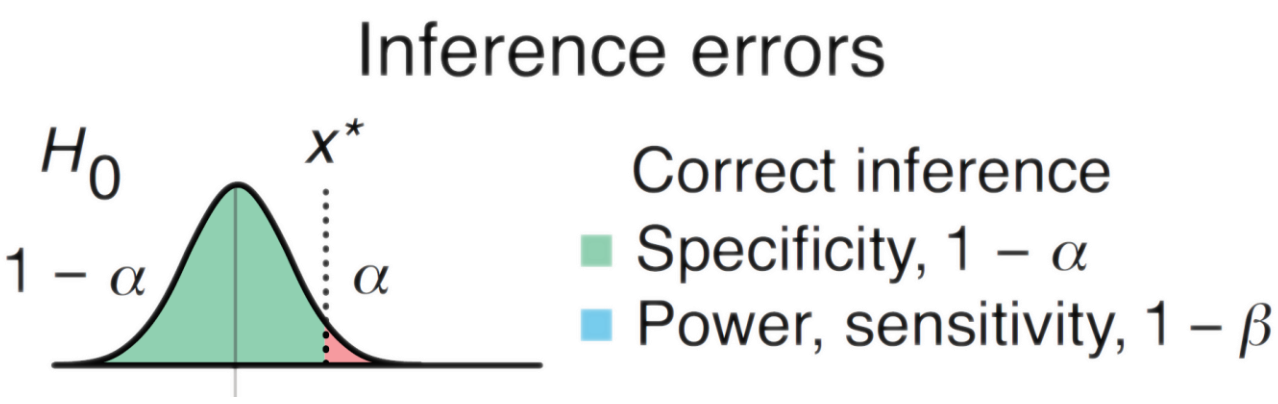
- H_0 : hypothèse nulle : $\mu_0 = 10$
- Distribution de la **moyenne** d'un échantillon de n individus

- μ_{obs} moyenne observé sur l'échantillon de n individus expérimentés ; $\mu_{obs} = 12$

- P-value = Prob($\mu_{obs} > 12 \mid H_0$)
- P-value ~ 2% < 5%
  => on rejette H_0

Points of significance: Power and sample size

P-value = p(False Positive)



		DECISION	
		Reject H_o	*Fail to Reject H_o*
UAL	*H_o True*	Type I Error *Producer Risk* α-Risk False Positive	**Correct Decision** Confidence Interval = 1- α

[Alpha and Beta Risks]

Test d'hypothèse

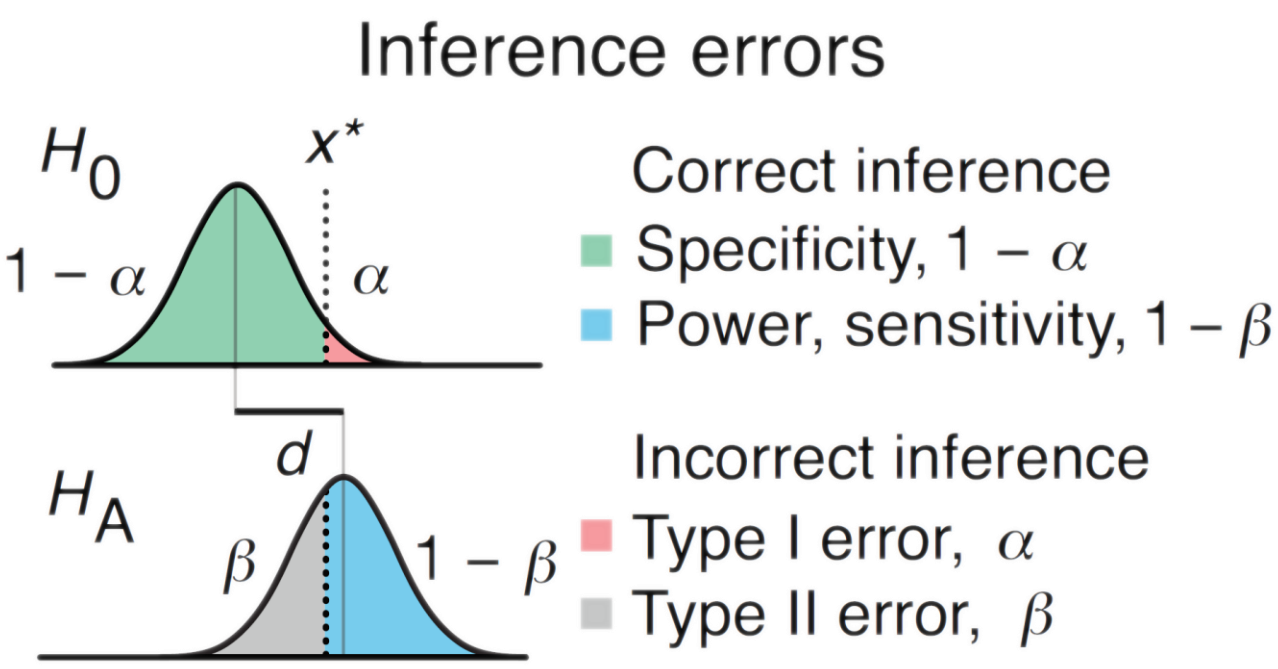
- Rejeter (ou non) l'absence de différence
- P-value = probabilité de trouver une valeur plus extrême si l'hypothèse nulle est vraie
- P-value ~ risque de croire en une différence alors qu'il n'y en a pas
- P-value ~ risque de croire en une différence alors qu'il s'agit d'une observation liée à un échantillonnage "exceptionnel"

NHST mis-conceptions

- P-value is the probability that the null hypothesis is false. NO!
- A low P-value value indicates a large effect. NO!
  => a low P-value can occur with small effect sizes, particularly if the sample size is large.
- A non-significant outcome means that the null hypothesis is probably true. NO!
  => the data do not conclusively demonstrate that the null hypothesis is false.

A Dirty Dozen: Twelve P-Value Misconceptions - Steven Goodman

Power = p(True Positive)



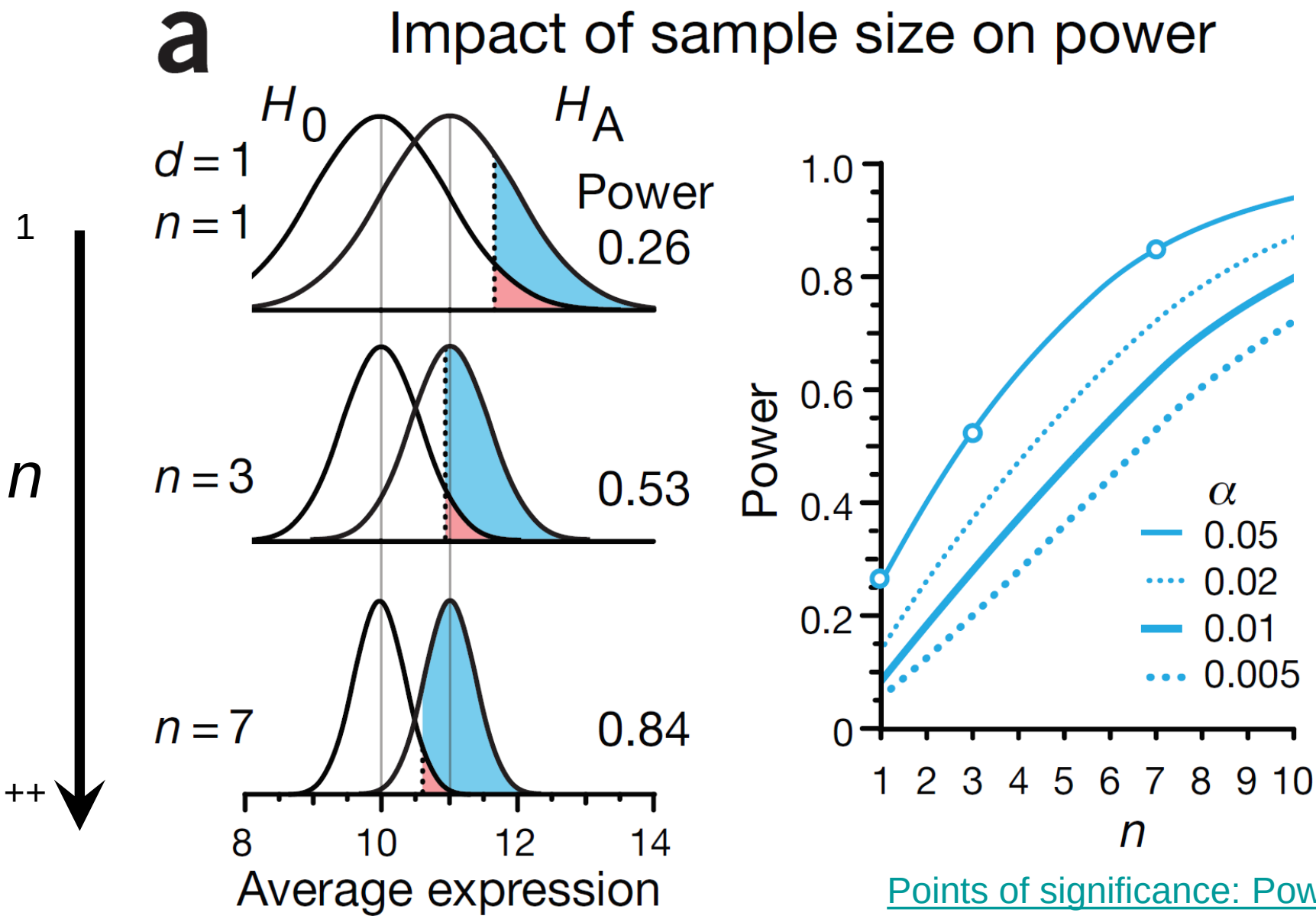
Sampling distribution of mean

		DECISION	
		Reject H_o	*Fail to Reject H_o*
ACTUAL	*H_o True*	Type I Error *Producer Risk* α-Risk False Positive	**Correct Decision** Confidence Interval = 1- α
	H_a True	**Correct Decision** **Power = 1- β**	Type II Error *Consumer Risk* β-Risk False Negative

H_o: Null Hypothesis H_a: Alternative Hypothesis

[Nature - Points of Significance - Decision and Risks]

N, d, and Power

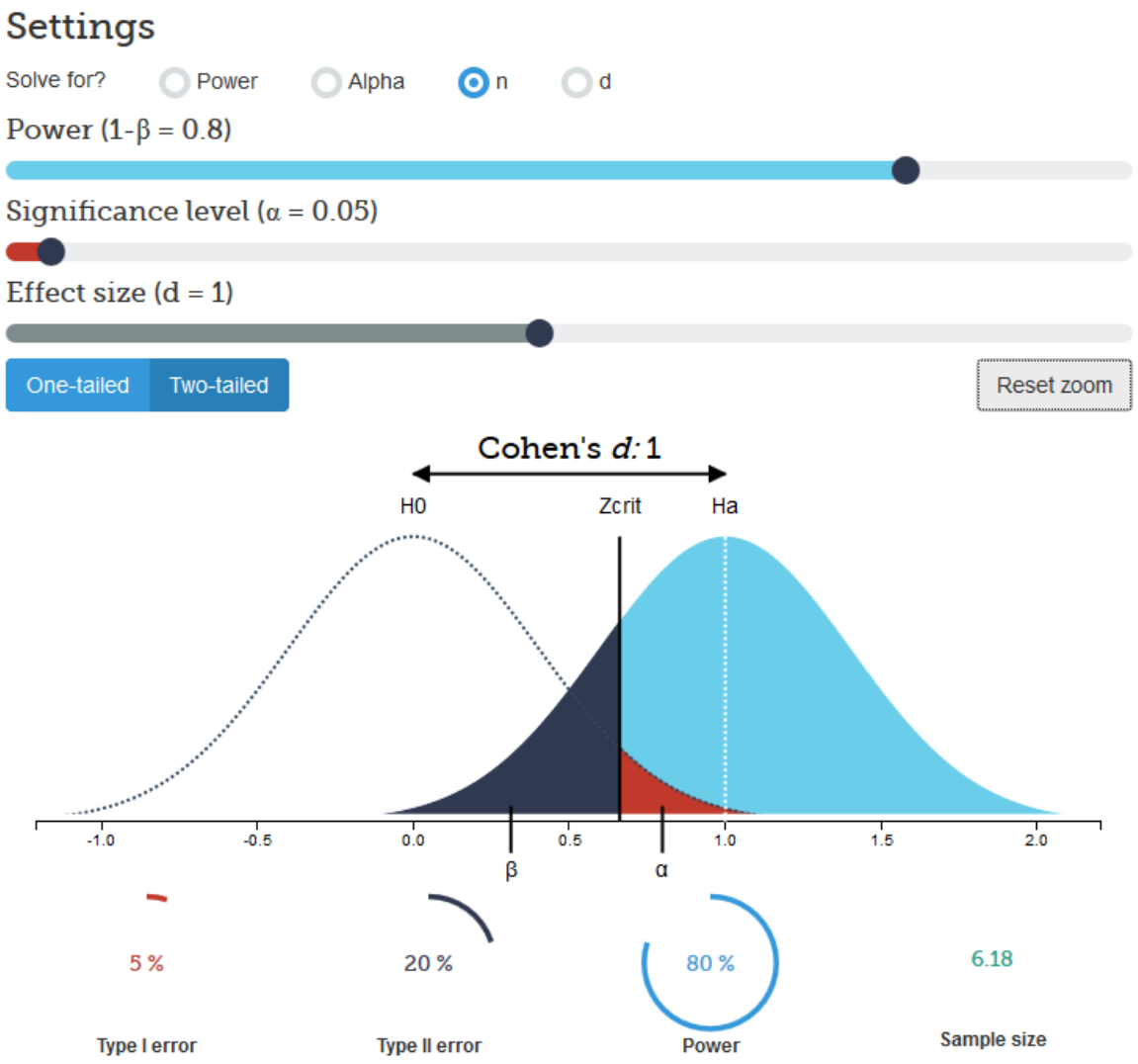


n is acting as the resolution of a microscope: H0 and HA become easier to separate, because the uncertainty decreases, but the difference does not change.

H_0 and H_A are assumed normal with $\sigma = 1$.
(a) Increasing n decreases the spread of the distribution of sample averages in proportion to $1/\sqrt{n}$. Shown are scenarios at $n = 1$, 3 and 7 for $d = 1$ and $\alpha = 0.05$. Right, power as function of n at four different α values for $d = 1$. The circles correspond to the three scenarios.

Points of significance: Power and sample size

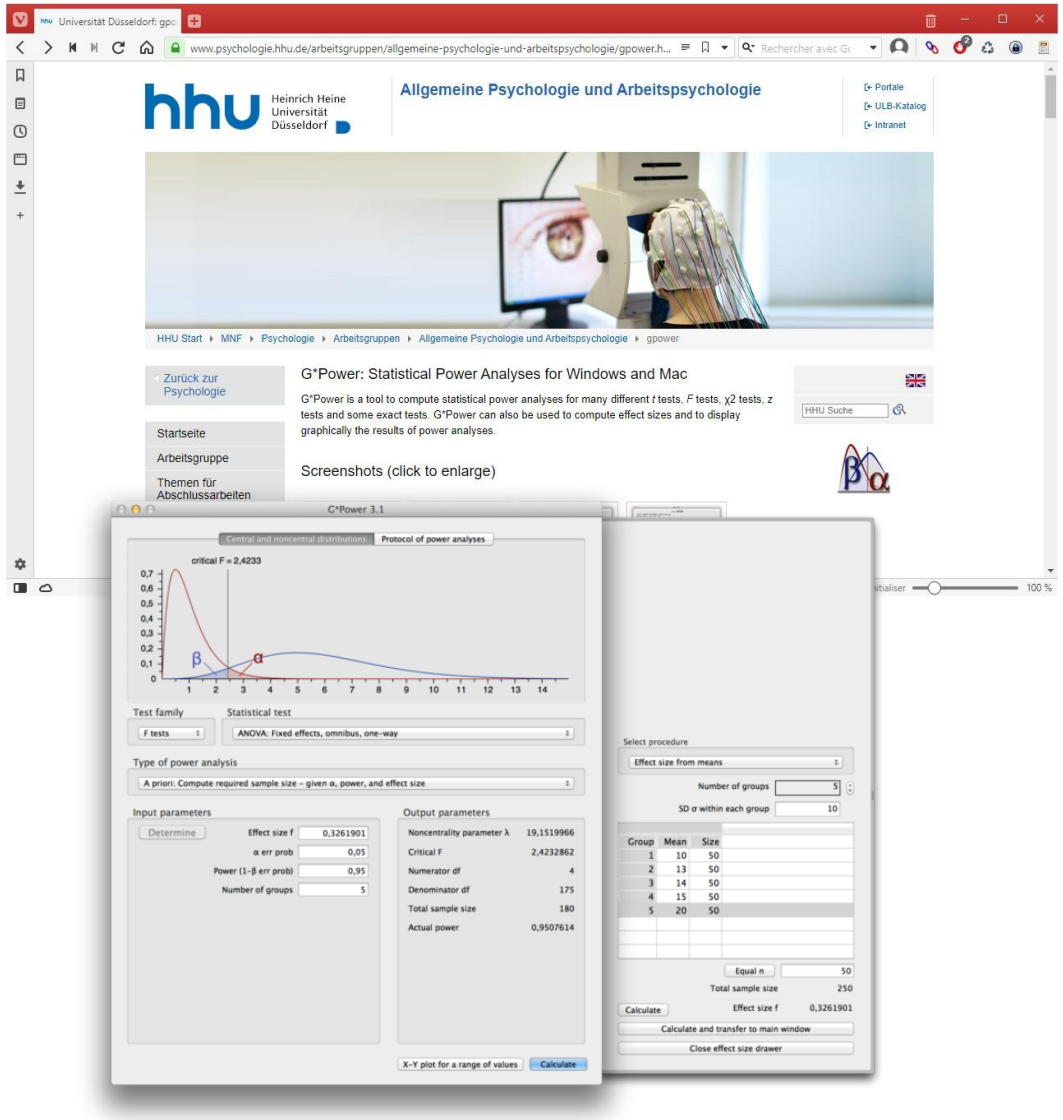
Démo interactive



Understanding Statistical Power and Significance Testing
http://rpsychologist.com/d3/NHST/

Calcul de puissance

- gpower.hhu.de
- test statistique appliqué
- taille de l'effet
  - relatif => Cohen
- étude pilote ou effet minimal



Grille de tests

Conditions d'application	variable gaussienne ou N > 30 ANOVA : variances homogènes	variable non gaussienne de préférence N >= 5
Échantillons	**test paramétrique**	**test non paramétrique**
Un seul échantillon	Test t de Student à une norme	Test Wilcoxon à une norme
Deux échantillons indépendants	Test t de Student	Test de permutations Test de Mann - Whitney Test de Kruskal - Wallis
Plus de deux échantillons indépendants	ANOVA 1 facteur ANOVA 2 facteurs	Test de permutations Test de Kruskal - Wallis
Deux Séries appariées	Test t de Student apparié	Test du signe Test de Wilcoxon
Plus de deux séries appariées	ANOVA à mesures répétées	Test de Friedman
Mesure de l'association entre variables	Coefficient de corrélation de Pearson	Coefficient de corrélation par rangs de Spearmann Coefficient de concordance de Kendall
Test de normalité	Test Shapiro - Wilk	Non Applicable

Guide de choix de test statistique XLSTAT

MicrOsiris BiostaTGV

Critères de choix

- Loi normale ou Inconnue ⇔Paramétrique ou Non Param.
  - hypothèses à vérifier a priori ou a posteriori
  - normalité, homogénéité des variances
- 2 échantillons ou Plus
  - si 2, 2-tail ou 1-tail
  - si Plus, recherche 2 à 2 ou à une référence, cf ANOVA
- Individus Appariés ou Indépendants
- Echantillon vs Référence
  - cas d'une référence en neuro ou de la normalisation au contrôle
- Analyse de survie

one-side vs two-side

- Si on connait a priori (avant l'expérience) le sens du changement, on testera que ce sens (one-side).
- Si on n'a pas d'a priori sur le changement, on testera les deux côtés (two-sides).
- Idem lorsqu'un piéton traverse une voie, sens unique ou double sens.
- Le risque de 5% sera donc que d'un côté, ou réparti à part égale de chaque côté (2.5% et 2.5%). Le seuil d'acceptation (X^*) sera donc plus proche de μ_0 pour le test d'un seul côté.

Appariement

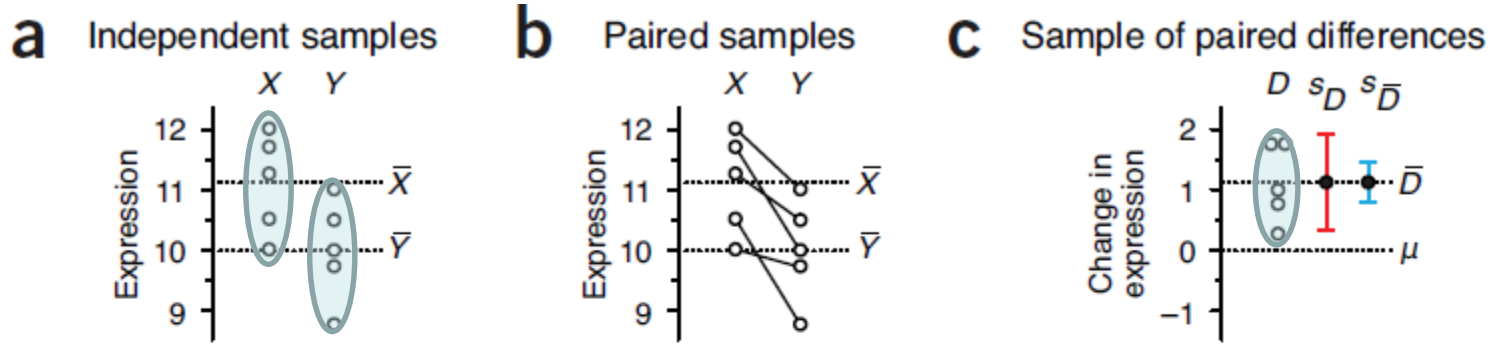


Figure 3 | The paired *t*-test is appropriate for matched-sample experiments. (**a**) When samples are independent, within-sample variability makes differences between sample means difficult to discern, and we cannot say that X and Y are different at $\alpha = 0.05$. (**b**) If X and Y represent paired measurements, such as before and after treatment, differences between value pairs can be tested, thereby removing within-sample variability from consideration. (**c**) In a paired test, differences between values are used to construct a new sample, to which the one-sample test is applied ($\bar{D} = 1.1$, $s_D = 0.65$).

Comparing samples—part I

Appariement

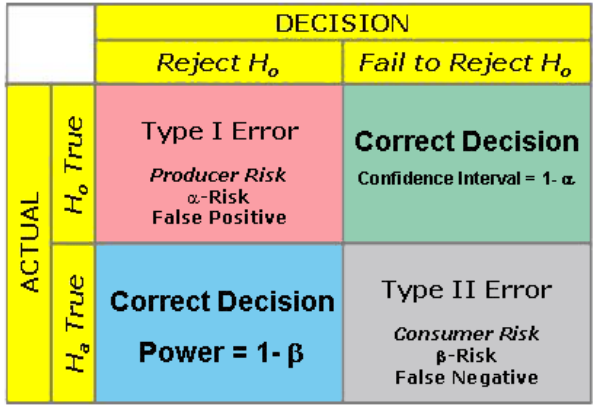
- gommer l'effet "individu"
- série temporelle

- ANOVA à mesures répétées
- Modèles mixtes
  - si valeurs manquantes

La statistique, c'est PAS compliqué !

- Estimer
  - Modéliser, expliquer
  - Vérifier l'adéquation du modèle aux mesures

- Décider
  - Risquer
  - α : déclarer une différence à tort ⇔ p-value
  - β : manquer une différence à tort ⇔ puissance



		DECISION	
		Reject H_o	*Fail to Reject H_o*
ACTUAL	*H_o True*	Type I Error *Producer Risk* α-Risk False Positive	**Correct Decision** Confidence Interval = 1- α
	H_a True	**Correct Decision** **Power = 1- β**	Type II Error *Consumer Risk* β-Risk False Negative

H_o: Null Hypothesis H_a: Alternative Hypothesis

- La complexité vient de la multitude des apports au cours de développement de cette science

La statistique, c'est PAS compliqué !

- Test d'hypothèse
  - assigner un modèle aux données
    - design (2, 3 ou 4 groupes, régression linéaire), nature du bruit, effets contrôlés
  - vérifier les conditions d'application a priori ou
    a posteriori l'adéquation du modèle aux mesures (résidus)
  - calculer la valeur d'un indice statistique adéquat
  - placer cette valeur dans la distribution de cet indice sous l'hypothèse nulle
  - mesurer la proportion d'indices plus grand : p-value

La statistique, c'est PAS compliqué !

- P-value
  - garde-fou qui écarte l'aléa
  - autorise l'interprétation
- Interpréter, c'est conclure scientifiquement sur
  - l'importance de l'effet,
  - pas sur l'importance de la p-value

To consult the statistician after an experiment
is finished is often merely to ask him
to conduct a post mortem examination.
He can perhaps say what the experiment died of.
Ronald Fisher 1938

Approfondir

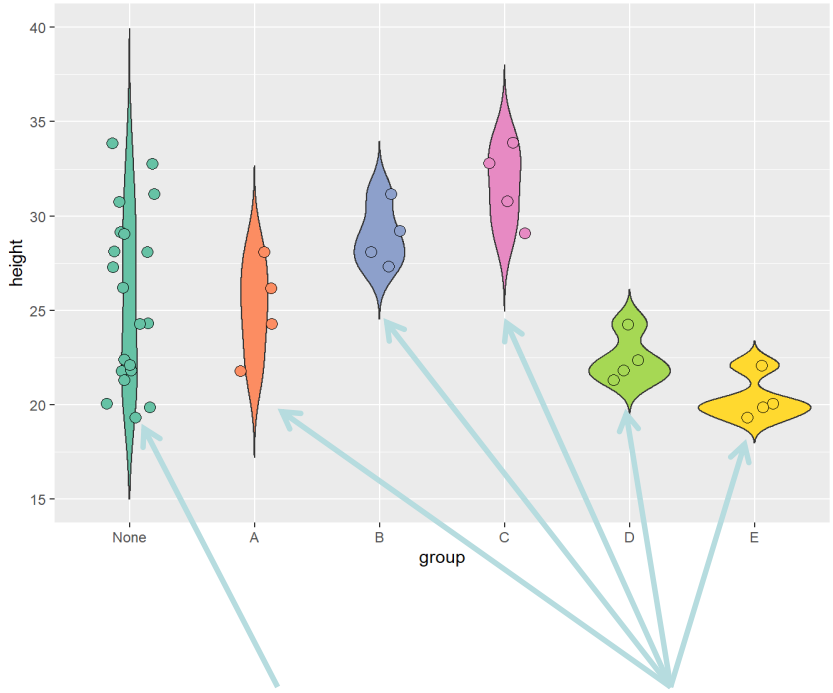
- Créer des projets synchronisés, partagés et versionnés
  - git thinkR
- Utiliser l'incontournable (?) tidyverse
  - thinkR
- Se perfectionner en Rmarkdown => Quarto
  - astuces thinkR
- Informer l'utilisateur
  - cli thinkR

MULTI-GROUPES
1 FACTEUR

ANOVA 1-way

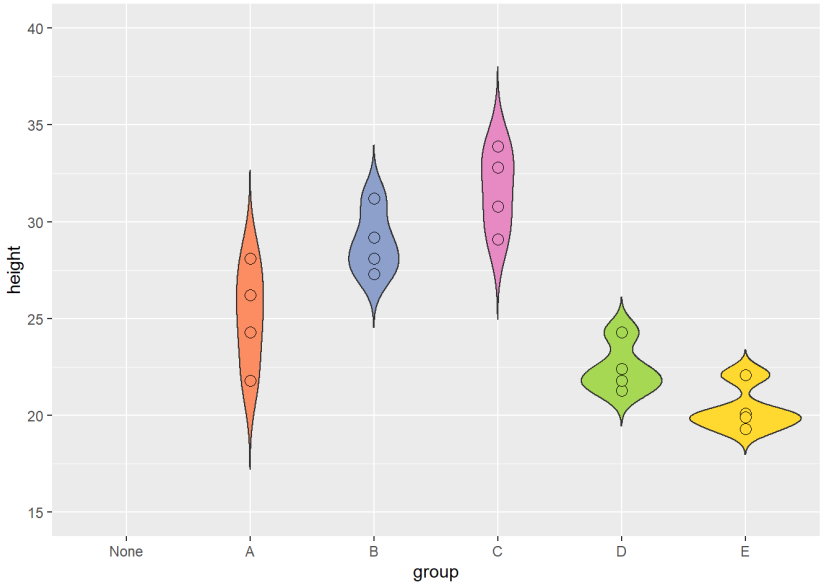
- multi-group comparison
- overall comparison
- H_0 : all means are equal (same population)
- all groups share the same dispersion: within variability is pooled

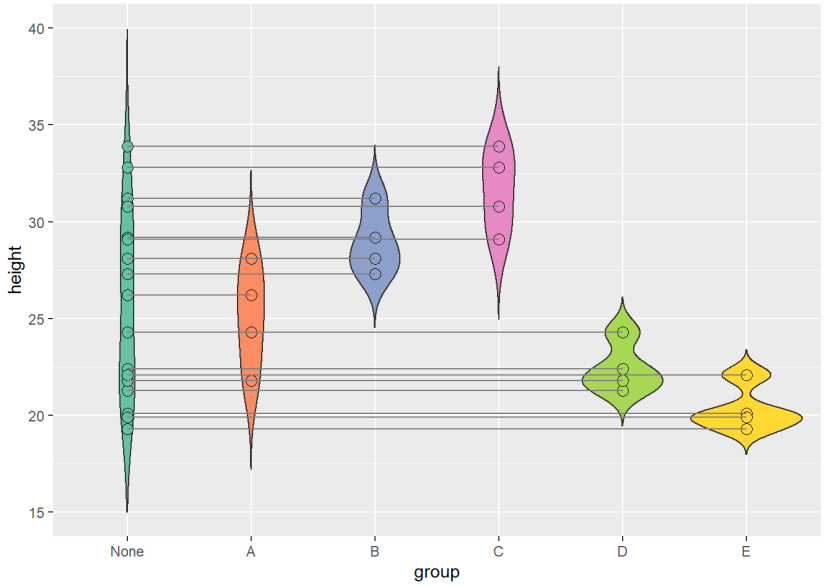
$$F = \frac{Variance\ between\ groups}{Variance\ within\ groups} \sim F(\nu_1, \nu_2)$$

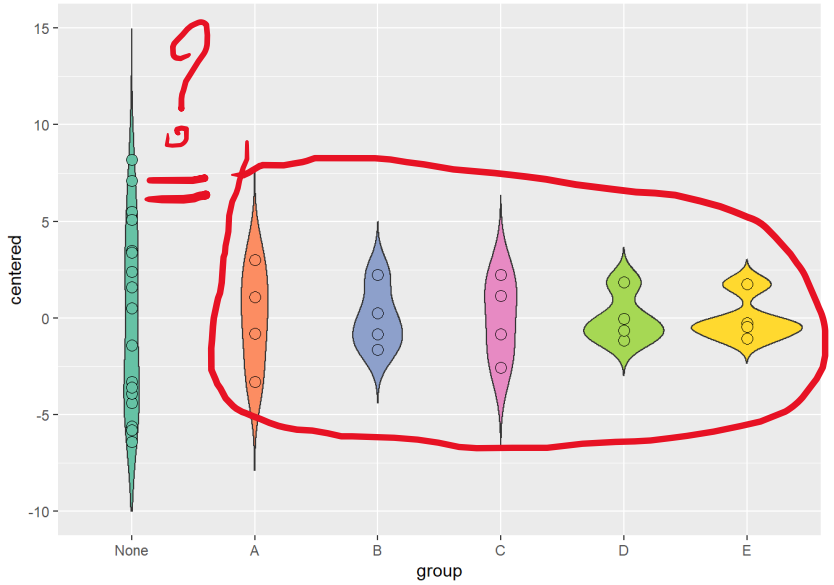


$$F = \frac{Total\ variance - Within\ variance}{Within\ variance}$$

ANOVA 1-way


40
35
30
25
20
15
height
None
A
B
C
D
E
group


40
35
30
25
20
15
height
None
A
B
C
D
E
group


15
10
5
0
-5
-10
centered
None
A
B
C
D
E
group

ANOVA 1-way

- **Conditions**
- normalité intra-groupe
- homogénéité des variances
  - test de Levene (ou Barlett)
- vérifier les hypothèses a posteriori
  - résidus standardisés
- effets additifs

- **Avantages**
- tous les individus contribuent à l'estimation de la variance

- **Inconvénients**
- où sont les différences ?
  - t-tests en série ?

http://jebrane.perso.math.cnrs.fr/m1psy/Tests_%20egalite_%20variances.pdf

ANOVA 1-way

1) normalité

```
##  Shapiro-Wilk normality test
##
## data:  aov.ex1$residuals
## W = 0.96561, p-value = 0.6609
```

ok !

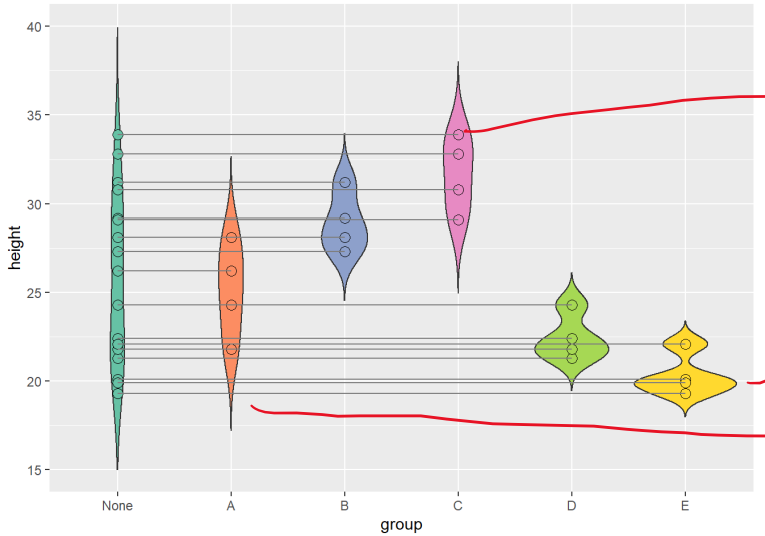
A priori

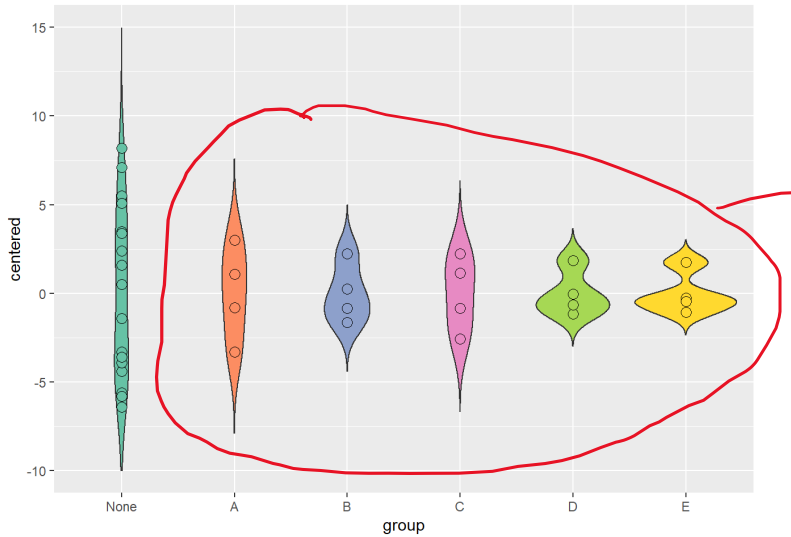
2) homogénéité des variances

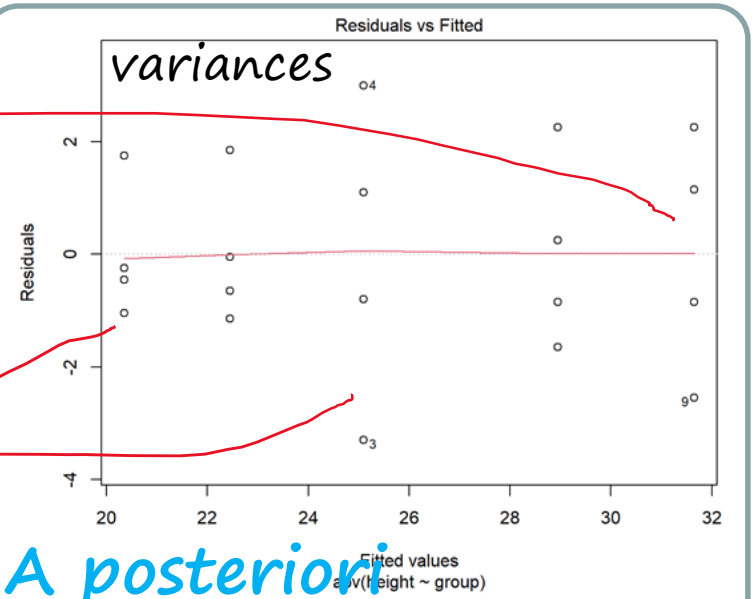
```
## Levene's Test for Homogeneity of Variance
##  (center = median)
##       Df F value Pr(>F)
## group  4  1.1885 0.3558
##       15
```

ok !

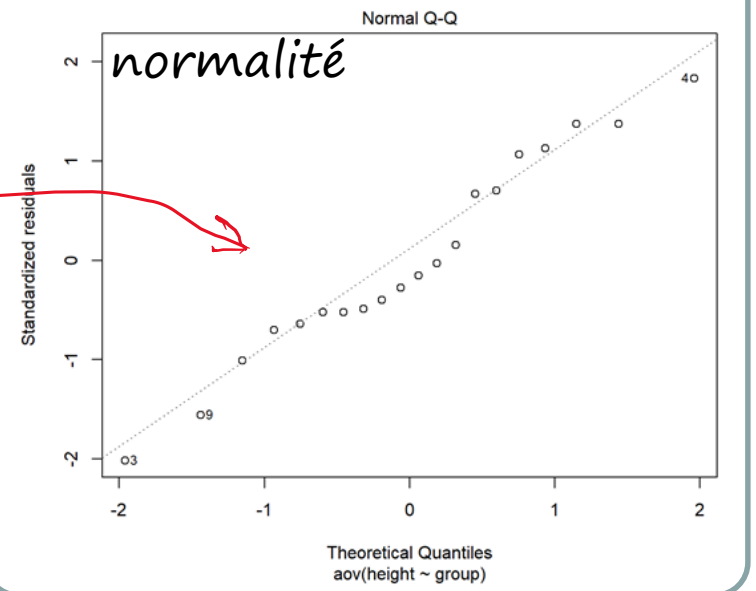
```
## Bartlett test of homogeneity of variances
## data:  height by group
## Bartlett's K-squared = 2.3389,
## df = 4, p-value = 0.6737
```







A posteriori



Méthodologie

- visualiser les données, décrire les groupes, identifier des valeurs extrêmes
- calculer le modèle
- vérifier les hypothèses =>>
- variance factorielle est significativement supérieure à la variance résiduelle ?
- identifier les groupes par des tests post-hoc

- vérifier les hypothèses par des tests numériques et des représentations graphiques

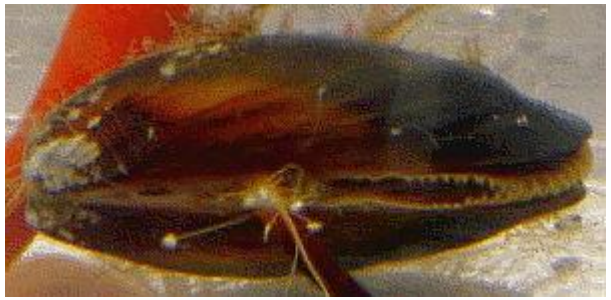
1) Indépendance des résidus
   - non corrélés entre eux
   - non corrélés au facteur étudié
2) Normalité des résidus
3) Homogénéité des variances
   - si non, correction de Welch

- alternative non paramétrique : Kruskal-Wallis

ANOVA 1-way

Dans une à un facteur, il y a une variable de mesure et une variable nominale. On effectue plusieurs observations de la variable de mesure pour chaque valeur de la variable nominale.

Voici quelques données sur une mesure de la coquille (ici la "longueur AAM") chez la moule Mytilus trossulus provenant de cinq endroits : Tillamook, Oregon ; Newport, Oregon ; Petersburg, Alaska ; Magadan, Russie ; Tvarminne, Finlande (McDonald et al., 1991).



fichier sur AMUBox

Tillamook	Newport	Petersburg	Magadan	Tvarminne
0.0571	0.0873	0.0974	0.1033	0.0703
0.0813	0.0662	0.1352	0.0915	0.1026
0.0831	0.0672	0.0817	0.0781	0.0956
0.0976	0.0819	0.1016	0.0685	0.0973
0.0817	0.0749	0.0968	0.0677	0.1039
0.0859	0.0649	0.1064	0.0697	0.1045
0.0735	0.0835	0.105	0.0764	
0.0659	0.0725		0.0689	
0.0923				
0.0836				

Calcul

L'idée est de calculer la moyenne des observations au sein de chaque groupe, puis de comparer la variance entre ces moyennes à la variance moyenne au sein de chaque groupe.

Sous l'hypothèse nulle que les observations des différents groupes ont toutes la même moyenne, la variance pondérée entre les groupes sera la même que la variance à l'intérieur des groupes. Plus les moyennes sont éloignées, plus la variance entre les moyennes augmente.

La statistique du test est le rapport de la variance entre les moyennes divisée par la variance moyenne à l'intérieur des groupes. La forme de la distribution F dépend de deux degrés de liberté, les degrés de liberté du numérateur (variance entre les groupes) = nombre de groupes moins un, et les degrés de liberté du dénominateur (variance à l'intérieur des groupes) = nombre total d'observations moins le nombre de groupes.

Pour l'exemple, il y a 5 groupes et 39 observations, de sorte que les degrés de liberté du numérateur sont 4 et les degrés de liberté du dénominateur sont 34.

	sum of squares	d.f.	mean square	F_s	P
among groups	0.00452	4	0.001113	7.12	2.8×10^{-4}
within groups	0.00539	34	0.000159		
total	0.00991	38			

On peut simplement écrire "Les moyennes étaient significativement hétérogènes (ANOVA à un seul facteur, F(4;34) = 7,12, P=2,8×10-4)". Les degrés de liberté sont indiqués en indice de F, avec le numérateur en premier.