Session 3:

- statistiques pour les données omiques
- RStudio et Rmarkdown

Teachers: Claire Vandiedonck, Antoine Bridier-Nahmias
Helpers: Anne Badel, Clémence Réda, Olivier Sand, Jacques van Helden

Plan de la session 3:

1. Statistiques pour les données omiques
    a. Rappels de stats de base
    b. Problèmes de la dimensionalité des données omiques

2. Rstudio et R markdown

1. Statistiques pour les données omiques

Some French-English terms

- barplot = diagramme en bâtons
- co-variate = covariable
- confidence interval (CI) = intervalle de confiance
- density probability =densité de probabilité
- likely = probable
- mean = moyenne
- pairwise = apparié
- power = puissance
- random variable = variable aléatoire
- random/sampling fluctuation = variation d'échantillonnage
- sample = échantillon
- significance = signification
- standard deviation = écart type = racine carrée de la variance
- standard error = standard deviation of the mean = écart type de la moyenne = écart-type rapporté à la racine carrée de la taille de l'échantillon
- threshold = seuil
- variance = variance = dispersion des données autour de la moyenne

Why using stats?

Making sense of data

↳ Aim: identify variables whose variation levels are associated with a phenotype or a covariate of interest (eg: response to stress, to a treatment, survival, mutation, tumor class, time…)

Variable to explain ~ explanatory variables + covariates + residual error

Problems addressed by statistics:

1. estimation: of the effects of interest and of how they vary
2. testing: = assessing the statistical significance of the observed effects

Deux difficultés dans la mise en évidence d'un effet

grande masse
de données

issues d'échantillons
et non de la population
en partie cachée

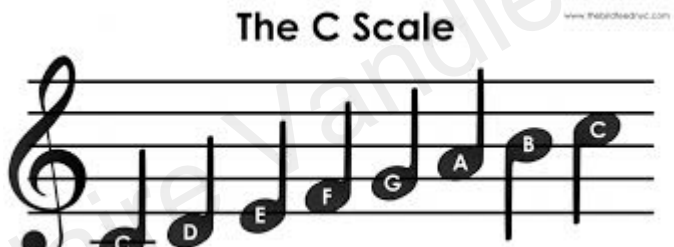
1.1. Random variable and sampling

Traits/Variables

Qualitative

- Nominal = categorical

- Ordinal = rankable



Quantitative = variable

- continuous: uncountable items

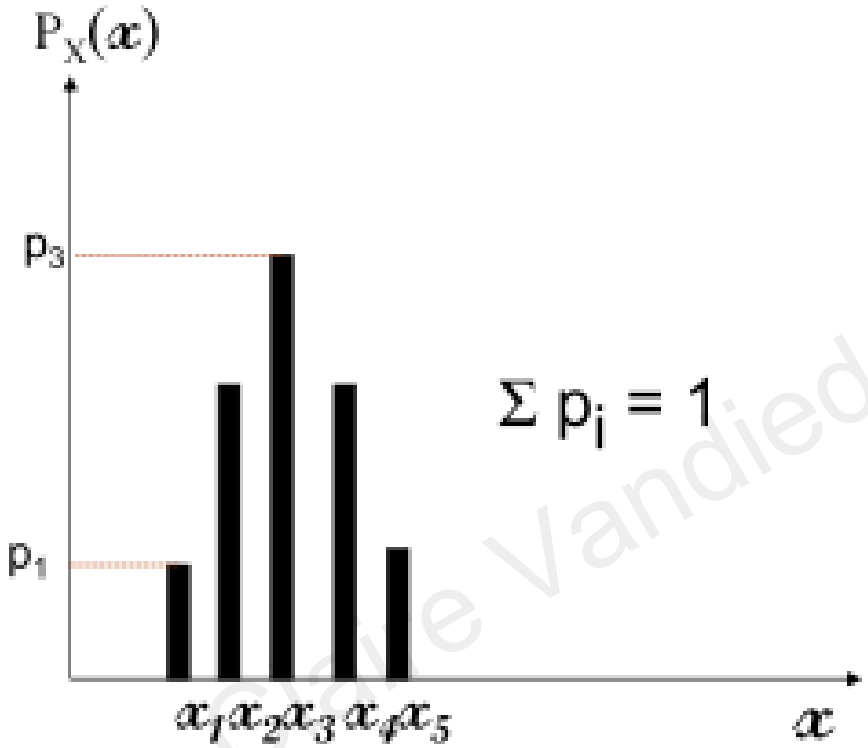
- discrete : countable items

Random variable

Probability associated to the each value of the variable

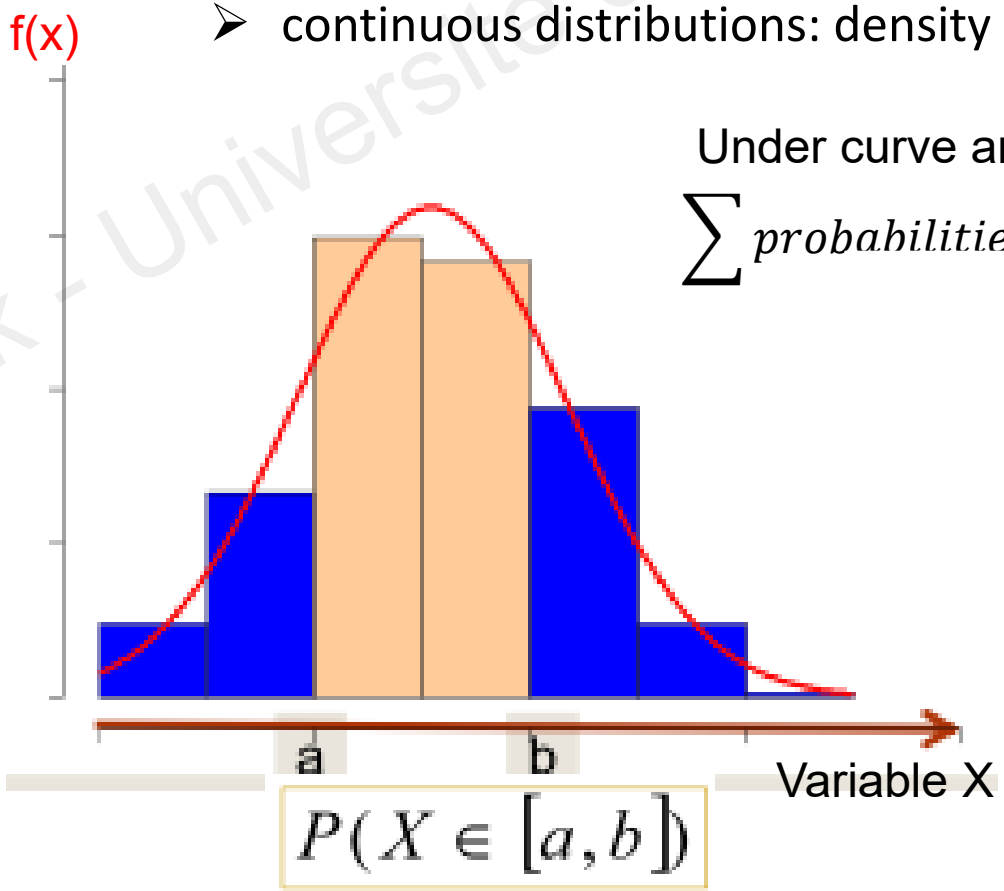
↳ characterized by a distribution function of density probability

- discrete distributions = barplots



$\Sigma\, p_i = 1$

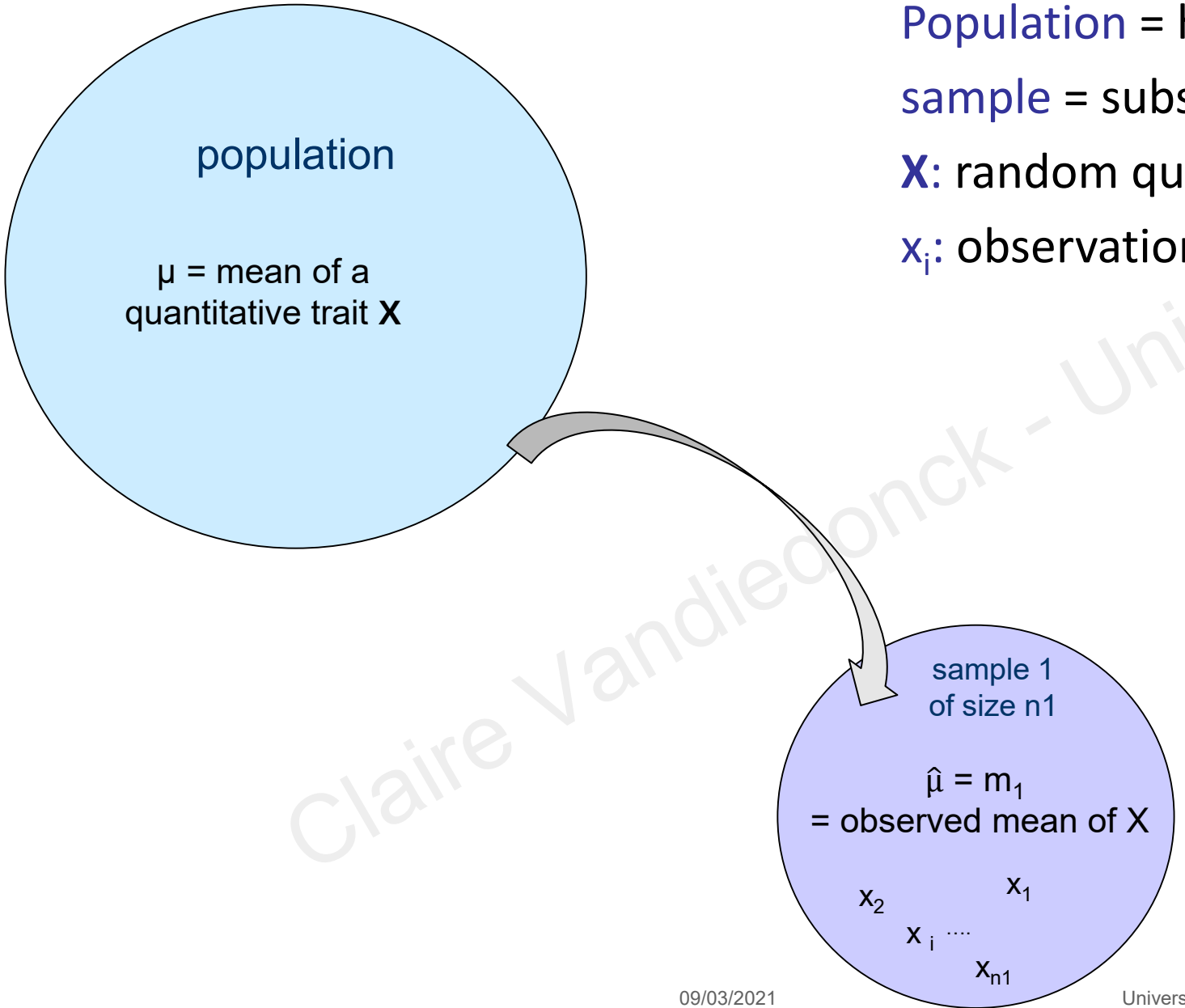
- continuous distributions: density probability



Under curve area =

$$\sum probabilities = 1$$

Population versus sample



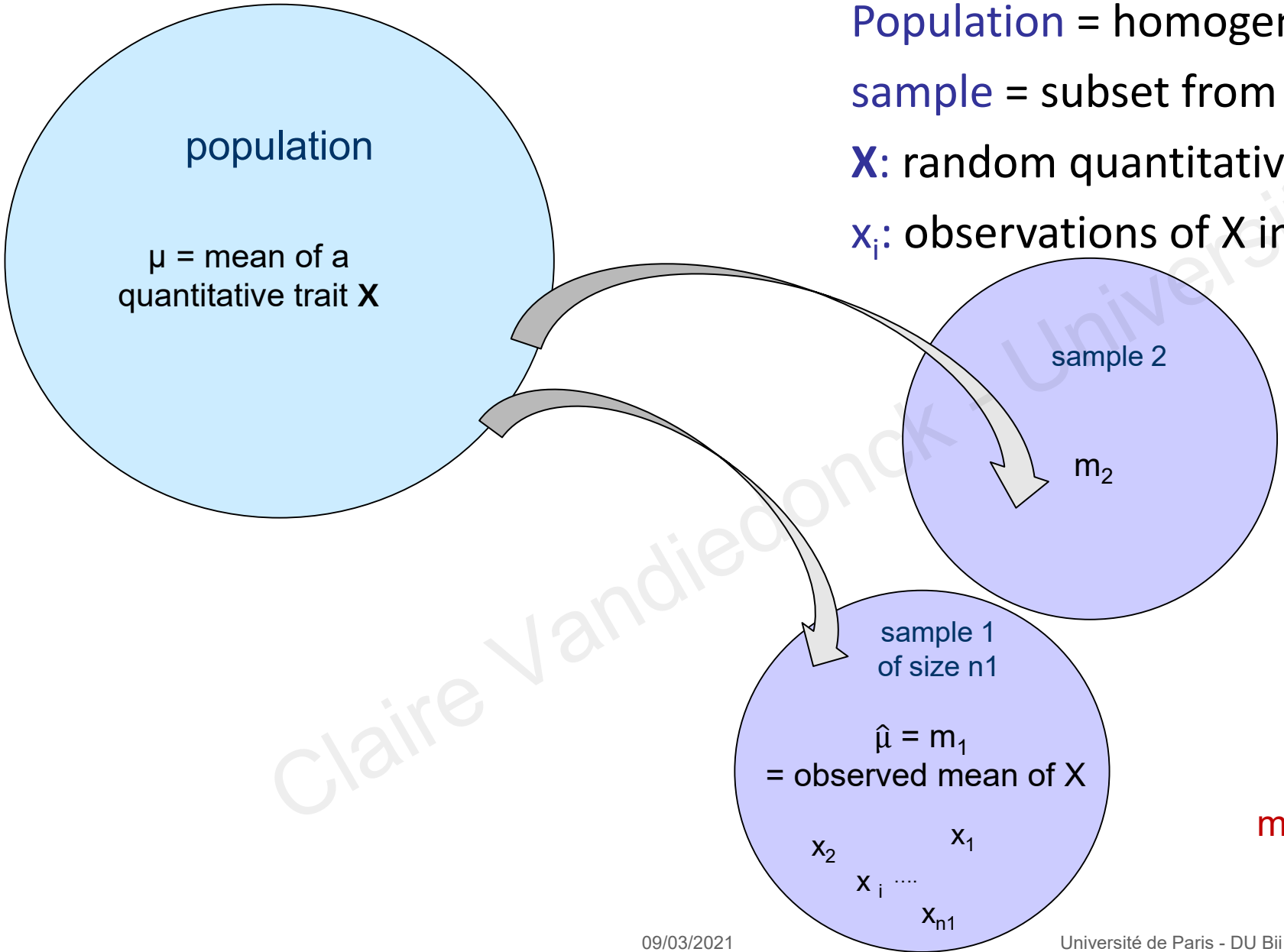
Population = homogeneous set

sample = subset from the population

X: random quantitative variable

x_i: observations of X in sample i of size n_i

Population versus sample



Population = homogeneous set

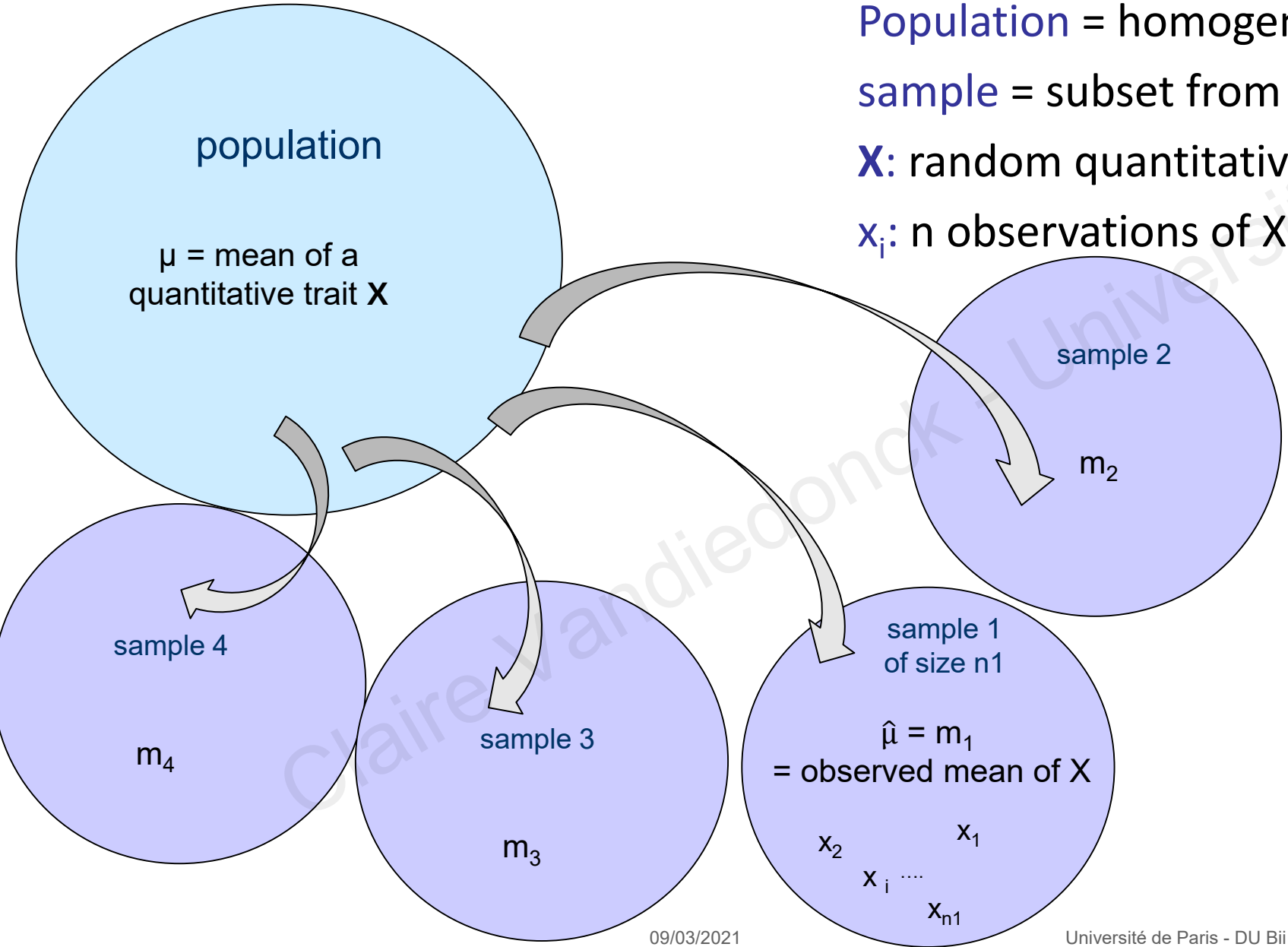
sample = subset from the population

X: random quantitative variable

x_i: observations of X in sample i of size n_i

m1 and m2 may differ due to sampling fluctuation

Population versus sample -> sampling fluctuation



Population = homogeneous set

sample = subset from the population

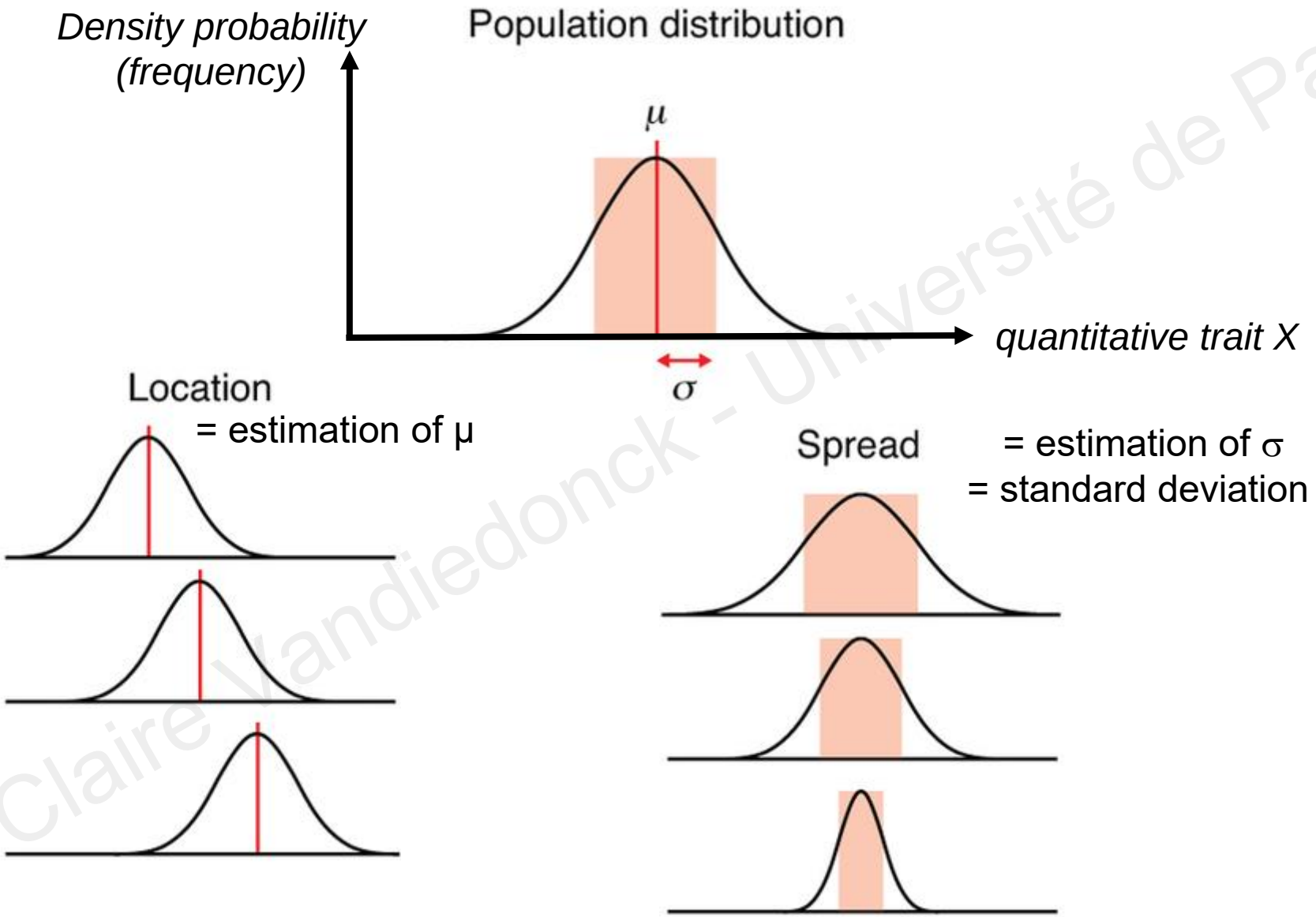
X: random quantitative variable

x_i: n observations of X in sample i of size n_i

⇒ m_i value fluctuates between samples

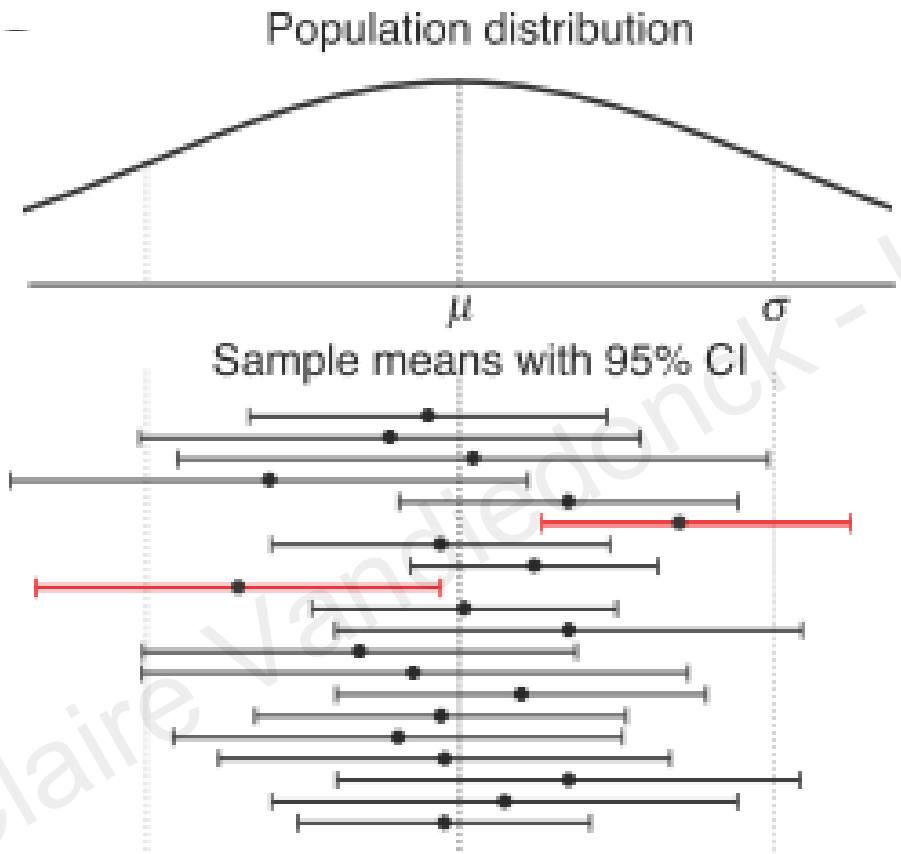
⇒ Values of a random variable $M_n = \overline{X} = \frac{\sum X}{n}$

1st aim : estimation of population parameters



Estimation with confidence intervals

95% of intervals are expected to span the mean while the other 5% (in red here) do not



$$IC_{1-\alpha} \text{ of } \mu = \left[m \pm \varepsilon_{\alpha} \sqrt{\frac{s^2}{n}} \right]$$

Live: sampling fluctuation!

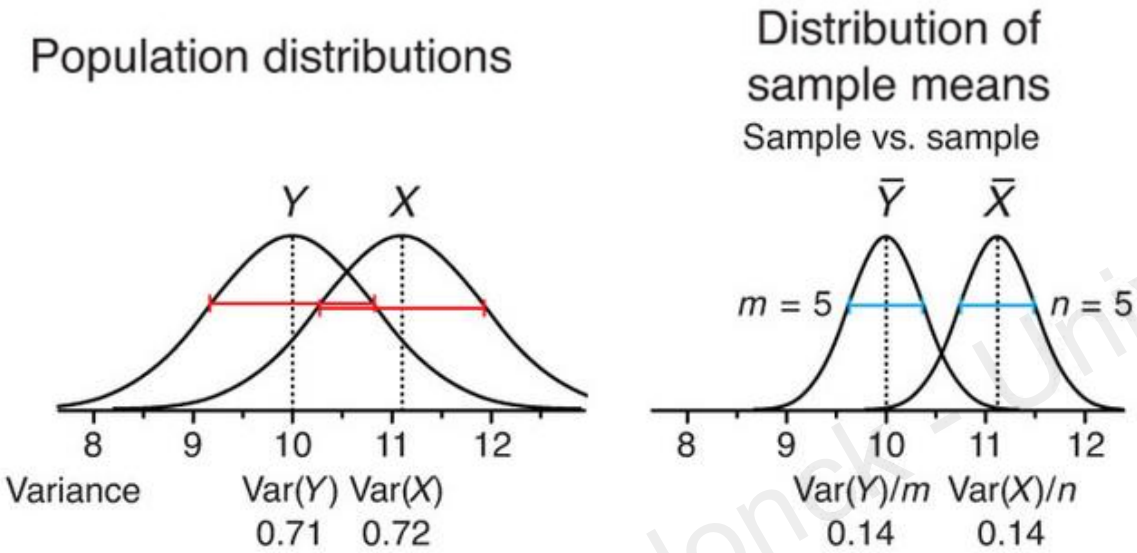
Sampling variation with a Shiny application

http://shiny.calpoly.sh/Sampling_Distribution/

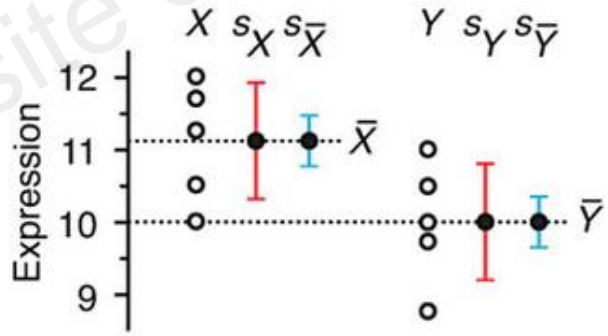
1.2. Statistical tests

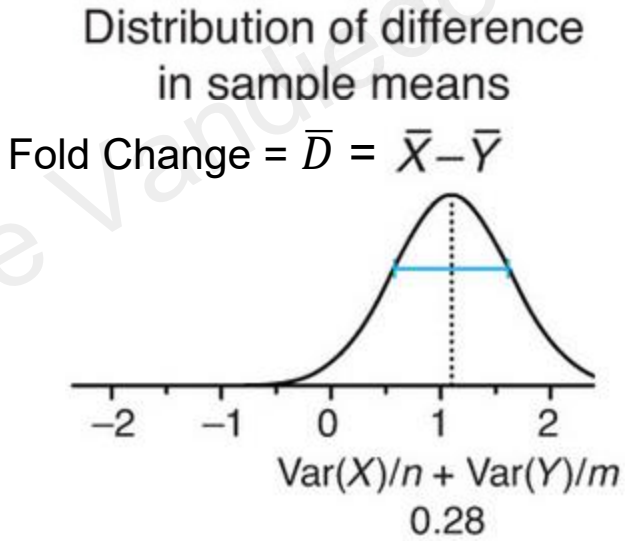
2nd aim: Comparing pouplation parameters

Comparing means of 2 populations X and Y



Two samples data of size 5





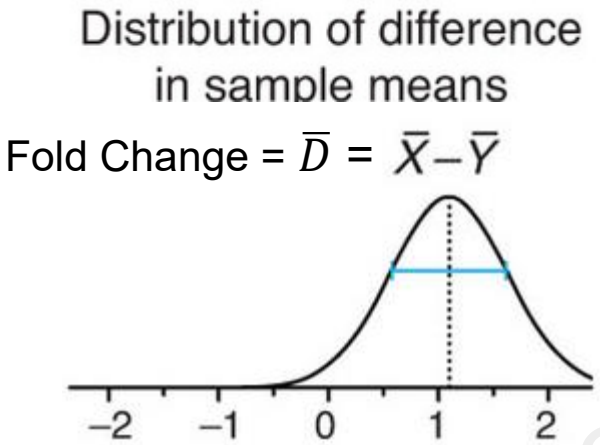
The difference of the means

$\bar{Y}$ - $\bar{X}$ = $\bar{D}$ is also a random variable

➢ Which distribution is followed by this difference $\bar{D}$?

2nd aim: Comparing pouplation parameters

Comparing means of 2 populations X and Y



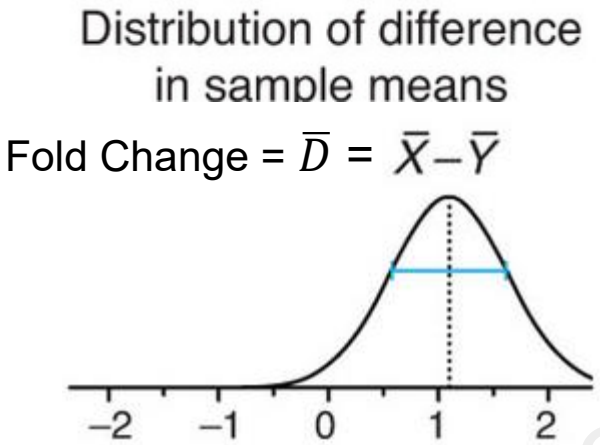
The difference of the means
$\overline{Y}$ - $\overline{X}$ = $\overline{D}$ is also a random variable

➢ Which distribution is followed by this difference $\overline{D}$?

H0: no difference = the means are identical
H1: there is a difference

2nd aim: Comparing pouplation parameters

Comparing means of 2 populations X and Y



The difference of the means
$\bar{Y}$ - $\bar{X}$ = $\bar{D}$ is also a random variable

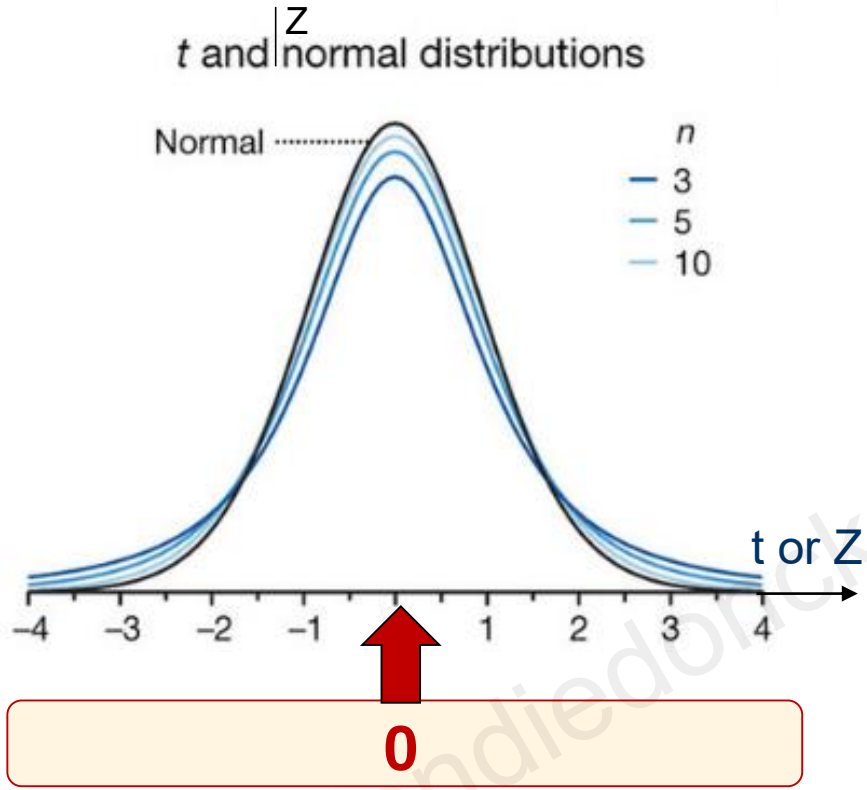
- Which distribution is followed by this difference $\bar{D}$?

H0: no difference = the means are identical
H1: there is a difference

0

Under H0: $\Delta = \mu 1 - \mu 2 = 0$
= the expected value (esperance) when there is no difference

Distribution of the differences of the means under H0



Under H0: $\Delta = \mu 1 - \mu 2 = 0$
= the expected value (esperance)
when there is no difference

$\overline{D}$ can be centered on Δ
and reduced by its standard deviation

$$\text{Z or } t = \frac{(\overline{X}-\overline{Y}) - (\mu_1-\mu_2)}{s_{\overline{X}-\overline{Y}}}$$

where $s^2_{\overline{X}-\overline{Y}} = s^2_{\overline{X}} + s^2_{\overline{Y}}$
$\approx s^2_p/n + s^2_p/m$

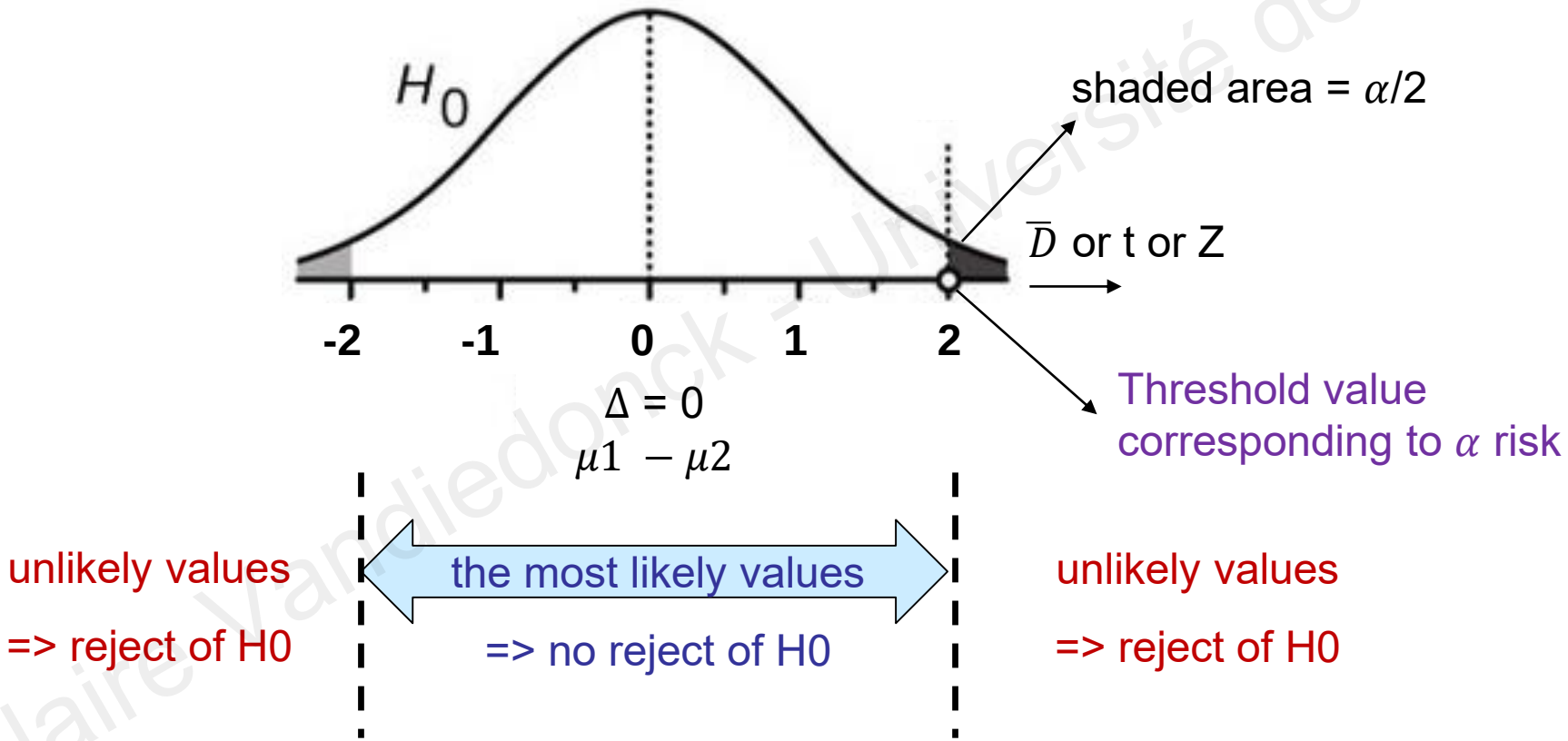
H0: no difference
H1: there is a difference

⇒ Z or t is a also random variable centered on 0 under H0

↳ How likely under the null hypothesis is the difference/statistics you observe?

Test theory: rejection criteria

Probability of observing $\overline{D}$ or t or Z under H_0



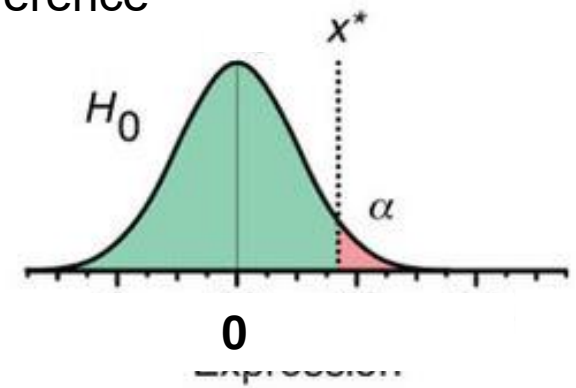
- Boundaries of the no reject area determined by alpha risk

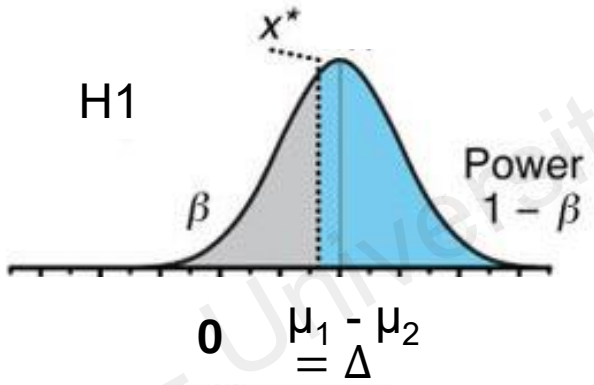
Test theory: alpha and beta risks

= no difference

$\mu_1 = \mu_2$

$\Delta = 0$

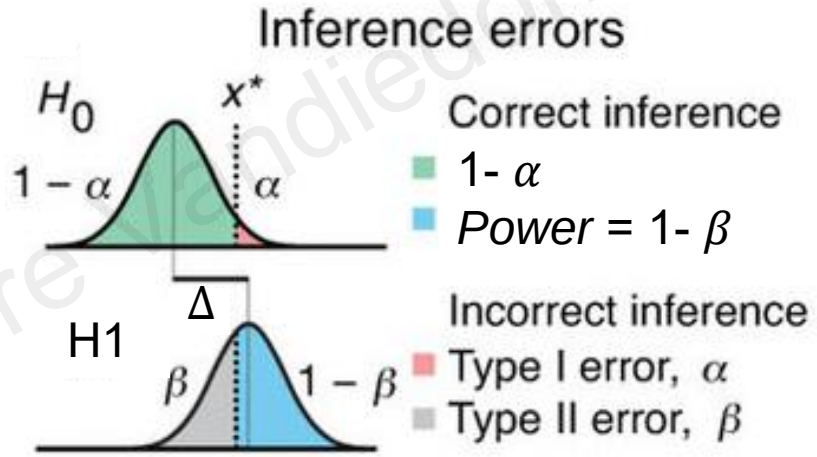




= difference

$\mu_1 \neq \mu_2$

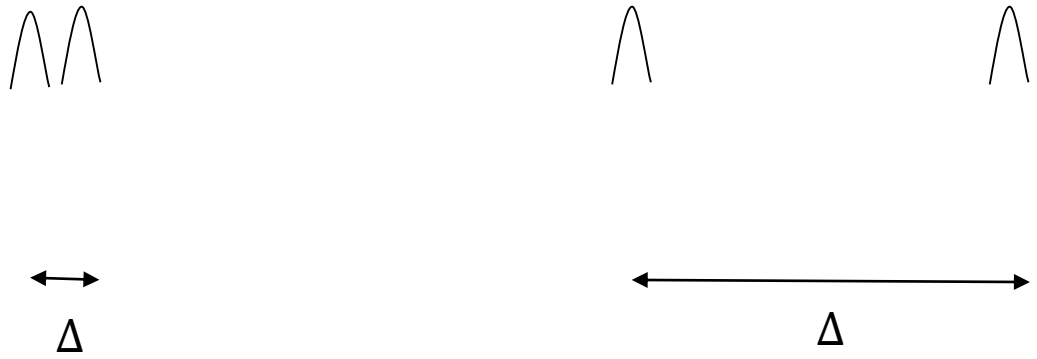
$\Delta \neq 0$



➢ Test decision	Reality H_0	Reality H_1
no reject of H_0	$1 - \alpha$ (TN)	β (FN)
reject of H_0	α (FP)	$1 - \beta$ (TP)

What does impact power?

1. Power increases with effect size (Δ)

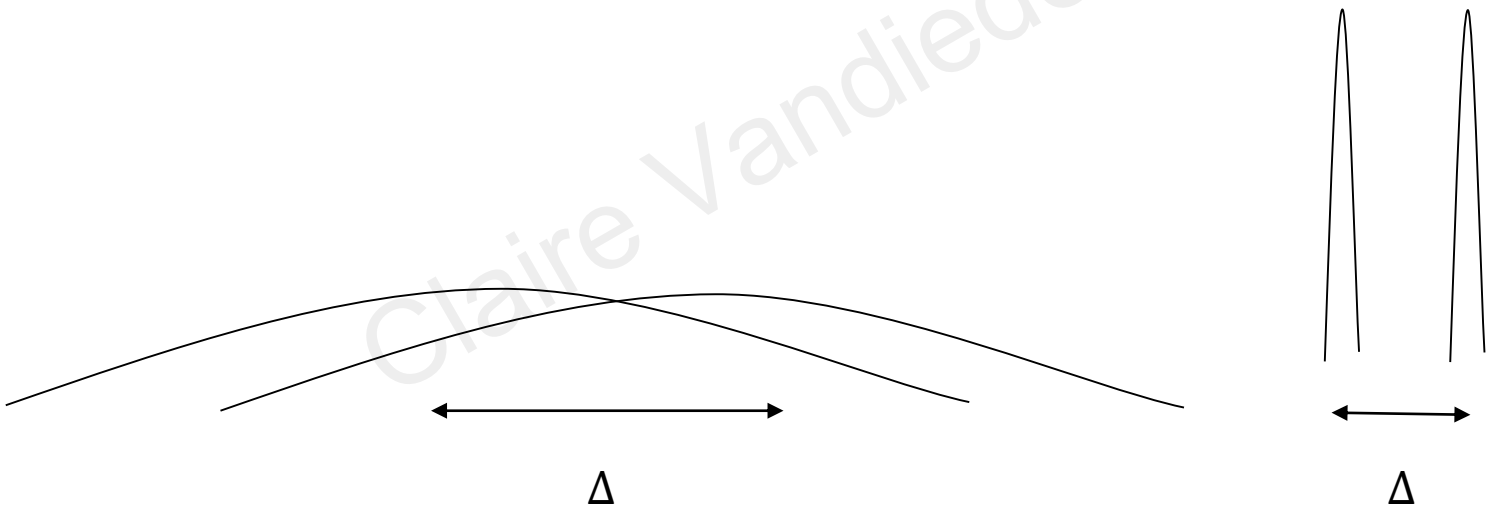


$$Z \text{ or } t = \frac{\overbrace{(\bar{X}-\bar{Y})}^{\bar{D}} - \overbrace{(\mu_1-\mu_2)}^{\Delta}}{s_{\bar{X}-\bar{Y}}}$$

where $s^2_{\bar{X}-\bar{Y}} = s^2_{\bar{X}} + s^2_{\bar{Y}}$

$\approx s^2_p/n + s^2_p/m$

2. Power increases when standard deviation decreases



3. Power increases with sample size

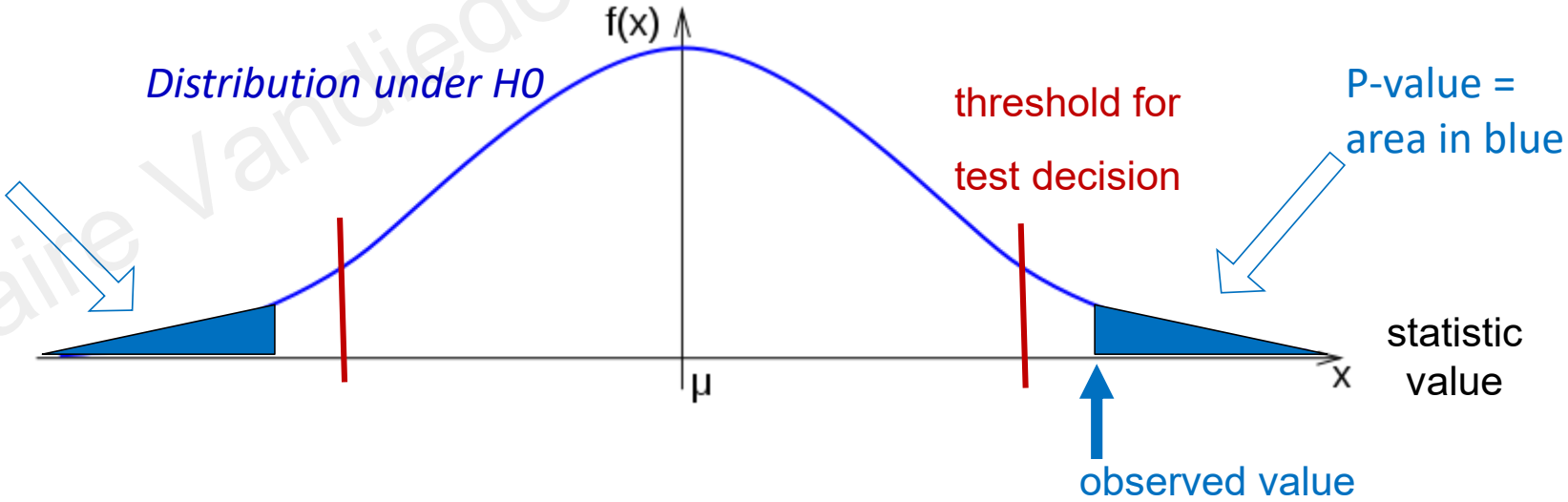
if n increases, Z increases

P-values

The p-value is defined as the probability to obtain, under H0, a value of the statistic (Student t, Z, Chi²…) at least as extreme as the observed value

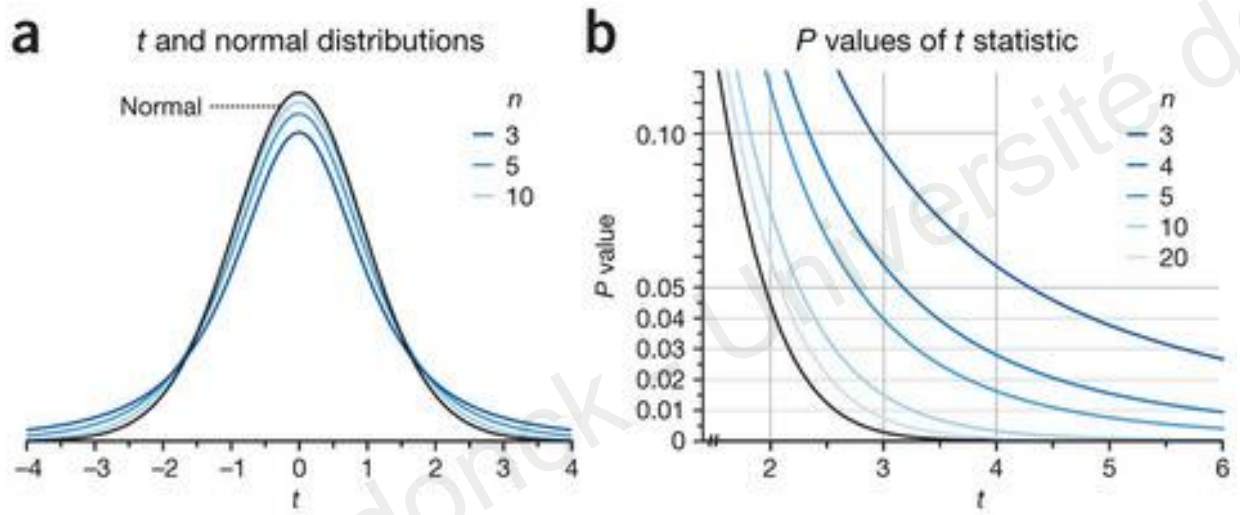
$$\textbf{pvalue} = P(\,|\text{statistics}| > \text{observed value} / H_0) \leq \alpha$$

- **report always your stat to have the direction effect + give CI of estimated effect size**
- p-value is automatically computed by software but only to report if reject of H0, i.e significant test at the α risk (otherwise report NS for not significant)
- the higher your |stat| , the lower your-pvalue



P-value in a Student test

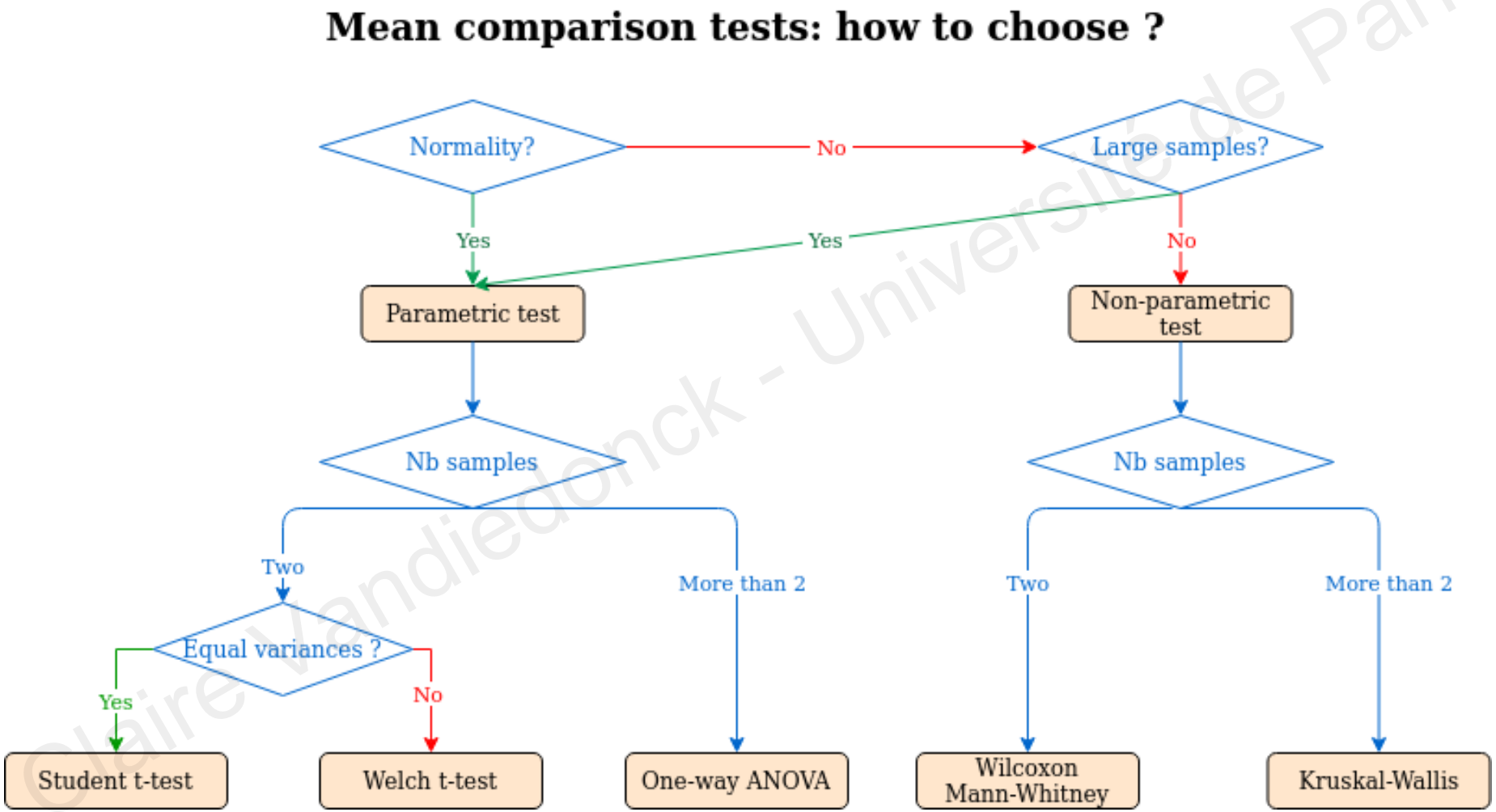
- the higher your stat (eg.|t|) , the lower your p-value, the higher your significacnce



t Table

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002
df										
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144

Which statistical test to use?



Comparing more than 2 populations

1. Perform a global test

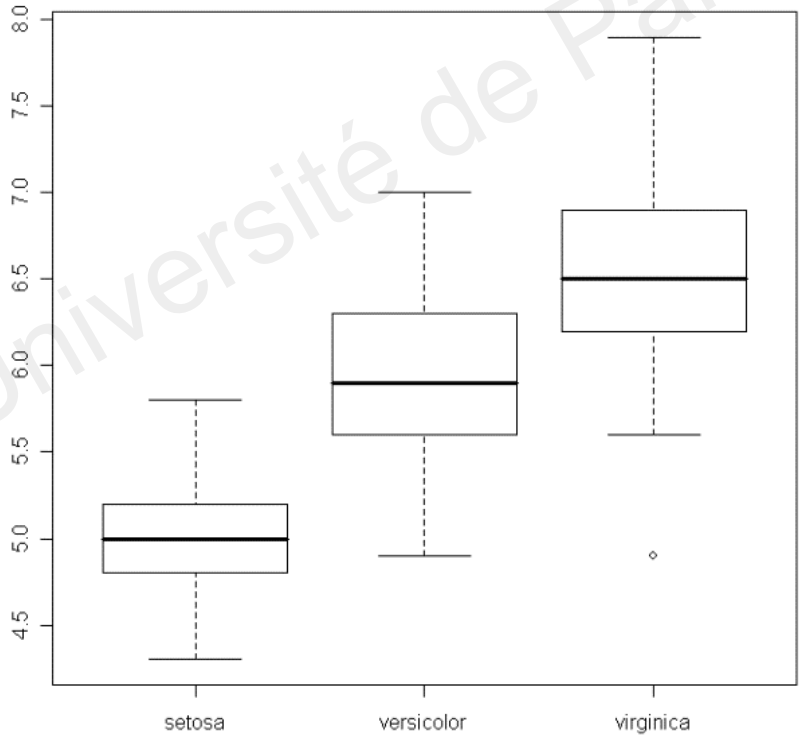
= one-way ANOVA

H_0 : all population means are equal

H_1: at least one of the means differs

- the test compares the ratio of the variance among the sample means to the variance of each sample

2. If significant, perform pair-wise comparisons = post-hoc tests



Linear regression: perfect for more cpmplex situation

It is useful to consider a model for the observed data (on a single trait)

See session 4!

$$Y = \mu + \alpha + \beta + \gamma + \ldots + \text{error}$$

eg. Microarray expression of a single gene Y =log2(intensity)

μ is the mean over all samples (all conditions)

error is the random error that is a mixture of measurement error and biological variability

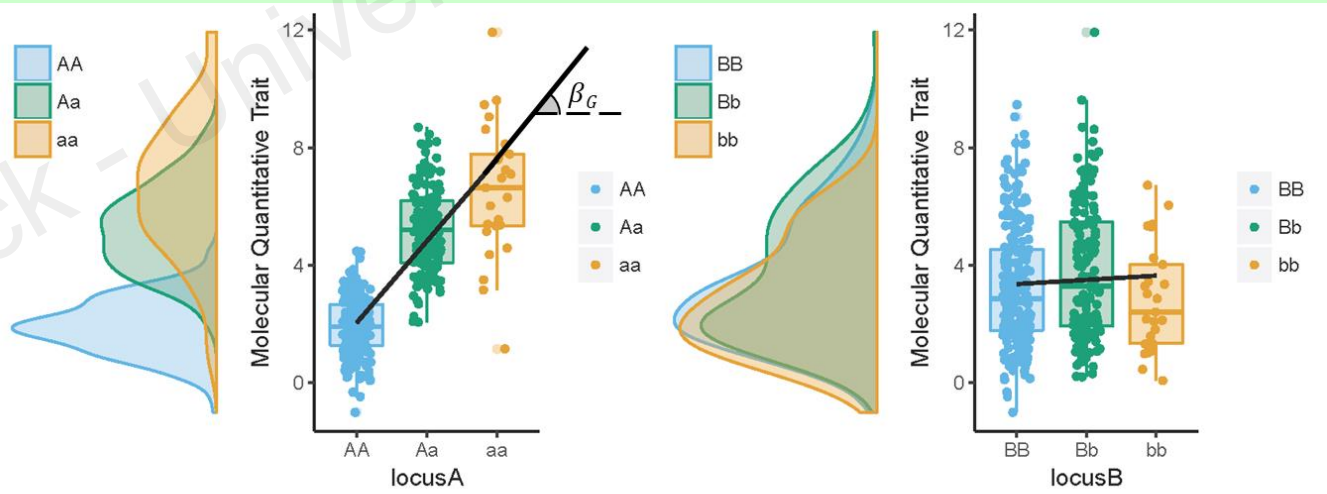
the other terms are systematic deviations from the mean, due to the factors of interest (treatments, tissue…) and technical effects (batch, platform,…)

➢ We test the simplest model:

$$H_0: Y = \mu + \text{error} \text{ while } \alpha, \beta \ldots = 0$$

Example: testing a genetic variant on expression

$$Y = \beta_0 + \beta_G X + \varepsilon \quad \begin{cases} H_0: \beta_G = 0 \\ H_1: \beta_G \neq 0 \end{cases}$$



$\beta_G \neq 0$
locus A is molecularQTL

$\beta_G = 0$
locus B is not a molecularQTL

=> Extendable to more complicated models with several factors and interactions

Practical in live !

1. Testing mean comparison for a single trait
-> impact of sample size, mean difference and variance

2. Multiple testing issue

1.3. Introduction to stat-omics : making sense of omic's data

Hétérogénéité des données omiques

Nature des données

- binaires (eg. présence ou absence d'un allèle ou d'un site de liaison)
- catégoriques (séquences de site consensus, isoforme exprimée)
- quantitative discrète (génotypes: 0, 1, 2)
- quantitative continue (niveau d'expression d'un gène ou d'une protéine)

Dimension des données (*exemples chez l'homme*)

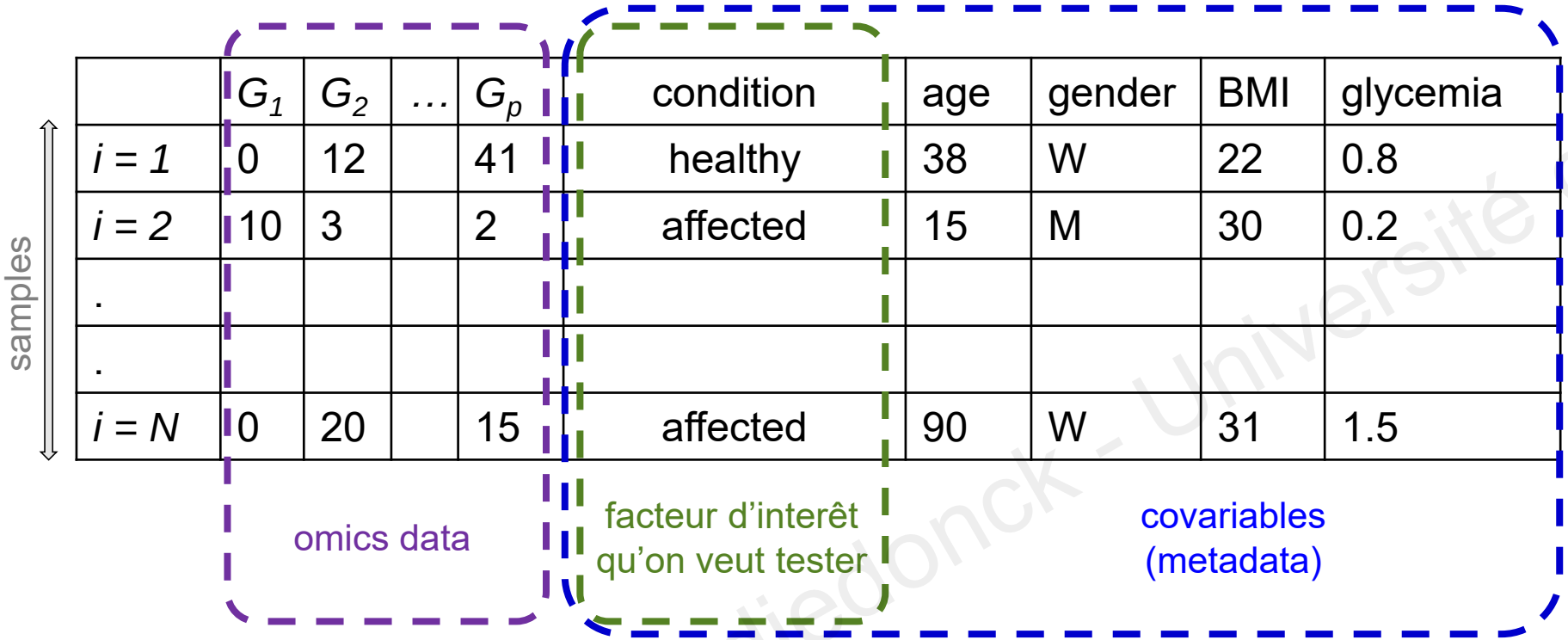
- génome ($4x10^6$ de variants bi-alléliques de type SNP)
- transcriptome (20-60 000 gènes, 200 000 transcrits)
- protéome (18 000 protéines, 293 000 peptides)

Données manquantes (4000 protéines)

Structure des données

- corrélations entre les variables mesurées (déséquilibre de liaison, co-expression…)
- corrélations entre les types de données

Des données non-omiques peuvent exister: covariables



	G_1	G_2	…	G_p	condition	age	gender	BMI	glycemia
i = 1	0	12		41	healthy	38	W	22	0.8
i = 2	10	3		2	affected	15	M	30	0.2
.									
.									
i = N	0	20		15	affected	90	W	31	1.5

- Par exemple, on peut avoir le niveau d'expression par gène pour chaque échantillon
- On peut aussi avoir des données cliniques pour les échantillons incluant le facteur d'intérêt qu'on veut tester et d'autres covariables qui pourraient impacter les niveaux d'expression

➢ **On souhaite expliquer les variations d'expression (variable expliquée) en fonction de covariables cliniques (variables explicatives)**

Variable to explain ~ explanatory variables + covariates + residual error

Quels facteurs peuvent expliquer la variation d'un trait?

Variation inter-groupes

1. Facteur/covariables d'intérêt => design experimental
   - ✓ conditions expérimentales testées: stimulus, traitement, temps, maladie…
   - ✓ variabilité génétique: mutation
   - ✓ tissus/type cellulaire…
   - ✓ etc...

2. Variation technique: réplicats techniques
   - ✓ experimental: lot, jour, expérimentateur, temperature ambiante…
   - ✓ multiplexage
   - ✓ variation de plate-forme
   - ✓ etc...

Variation intra-groupes

Variation biologique => réplicats biologiques

- ✓ fluctuation d'échantillonnage

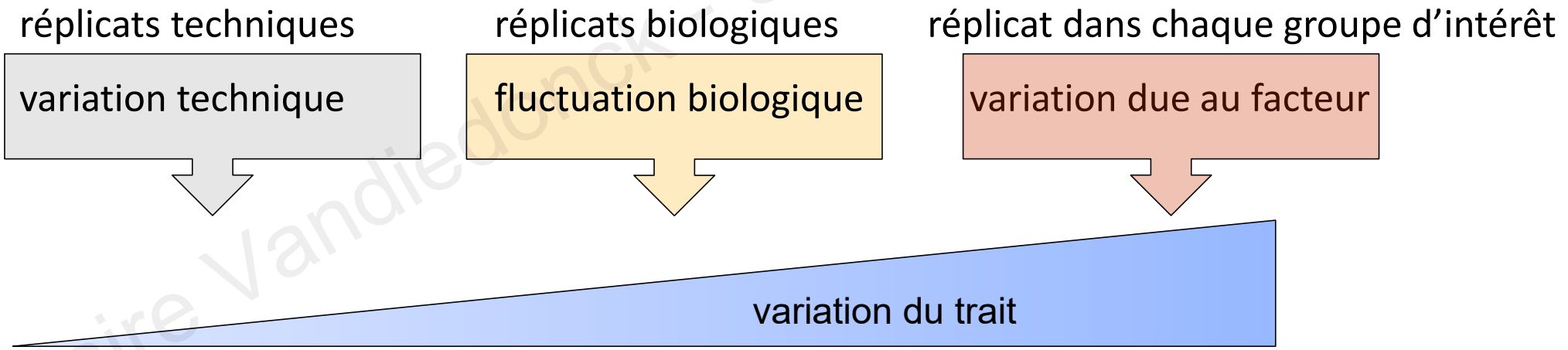
De l'importance d'un bon design experimental

Les différences entre les conditions peuvent uniquement être testées uniquement

si des **REPLICATS** sont inclus

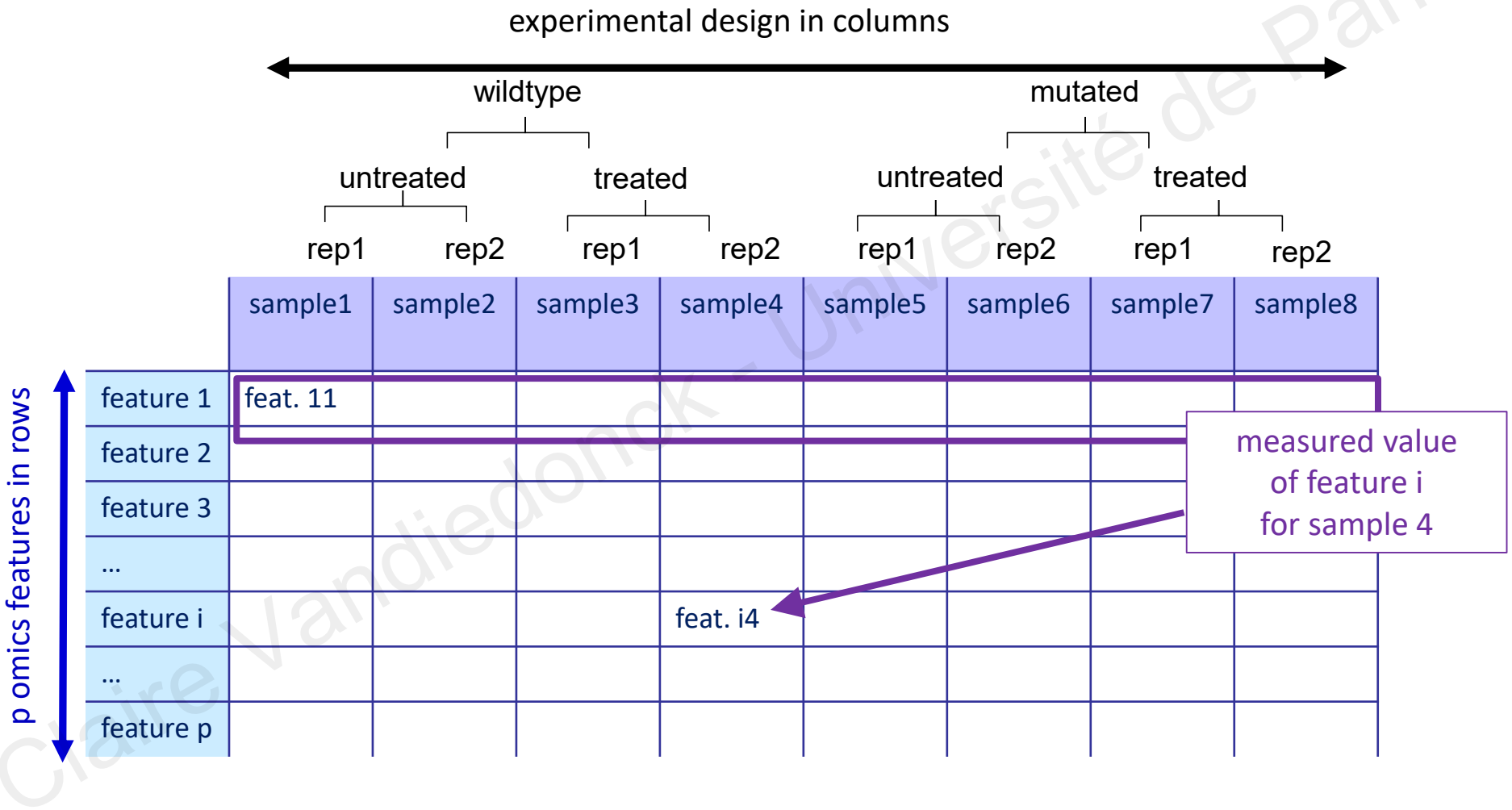
⇒ permettent de determiner quelles differences sont dues aux fluctuations aléatoires d'échantillonage

☞ Ideal scenario :



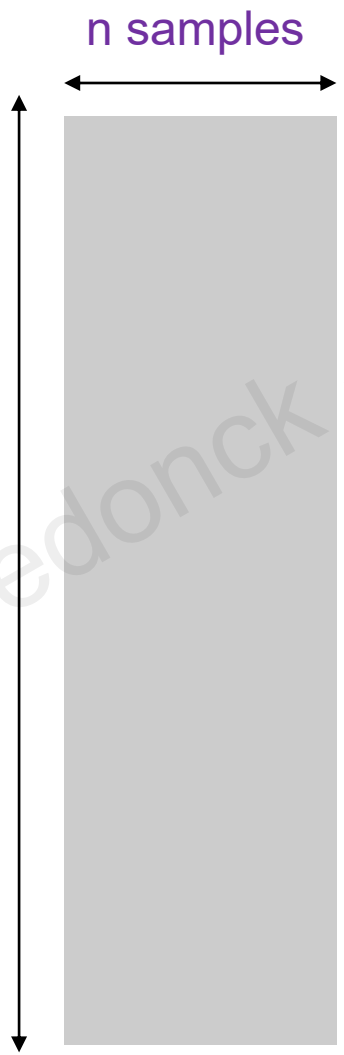
La structure des données omiques

Matrice de données



Les problèmes de diemnsionalité des données

p >> n

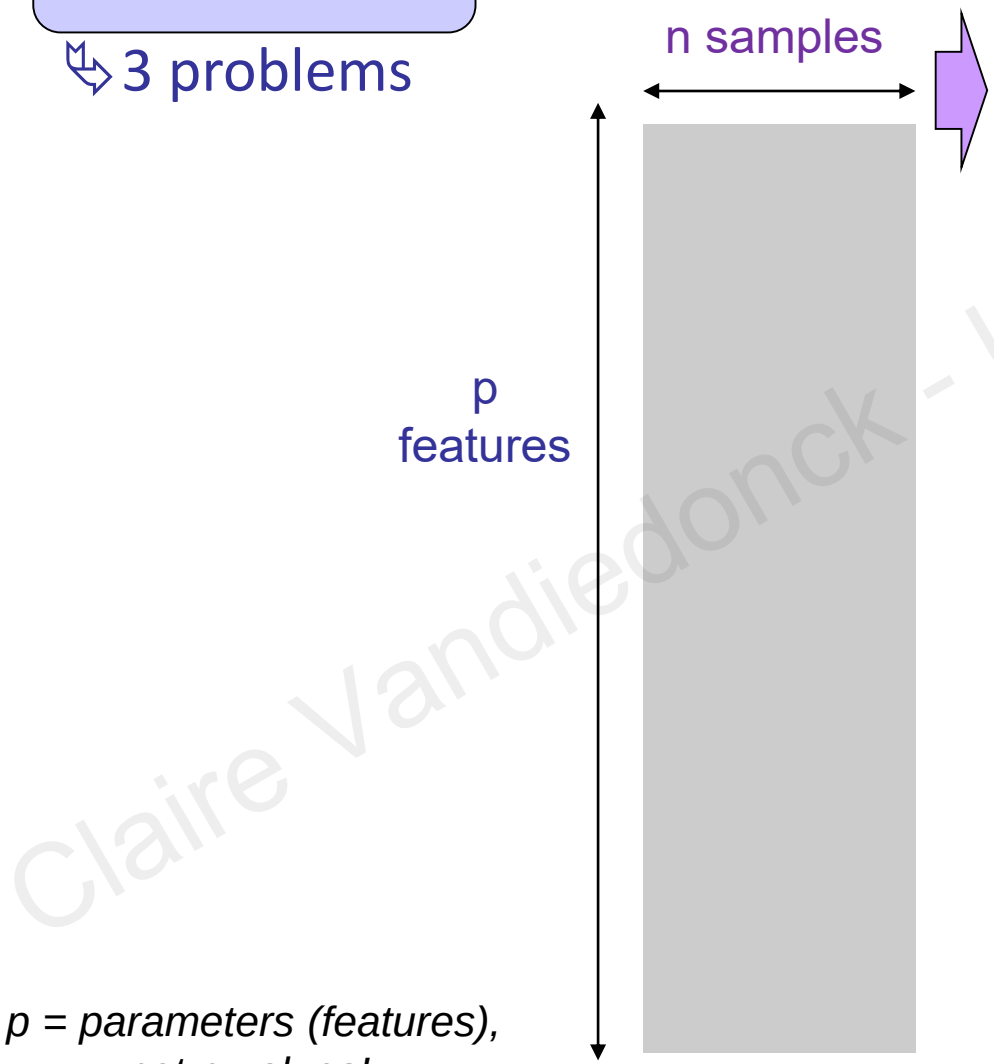


p = parameters (features), not p-values!

Les problèmes de diemnsionalité des données

p >> n

↳ 3 problems



n small:

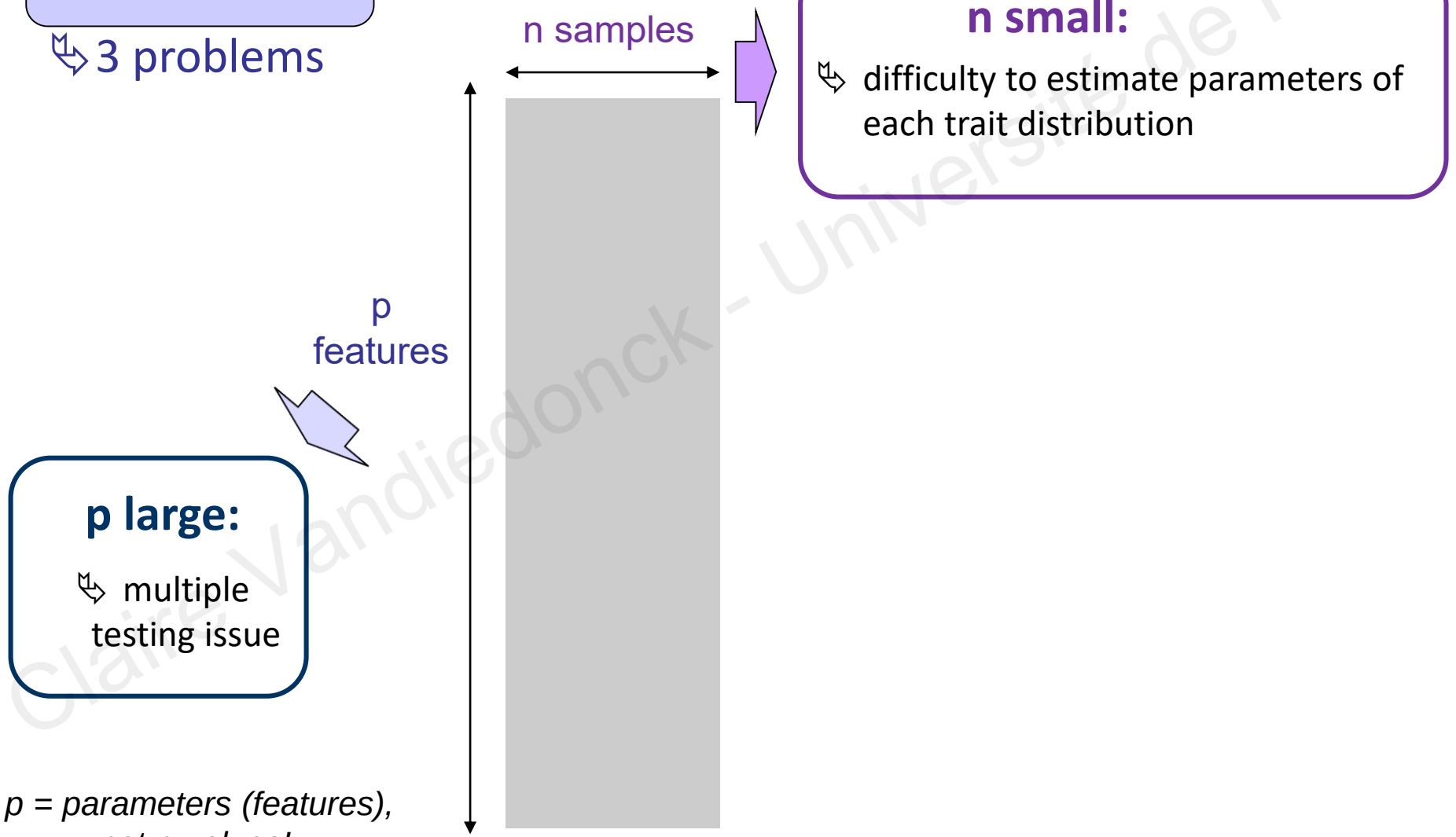
↳ difficulty to estimate parameters of each trait distribution

p = parameters (features), not p-values!

Les problèmes de diemnsionalité des données

p >> n

↳ 3 problems



n small:

↳ difficulty to estimate parameters of each trait distribution

p large:

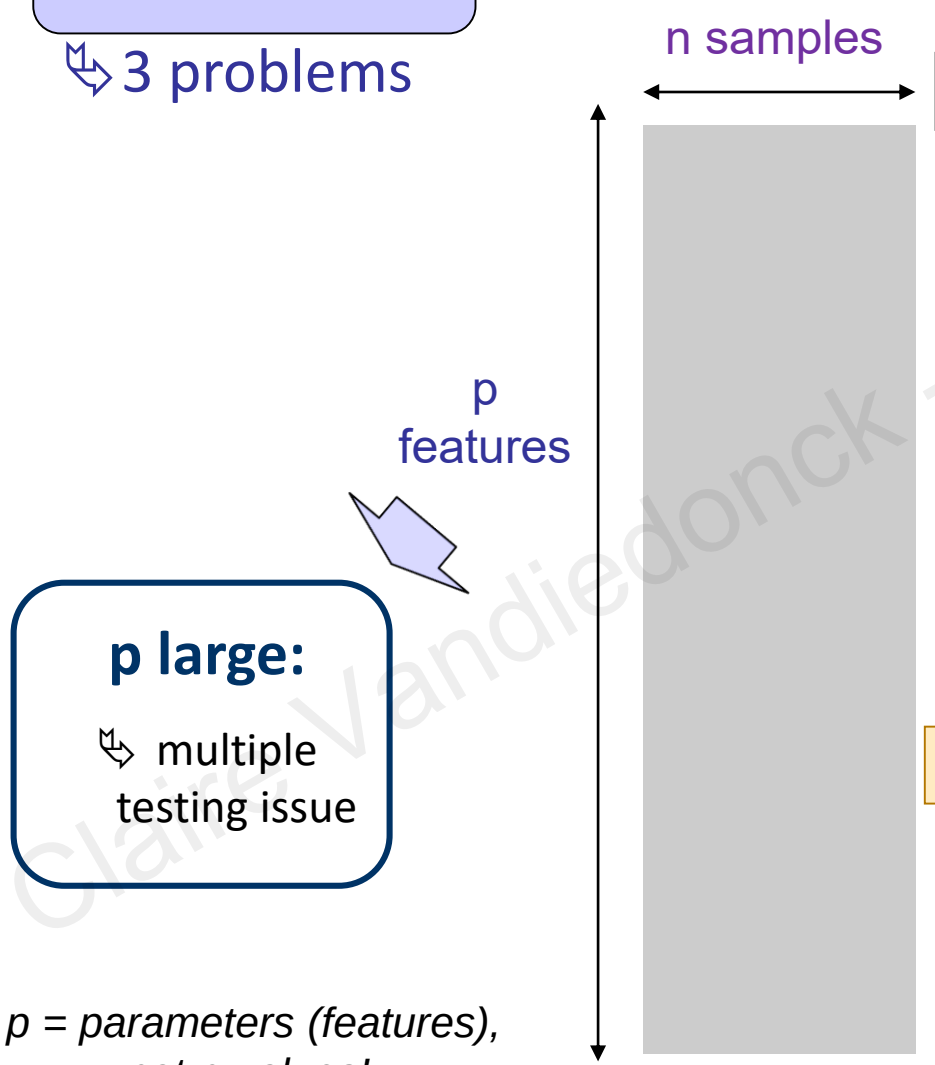
↳ multiple testing issue

p = parameters (features), not p-values!

Les problèmes de diemnsionalité des données

p >> n

↳ 3 problems



n small:

↳ difficulty to estimate parameters of each trait distribution

p large:

↳ multiple testing issue

Correlation between traits

↳ difficulty to estimate because n small

↳ redundancy: too many tests?

p = parameters (features), not p-values!

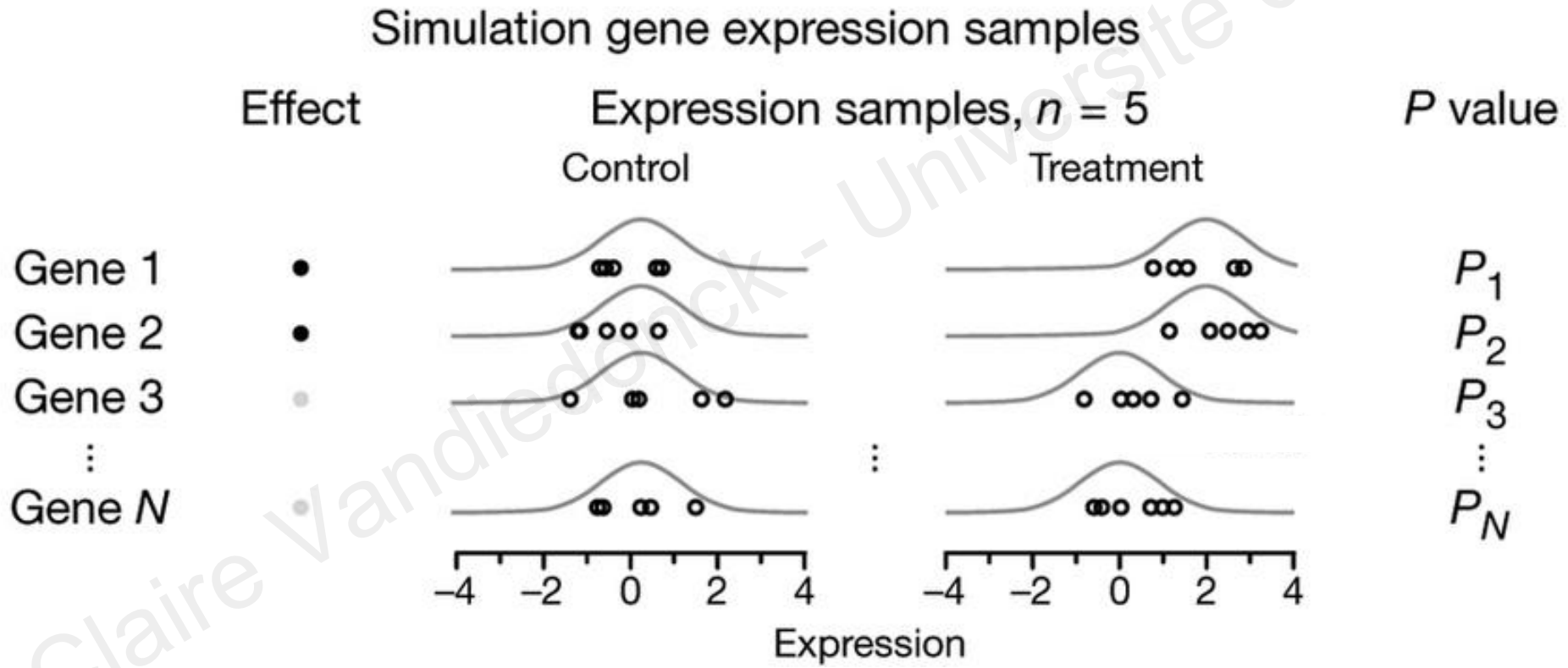
1.4. The 1st issue: multiple testing

The problem

We perform multiple tests = one per feature/trait

↳ for each feature, we either reject or not H0 at a risk α = PCER

= per-comparison error rate



Why is this problem so important?

Omics are big data:

A typical microarray or RNA-seq experiment: 10,000 genes

=> as many hypothesis tests

Just one hypothesis test:

For an $\alpha = 0.05$, we tolerate to reject H_0 wrongly 5% of the times

↳ but for 10,000 tests the number of false positives goes up to 500

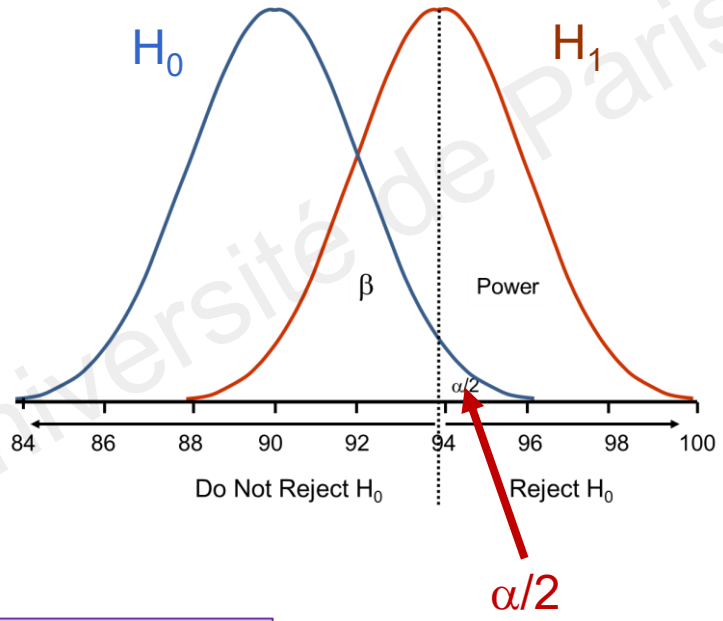
=> too many!!!

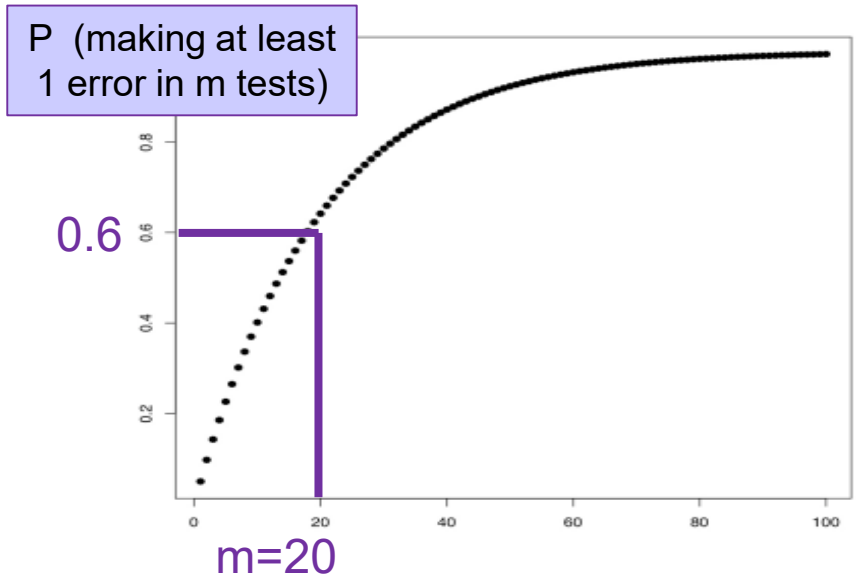
Expected value (**e-value**)

- Expected number of FP = E(FP)= $m\alpha$

Family-wise error rate (FWER)

- P(making an error) = α
- P(not making an error) = $1 - \alpha$
- P(not making an error in m tests)= $(1-\alpha)^m$
- **FWER** = P(making at least 1 error in m tests) = $1 - (1-\alpha)^m$





Counting errors

Decision on H_0	H_0 True	H_1 True	
reject	V (incorrect)	S	R
do not reject	U	T (incorrect)	m-R
	m_0	$m-m_0$	m

m = number of tests

R = number of rejected H_0

m_0 = number of true H_0

➢ only m and R are observed!

V = number of type I errors = false positives

By the way, where are:

the false negatives?

the true positives?

the true negatives?

Counting errors

Decision on H_0	H_0 True	H_1 True	
reject	V (incorrect)	S	R
do not reject	U	T (incorrect)	m-R
	m_0	$m-m_0$	m

m = number of tests

R = number of rejected H_0

m_0 = number of true H_0

➢ only m and R are observed!

V = number of type I errors = false positives

By the way, where are:

the false negatives?

the true positives?

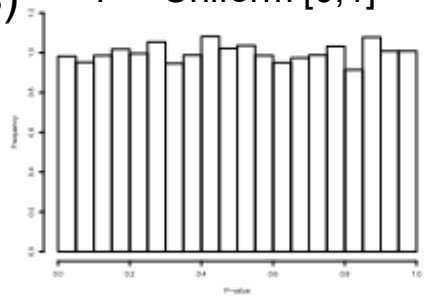
the true negatives?

	H0 True	H1 True
Reject H0	FP	TP
No reject	TN	FN

Controlling the type I error rate

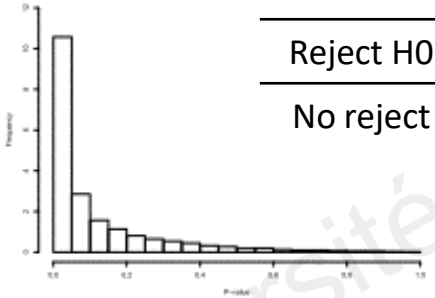
f(pvalues)

P ~ Uniform [0,1]



pvalues

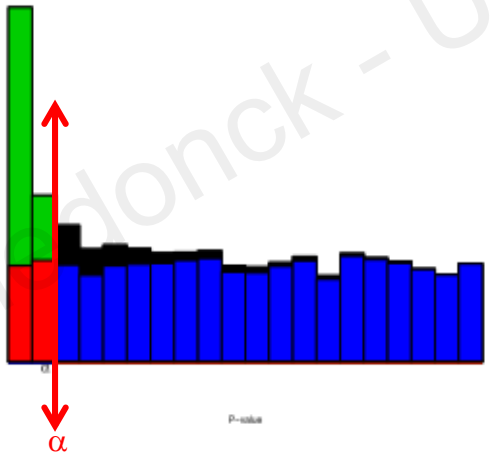
p values for null genes



p values for non-null genes

	H0 True	H1 True
Reject H0	FP	TP
No reject	TN	FN

p values for all genes



True positives = S
False positives = V
True negatives = U
False negatives = T

Where to set the threshold of significance to control the type I error rate?

=> Trade-off between type I error and power!!

☞ Storey JD, Tibshirani R. Statistical significance for genomewide studies. PNAS. 2003 100:9440-5. PMID: 12883005; PubMed Central PMCID: PMC170937.

Bonferroni correction

Aim: to control the family-wise error rate (FWER):

= the error rate across the whole collection/family of hypothesis tests

= FWER = P(V ≥ 1) = probability of ≥ 1 false positive among all tests

↳ By "adjusting" the p value with the Bonferroni correction

set $\alpha' = \alpha/m$

reject hypotheses if $p < \alpha'$

✓ E.g. for a type I error rate of 0.05 per experiment (PCER)

and m= 10 000 tests: $\alpha' = 0.05/10{,}000 = 5x10^{-6}$

very popular

the problem for "Omics" experiments: very conservative

=> alternative approaches investigated: very active area of current research in statistics!

False Discovery Rate (FDR)

We focus on positive tests (H_0 rejected):

FDR = proportion of false positive among the set of rejected hypotheses (the "discoveries"):

- ✓ FDR = V/R

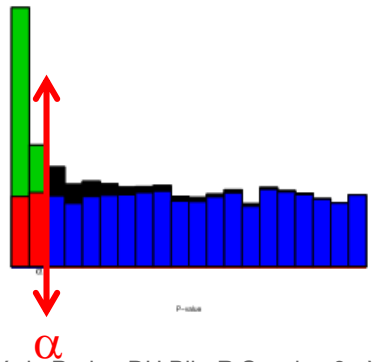
Decision on H_0	H_0 True	H_1 True	
reject	**V** (incorrect)	**S**	R
do not reject	**U**	**T** (incorrect)	m-R
	m_0	$m-m_0$	m

A related parameter
= the False Positive Rate (FPR)

- ✓ FPR = V/m_0

Decision on H_0	H_0 True	H_1 True	
reject	**V** (incorrect)	**S**	R
do not reject	**U**	**T** (incorrect)	m-R
	m_0	$m-m_0$	m

	H0 True	H1 True
Reject H0	FP	TP
No reject	TN	FN

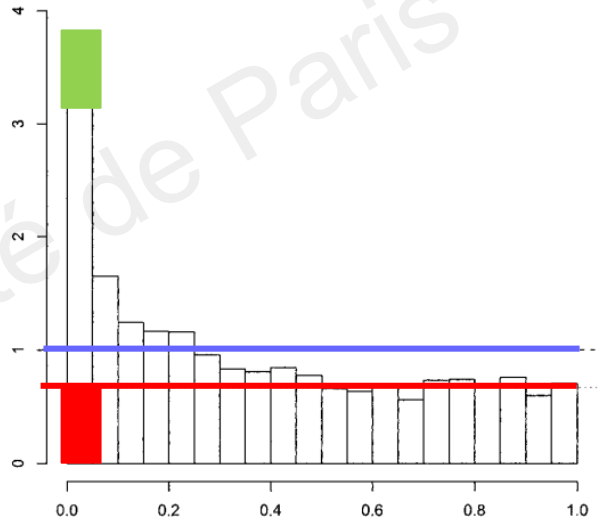


Q values

Qvalue of a gene = expected proportion of false positives when calling that gene significant

- ✓ the q-value depends on the p-value for the test of the gene and on the distribution of the entire set of p-values from the family of tests being considered (Storey and Tibshiriani 2003)
- ✓ Thus, in a microarray study testing for differential expression, if gene X has a q-value of 0.013 it means that 1.3% of genes that show p-values at least as small as gene X are false positives
- ✓ The maths:
  - π_0 : the proportion of true null tests (TN)
  - $\alpha m \pi_0$: the number of false positives (FP)
  - $\alpha m \pi_0 / R$: an estimate of the FDR (V/R)

	H0 True	H1 True	
Reject H0	FP	TP	R
No reject	TN	FN	m-R
	m_0	$m-m_0$	m



Benjamini-Hochberg procedure

To control FDR at level δ:

- ✓ order the unadjusted p-values: p1<p2<...<pm
- ✓ find the test with the highest rank, j, for which the p value,

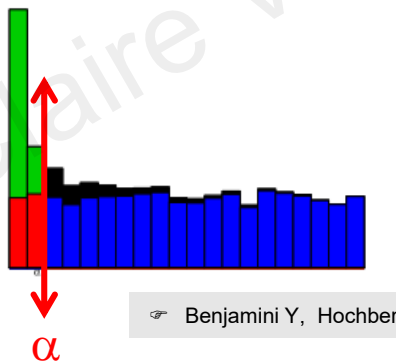
$$p_j \leq \delta \frac{j}{m}$$

- ✓ Declare the tests of rank ≤ j as significant

Example: m = 10 and δ = 0.05

Rank (j)	P-value	(j/m)× δ	Reject H_0 ?	Adj. P val: p_j x m /j
1	0.0008	0.005	1	0.008
2	0.009	0.010	1	0.045
3	0.018	0.015	0	0.06
4	0.030	0.020	0	0.075
5	0.032	0.025	0	0.064
6	0.048	0.030	0	0.08
7	0.350	0.035	0	0.5
8	0.781	0.040	0	0.976
9	0.900	0.045	0	1
10	0.993	0.050	0	0.993

Values expected for a uniform distribution of p_j between 0 and delta



☞ Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. JSTOR.1995, https://www.jstor.org/stable/2346101.

1.5. The 2nd issue: estimation of traits distribution (mean and variance)

To estimate or not to estimate?

1. No estimation when using non-parametric tests
   - less power if data fit with parametric distribution
   - not suitable for designs with several factors

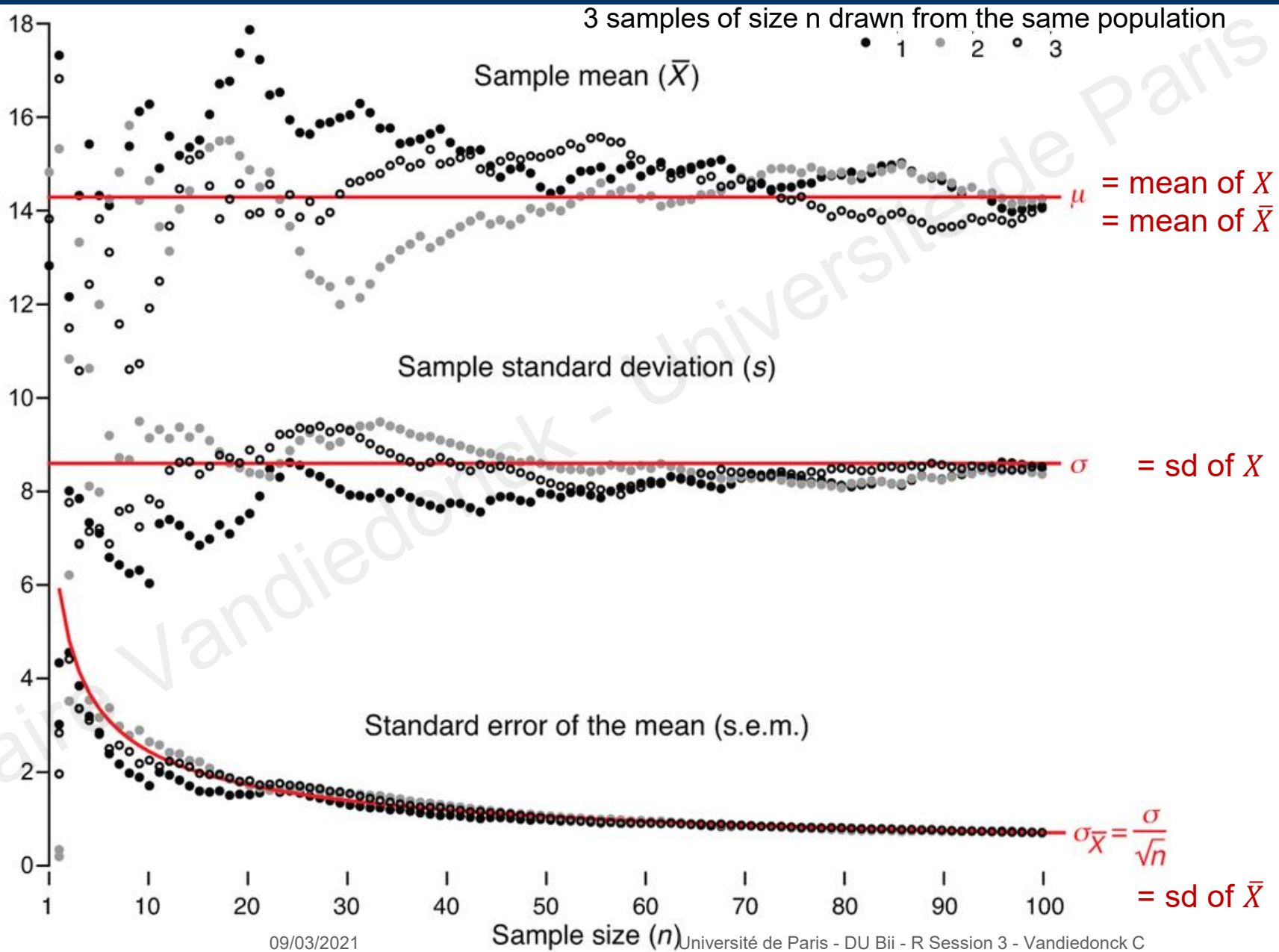
2. Random re-sampling

➢ approaching the distribution of p-values/statistics under null hypothesis by **permutation** (no replacement) of the levels of the factor of interest in the dataset => the empirical pvalue is the probability of observing the pvalue/statistic under the empirical distribution (cannot be lower than 1/1000 if 1000 permutations)

➢ estimating the CI of the distribution parameters by **bootstrap** (replacement) of the quantitated trait among all observed values within the dataset without changing the levels of the factor of interest
   - computationally intensive

3. Selecting a distribution law fitting the data
   - estimation of mean and variance
   - parametric tests

Better estimation when sample size is increased



Transcriptome data: several distribution laws

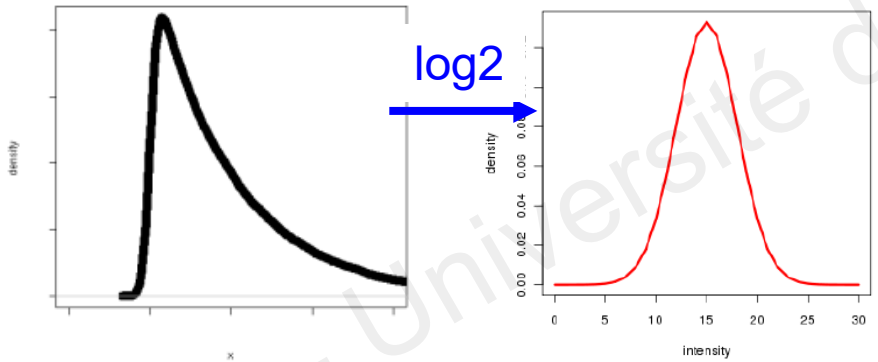
Microarrays

the abundance of each sequence depends on the fluorescence level

= intensities

↳ *continuous quantitative variables*

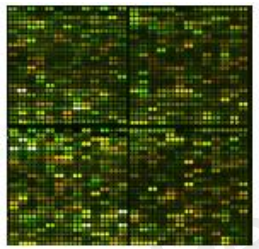
- ✓ asymmetrical distribution
- ✓ log2 intensities behave better



Student law

distribution asymétrique à droite
-> le passage en log2 donne souvent une distribution 'normale' (**1er sens de normalisation**)

2-color array:
2 samples hybridized



Y_{ga} = log2(Red/Green)

1-color array:
1 sample hybridized

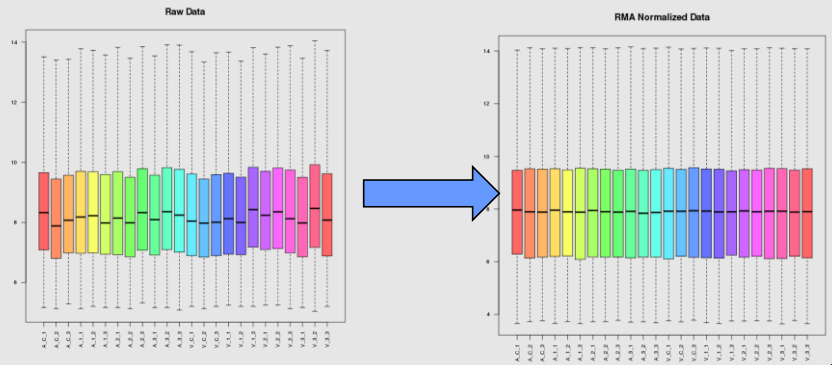
Y_{ga} = log2(intensity)

summarized over probes

gene g (or probe)

array a

Il est aussi nécessaire de **normaliser les échantillons entre eux (2ème sens de normalisation) pour pouvoir les comparer** (même échelle)



Estimating mean and variance in microarray experiments

Gene expression values are given by fluorescence intensities

- continuous variables
- assumed to fit a Student t distribution (after log2 transformation) of the difference mean

$$t_{\text{gene i}} = \frac{\bar{x}_i}{\tilde{s}_i/\sqrt{n}}$$

- but low number of replicates => difficult to estimate the variance

⇒ LIMMA (Linear Model for MicroArray experiments)

- uses a “moderated” t statistics using information from all genes (group of genes g like gene i) to estimate the variance

$$\tilde{t}_{\text{gene i}} = \frac{\bar{M}_i}{\tilde{s}_g/\sqrt{n}}$$

- allows for linear models
- design matrix => the factors to be accounted for in the model
- contrast matrix => which comparisons are of interest
- accounts for multiple testing: computes adjusted p-value (FDR B-H)

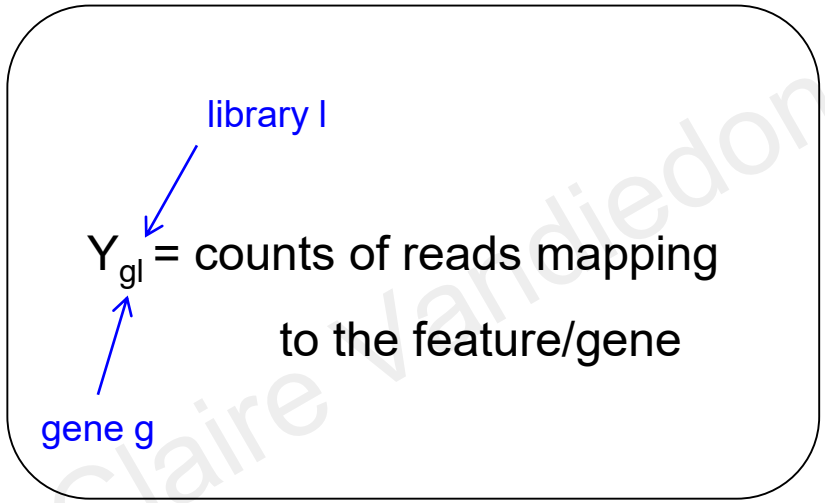
Transcriptome data: several distribution laws

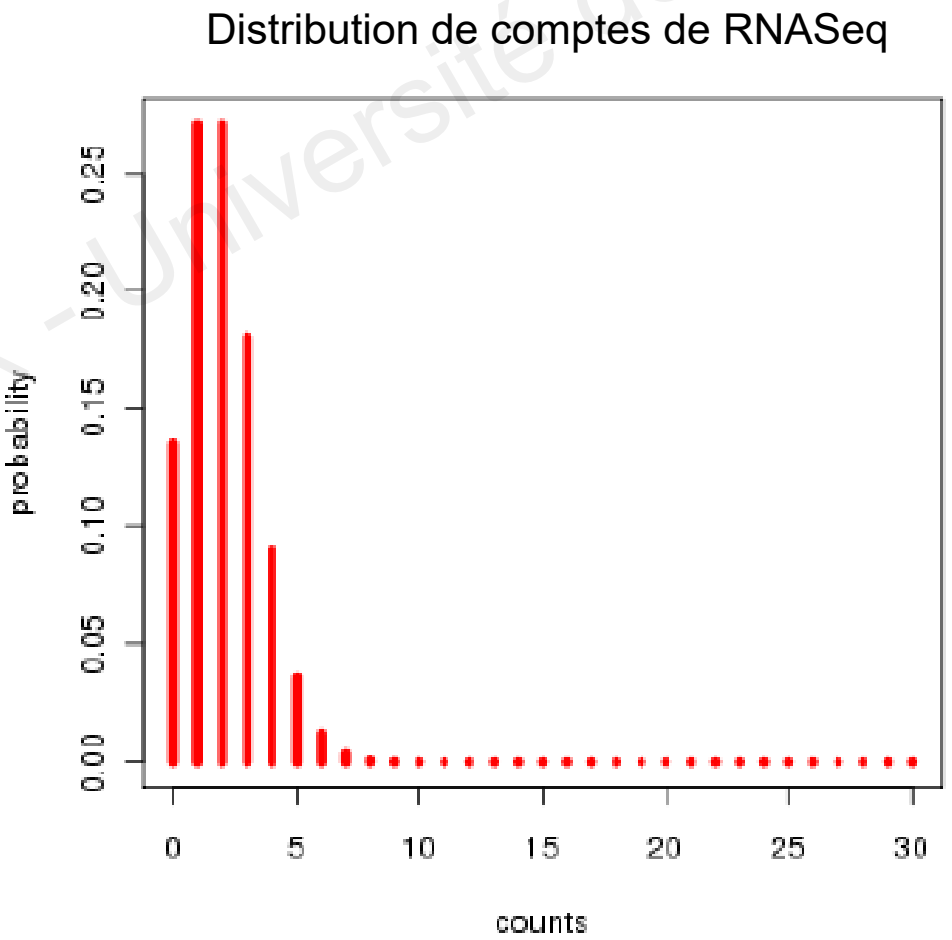
RNASeq

L'abondance des transcrits est mesurée par le nombre de lectures cartographies au niveau de la sequence génomique du transcrit

= comptes de lectures

↳ *Variable quantitative discrète*





=> Il faut utiliser la bonne loi de distribution (Poisson, Négative Binoimale…)

Estimating mean and variance in RNASeq experiments

In RNA-Seq, each feature (gene, exon, isoform) has an expression rate: each segment is sequenced with a low probability

Number of reads from gene g in library i can be captured by a Poisson model (Marioni et al. 2008)

$$r_{ij} \sim Poisson\ (\lambda_{ig} = \mu_{ig} k_{ig})$$

where

μ_{ig} is the concentration of the RNA

k_{ig} is a normalisation constant

$$\hat{\mu}_{ig} = \frac{r_{ig}}{k_{ig}}$$

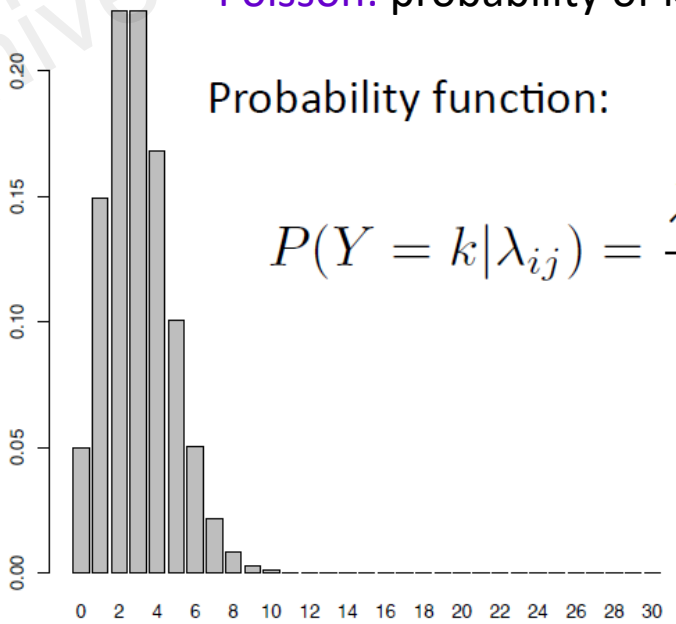
$$\lambda_{ig} = \mu_{ig} k_{ig} = E(r_{ij}) = Var(r_{ij})$$

↳ If n Xiid ~Poisson (λ), Σ Xi ~Poisson (nλ)

- Poisson: probability of k rare events

Probability function:

$$P(Y = k|\lambda_{ij}) = \frac{\lambda_{ij}^k e^{-\eta}}{k!}$$

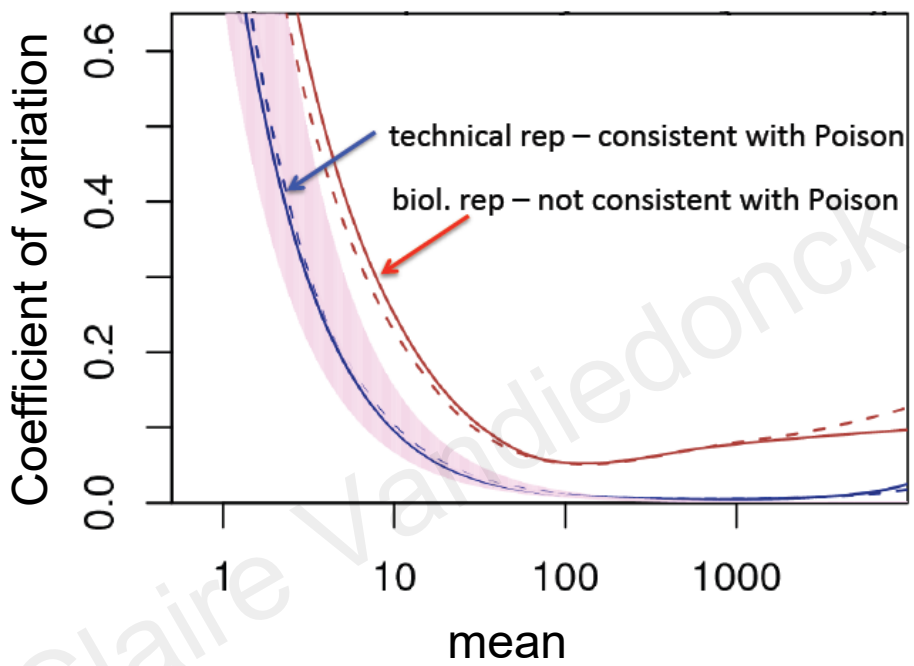


Need to account for extra variability

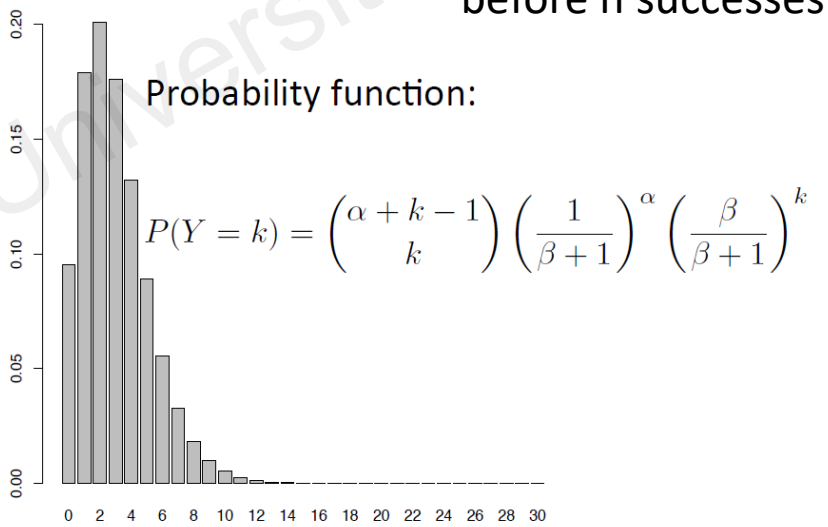
Poisson distribution accounts for technical variation

But biological noise induces an overdispersion

Convergence on a negative binomial model for count data



- Negative-binomial: probability of k failures before n successes



$$r_{ij} \sim NB\left(\alpha, \frac{1}{1+\beta}\right)$$

where α and β are the parameters of a gamma distribution followed by the rates of different samples

Modelling the variation

The example of DESeq and EdgeR

- generalized linear model fitting the negative binomial distribution:

$K_{ij} \sim NB(\mu_{ij}, \alpha_i)$

K_{ij} : counts of reads for gene i in sample j

α_i : gene-specific dispersion parameter

μ_{ij} : fitted mean

- $\mu_{ij} = s_j q_{ij}$

s_j : sample-specific size parameter

q_{ij} : a parameter proportional to the expected true concentration of fragments for sample j

- $\log_2(q_{ij}) = x_{j.} \beta_i$

β_i : the log2 fold change for gene i for each column (j.) of the model matrix X

1.6. The 3rd issue: reducing dimensionality

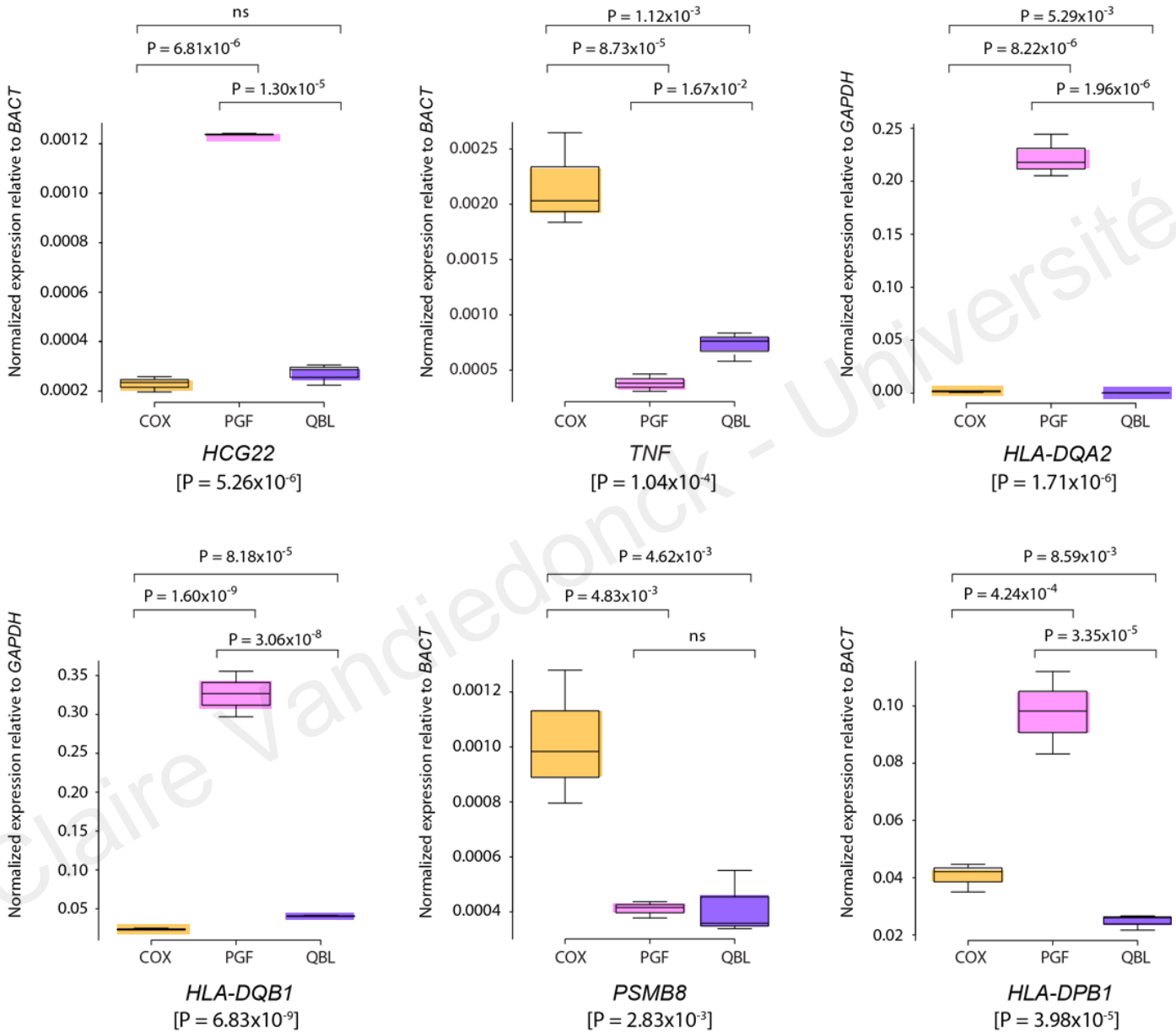
-> cf. next sessions 4 to 6

1.7. Results of a differential analysis

Table of results for differentially expressed (DE) genes

Gene Name	Class	log2 (Fold Change)			Adj.P.Val
		COX vs PGF	QBL vs PGF	QBL vs COX	
ZFP57	I	2.77	0.00	-2.76	1.22×10^{-14}
HLA-DPB2 *	II	-3.19	-3.02	0.17	2.89×10^{-12}
HLA-DQA2	II	-2.45	-1.62	0.82	1.91×10^{-11}
HLA-DQB2	II	-2.74	-2.58	0.16	3.21×10^{-11}
HLA-21 *	I	-2.52	0.36	2.87	1.32×10^{-10}
TNF	III	1.90	1.03	-0.87	4.79×10^{-10}
HLA-DPB1	II	-2.08	-0.90	1.18	6.44×10^{-10}
RPL32P1 *	II	-1.52	-1.19	0.33	2.07×10^{-09}
HLA-B	I	-0.06	-1.19	-1.13	6.59×10^{-09}
HLA-A	I	-1.51	-1.86	-0.35	2.30×10^{-08}
HLA-L *	I	-1.29	-1.47	-0.18	2.30×10^{-08}
XXbac-BPG254F23.6	II	-1.59	-1.59	0.00	2.50×10^{-08}
HCG22	I	-1.56	-1.26	0.30	2.96×10^{-08}
XXbac-BPG254F23.5	II	-1.42	-1.61	-0.19	1.33×10^{-07}
LTA	III	1.32	0.57	-0.75	2.04×10^{-07}
NCR3	III	0.87	0.95	0.08	4.95×10^{-07}
HLA-F	I	0.15	-0.90	-1.05	4.95×10^{-07}
HLA-DOA	II	-1.32	-0.89	0.43	5.07×10^{-07}
TAP1	II	0.97	0.08	-0.89	6.86×10^{-07}
LTB	III	-0.95	-0.06	0.89	7.02×10^{-07}
LST1	III	-0.18	0.48	0.66	9.42×10^{-07}
DAQB-335A13.8	I	0.61	-0.02	-0.63	1.12×10^{-06}
TCF19	I	1.11	0.62	-0.49	1.49×10^{-06}
CLIC1	III	1.22	0.57	-0.66	1.49×10^{-06}
HLA-DMA	II	-0.57	-0.89	-0.33	3.52×10^{-06}
BRD2	II	0.78	0.27	-0.51	3.60×10^{-06}
NRM	I	0.77	0.39	-0.38	4.48×10^{-06}
HLA-C	I	0.05	1.11	1.06	4.98×10^{-06}
PSMB9	II	0.42	-0.29	-0.71	6.05×10^{-06}
HCG27	I	0.56	0.06	-0.50	7.01×10^{-06}

Boxplots (or violoplots) for top DE genes



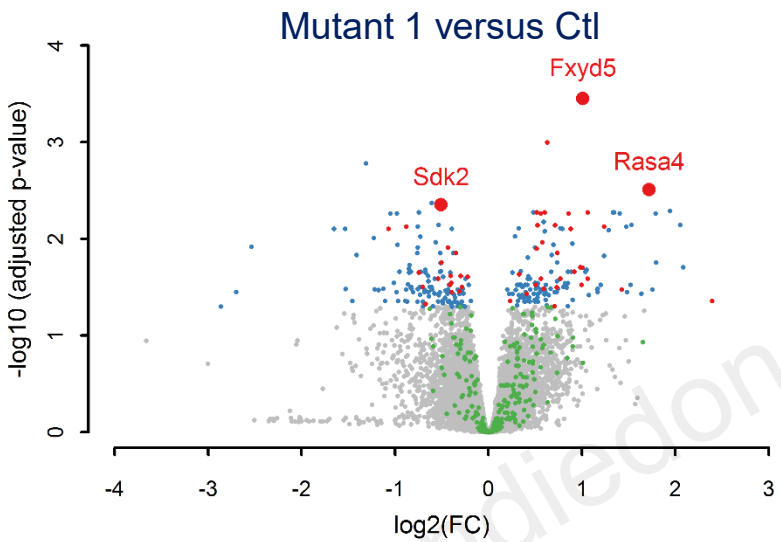
Graphical representation of DE genes

Volcano plots:

X = log2(Fold chnage)

Y = -log10 (pvalue)

- Exemple ici chez la souris avec 2 gènes KO versus Wild Type



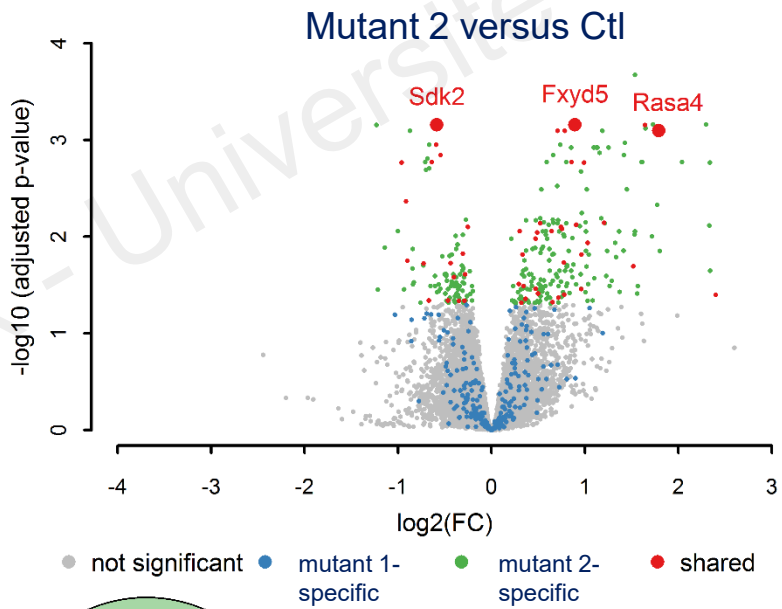
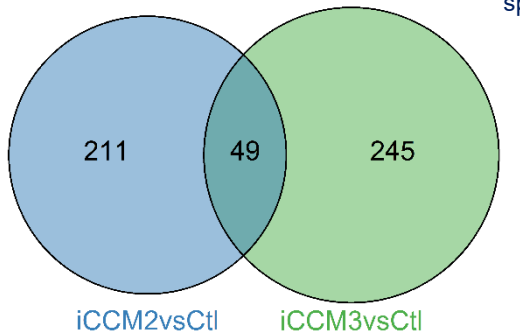


Diagramme de Venn

intersection des listes de gènes



1.8. Links

Nature series: http://www.nature.com/collections/qghhqm

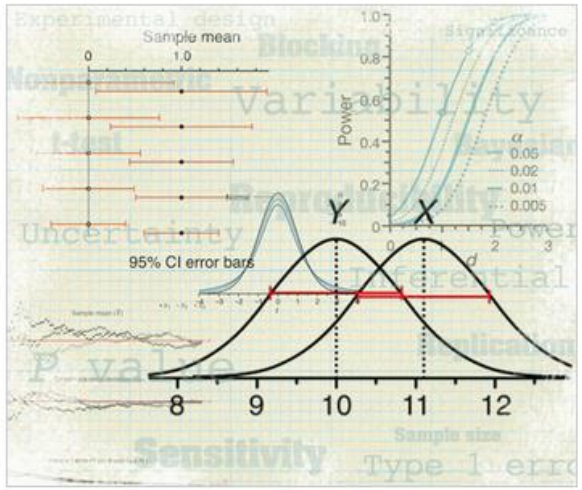
WEB COLLECTION

Search Go

Advanced search

Statistics for biologists

Home | Practical guides | Statistics in biology | Points of Significance | Other resources



There is no disputing the importance of statistical analysis in biological research, but too often it is considered only after an experiment is completed, when it may be too late.

This collection highlights important statistical issues that biologists should be aware of and provides practical advice to help them improve the rigor of their work.

Nature Methods' **Points of Significance** column on statistics explains many key statistical and experimental design concepts. **Other resources** include an online plotting tool and links to statistics guides from other publishers.

Image Credit: Erin DeWalt

Statistics in biology

Nature News | Editorial

Number crunch

Nature | Comments and Opinion

Research methods: Know when your numbers are significant

David L. Vaux

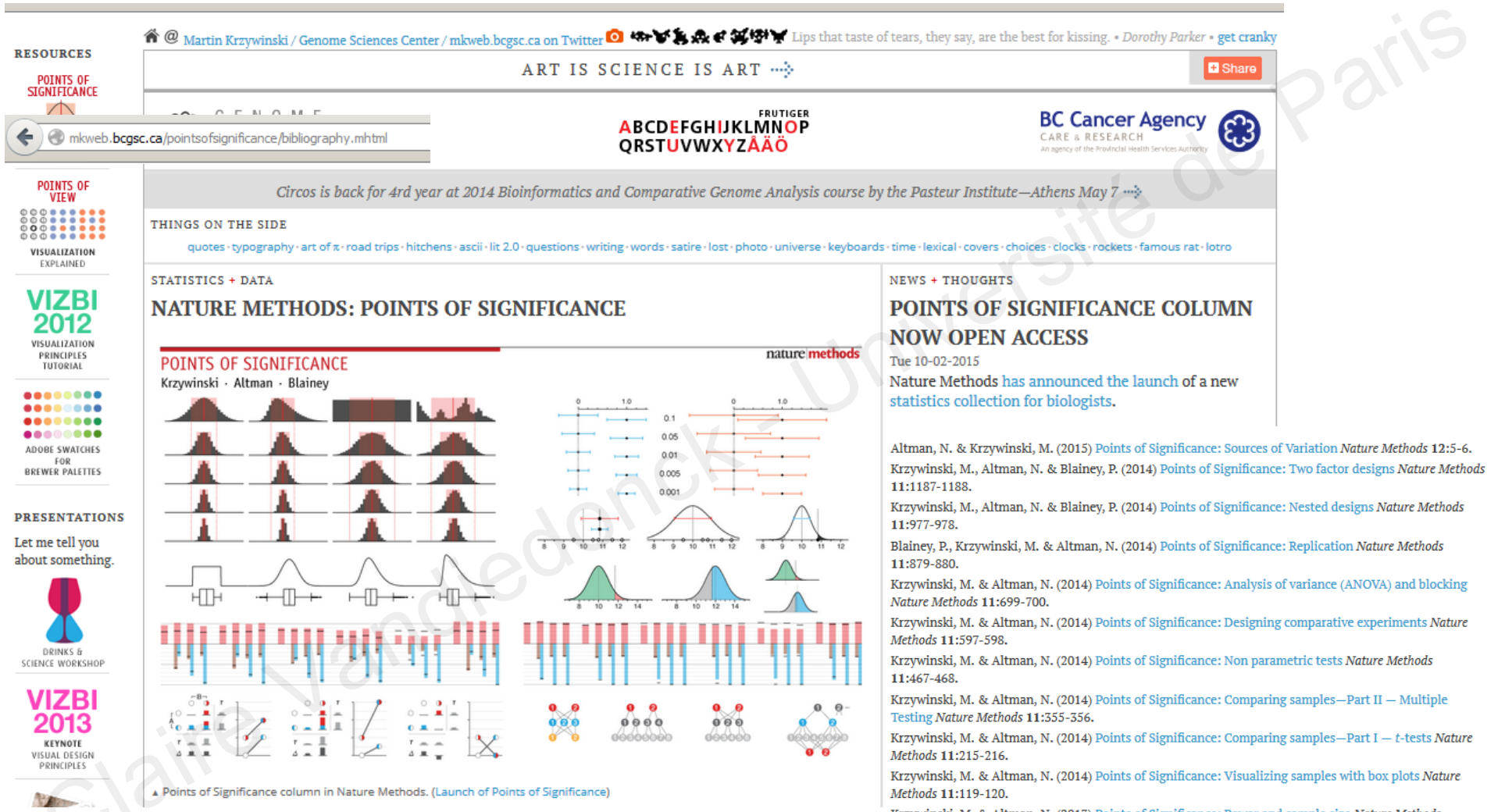
Top picks from nature news

Nature News | News

Scientific method: Statistical errors

Regina Nuzzo

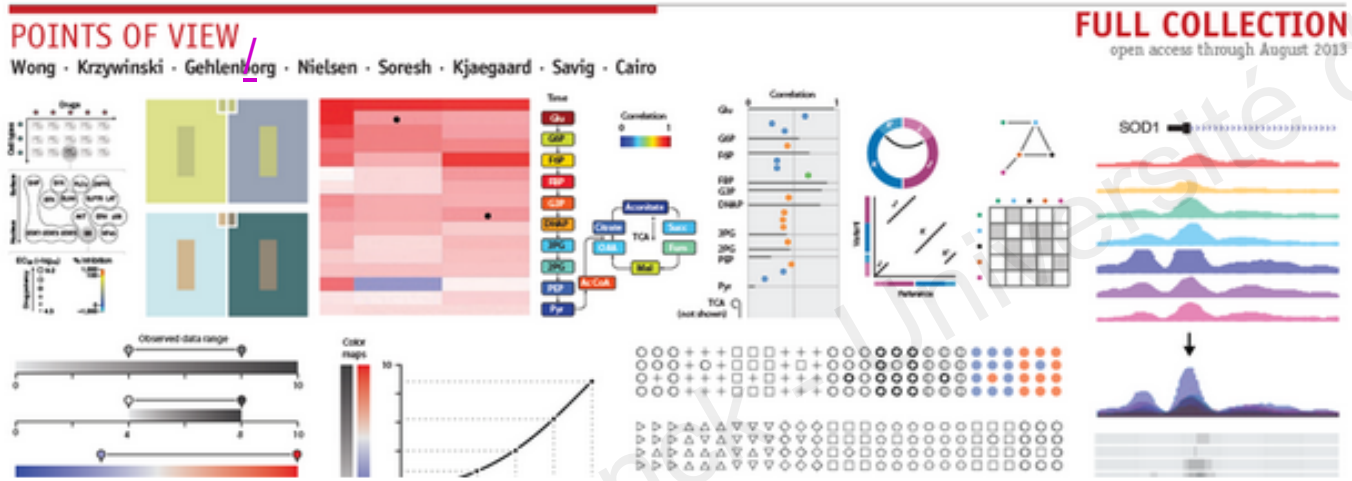
Points of significance: http://mkweb.bcgsc.ca/pointsofsignificance/



Points of view: http://mkweb.bcgsc.ca/pointsofview

COMMUNICATION + SCIENCE

NATURE METHODS: POINTS OF VIEW



▲ The full collection of a 35 Points of View column is now available. (3 years of Points of View)

PRACTICAL TIPS FOR EFFECTIVE FIGURES

POINTS OF VIEW — HISTORY

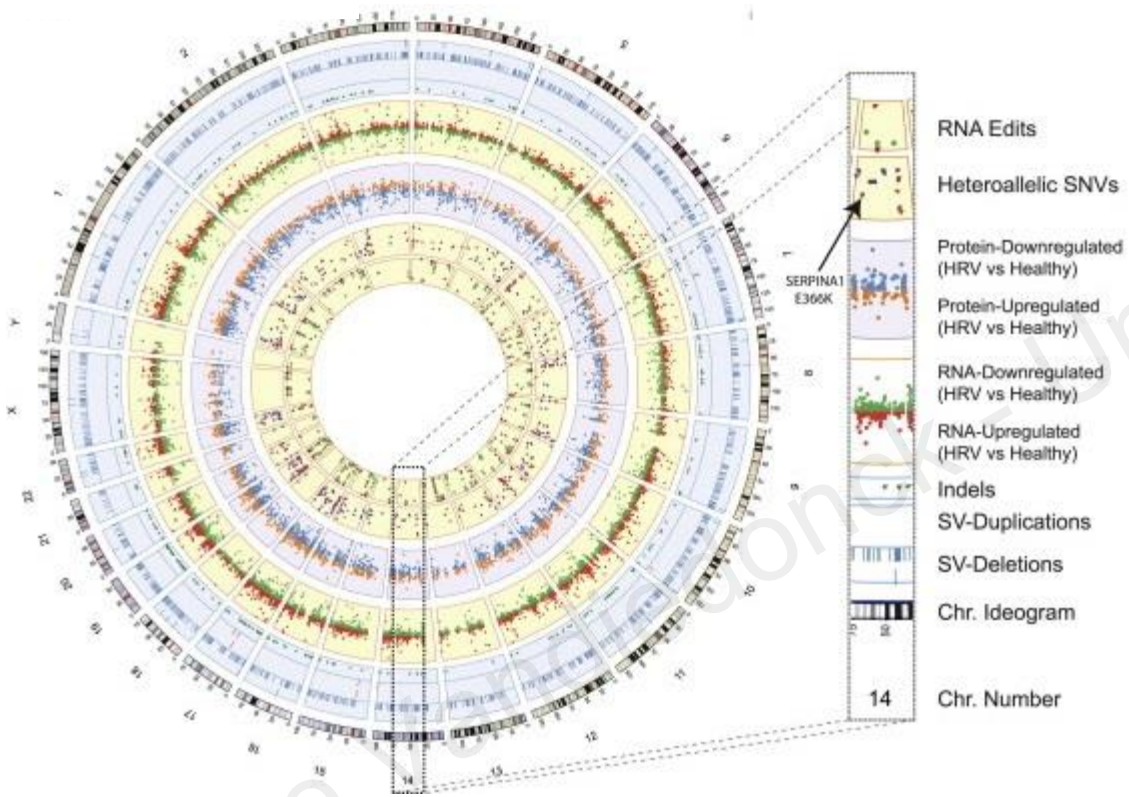
In its 2.5 year history, the PoV column has established a significant legacy— it is one of the most frequently accessed parts of Nature Methods. The reason I think is clear: the community sees the value in clear and effective visual communication and acknowledges the need for a forum in which best practices in the field are presented practically and accessibly.

Bang Wong, in collaboration with visiting authors (Noam Shoresh, Nils Gehlenborg, Cydney Nielsen and Rikke Schmidt Kjærgaard), has penned 29 columns in the period of August 2010 to December 2012, covering broad topics such as salience, Gestalt principles, color, typography, negative space, layout, and data integration.

When it was A.C. Greyling's turn to speak at a debate in which Christopher Hitchens and Richard Dawkins already made their points, Greyling said

Circos to represent genomic traits: http://circos.ca/intro/genomic_data/

Co-localisation

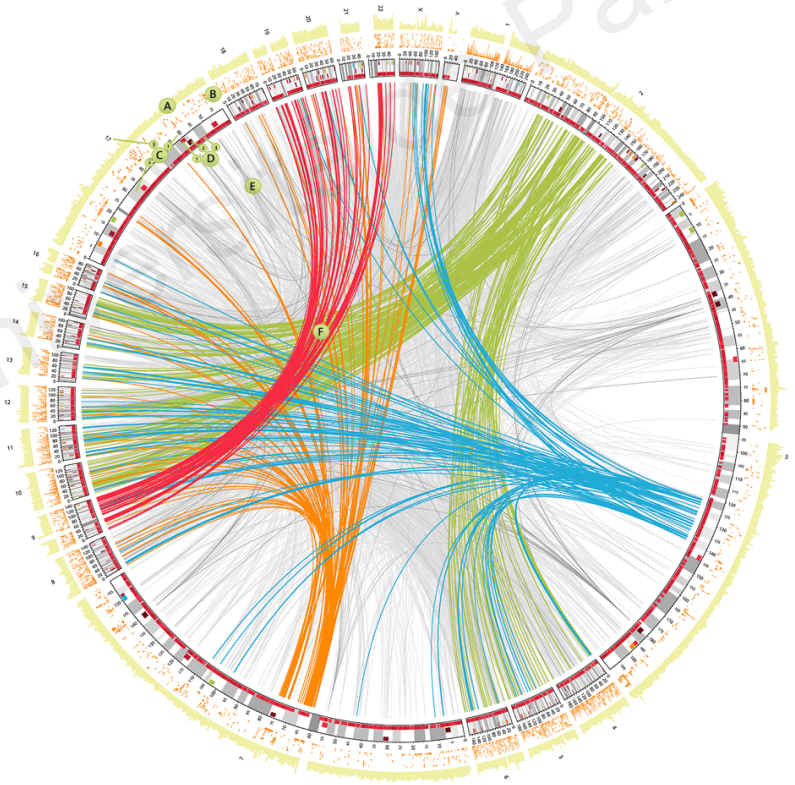


Cell 148, 1293–1307, March 16, 2012

Personal Omics Profiling Reveals Dynamic Molecular and Medical Phenotypes

Rui Chen,[1,11] George I. Mias,[1,11] Jennifer Li-Pook-Than,[1,11] Lihua Jiang,[1,11] Hugo Y.K. Lam,[1,12] Rong Chen,[2,12] Elana Miriami,[1] Konrad J. Karczewski,[1] Manoj Hariharan,[1] Frederick E. Dewey,[3] Yong Cheng,[1] Michael J. Clark,[1] Hogune Im,[1] Lukas Habegger,[6,7] Suganthi Balasubramanian,[6,7] Maeve O'Huallachain,[1] Joel T. Dudley,[2] Sara Hillenmeyer,[1] Rajini Haraksingh,[1] Donald Sharon,[1] Ghia Euskirchen,[1] Phil Lacroute,[1] Keith Bettinger,[1] Alan P. Boyle,[1] Maya Kasowski,[1] Fabian Grubert,[1] Scott Seki,[2] Marco Garcia,[2] Michelle Whirl-Carrillo,[1] Mercedes Gallardo,[9,10] Maria A. Blasco,[9] Peter L. Greenberg,[4] Phyllis Snyder,[1] Teri E. Klein,[1] Russ B. Altman,[1,5] Atul J. Butte,[2] Euan A. Ashley,[3] Mark Gerstein,[6,7,8] Kari C. Nadeau,[2] Hua Tang,[1] and Michael Snyder[1,*]

Interaction



Towards an increasing complexity of omics

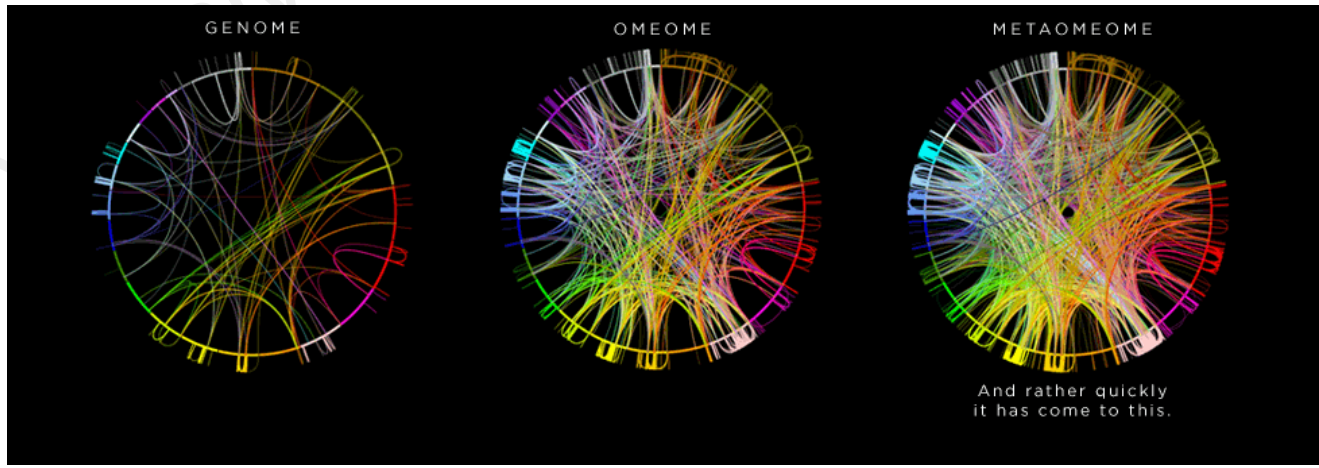
COMMENTARY *The Scientist* 15[7]:8, Apr. 2, 2001

'Ome Sweet 'Omics--
A Genealogical Treasury of Words

By Joshua Lederberg and Alexa T. McCray

antigenome	immunogenome	plastidome
bacteriome	immunome	plerome
basidiome	haptenome	proteinome
biome	karyome	proteome
cardiome	leptome	psychome
caulome	microbiome	regulome
chondriome	mnemome	rhabdome
cladome	mycetome	rhizome
coelome	neurome	stereome
epigenome	odontome	thallome
erythrome	osteome	tracheome
genome	pharmacogenome	transcriptome
geome	phenome	trichome
hadrome	phyllome	vacuome
histome	physiome	

Now!



Genomics and *Proteomics* are the buzzwords of the dawning millennium. There is no counting of www.-ics.com and www.-ix.com sites to be found on the Web. That most of these terms, old and new, have been contrived as slogans to attract attention, does not diminish their likely substance, and they are embedded in the advancing edge of science and technology. Defying the French Linguists' caveat, we may yet ask, where do terms such as *genome* and

https://lhncbc.nlm.nih.gov/system/files/pub2001047.pdf

2. RStudio et Rmarkdown

a. A live session!

b. Optional: A Rmd practical on statistics for omics data