

11 juillet 2019

STATISTIQUES

comprendre les notions de bases pour bien choisir son test



Institute for Advanced Biosciences

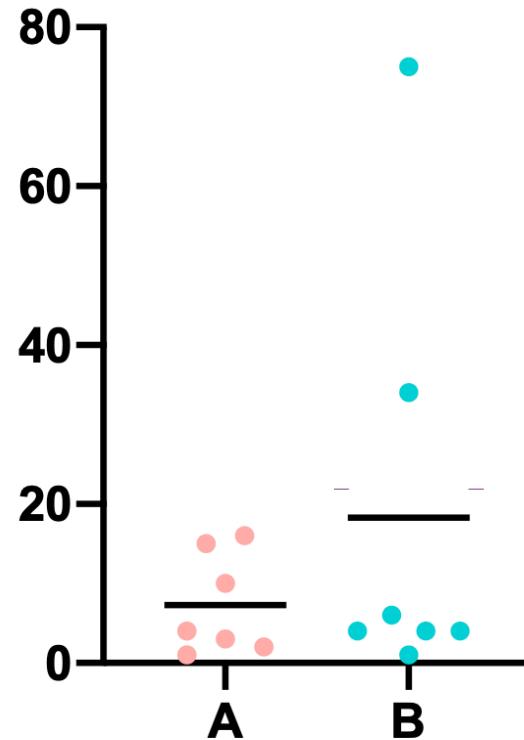
Définitions

Objectif :

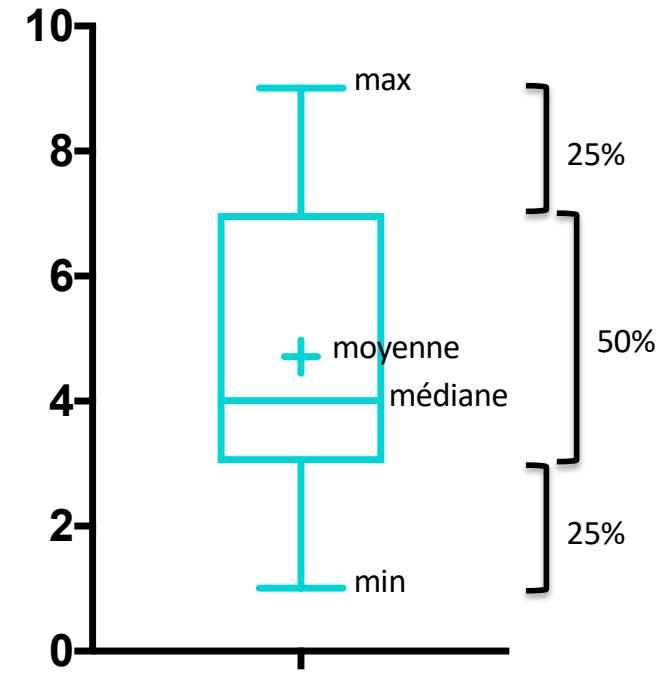
Est-ce que 2 groupes d'échantillons sont différent l'un de l'autre ?

Est-ce que 2 facteurs sont corrélés ?

Visualisation des données :



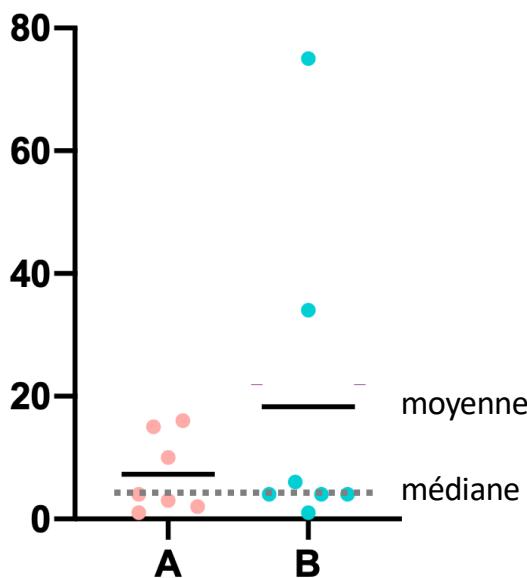
Scatter plot



Boite à moustache

Définitions

Que veut on comparer ? Moyennes ou Médianes ?

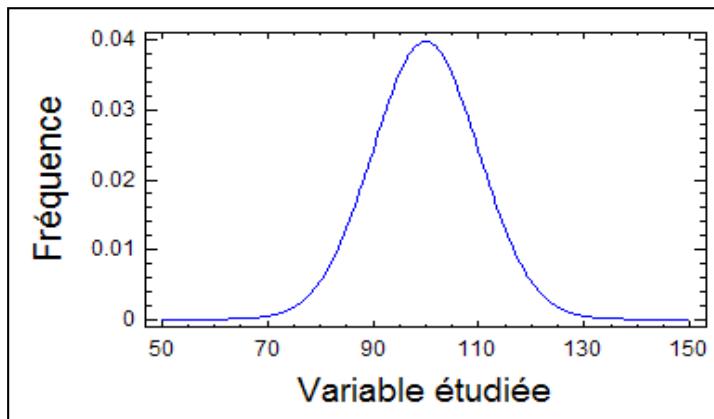
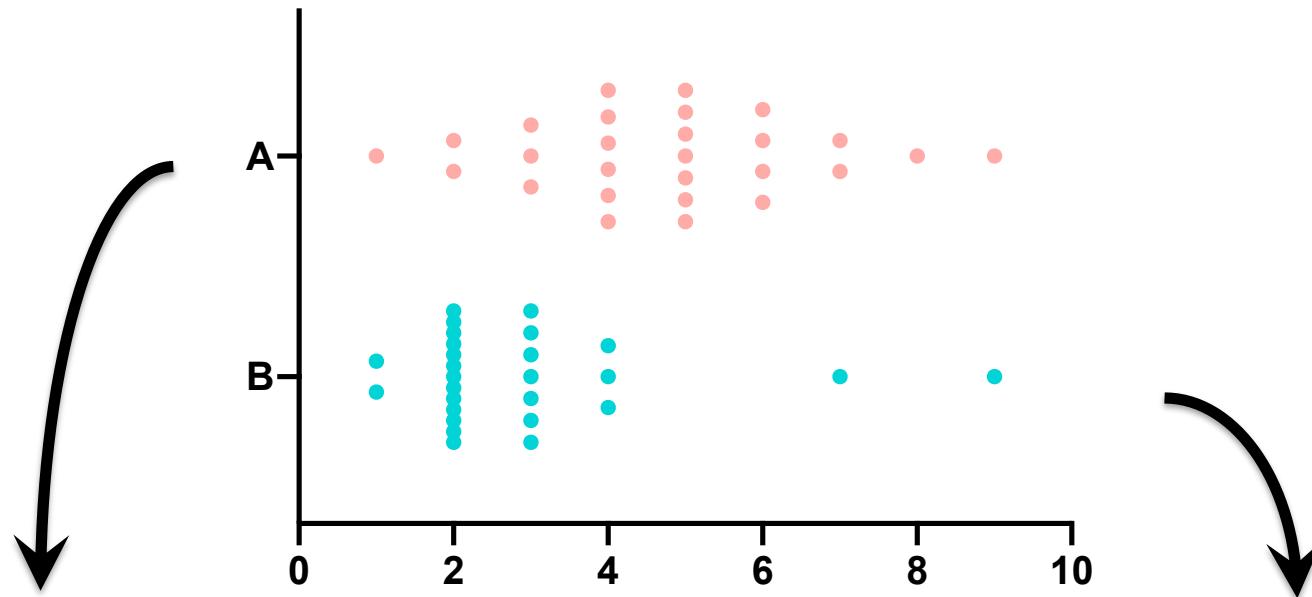


A	B
1	1
2	4
3	4
4	4
10	5
15	37
16	74
7,3	18,4

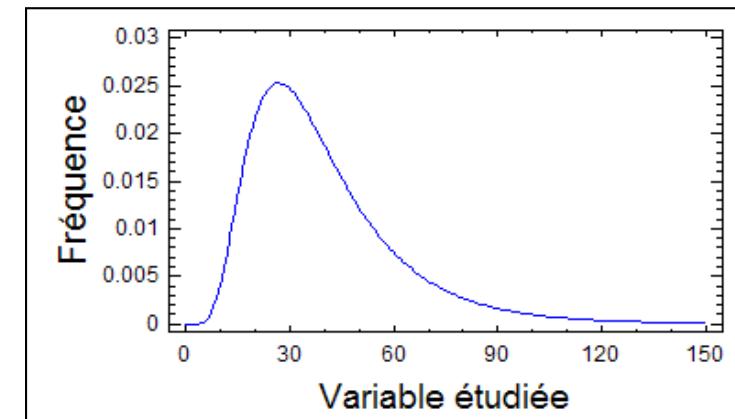
médiane

moyenne

La distribution



Normalité



Non Normalité

La distribution

Test de normalité : test de Shapiro wilk

The image shows two overlapping dialog boxes from SPSS:

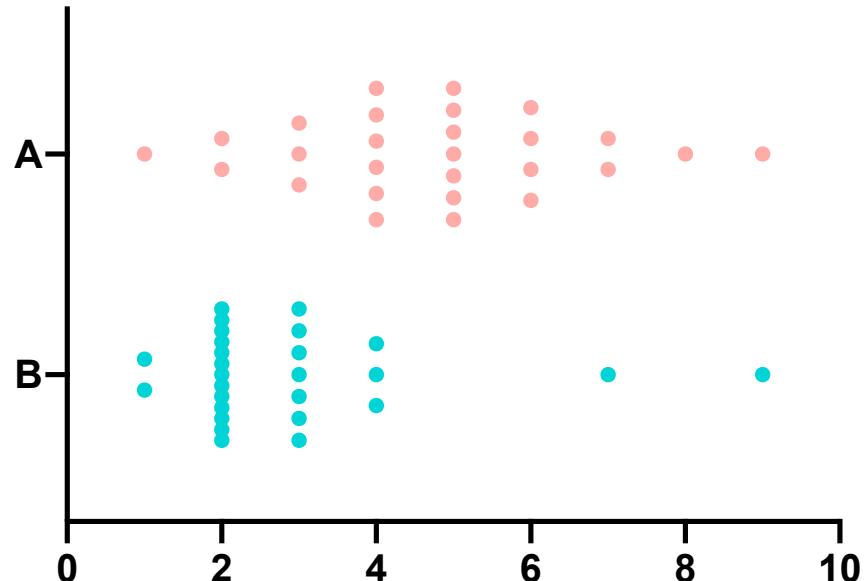
Analyze Data Dialog (Left):

- Category: Built-in analysis
- Section: Which analysis?
- Transform, Normalize... (checkbox)
- XY analyses (checkbox)
- Column analyses (checkbox):
 - t tests (and nonparametric tests) (checkbox)
 - One-way ANOVA (and nonparametric) (checkbox)
 - Column statistics** (checkbox, highlighted with a red oval)
 - Frequency distribution
 - ROC Curve
 - Bland-Altman method comparison
 - Correlation
 - Identify outliers
- Grouped analyses (checkbox)
- Contingency table analyses (checkbox)
- Survival analyses (checkbox)
- Parts of whole analyses (checkbox)
- Generate curve (checkbox)

Parameters: Column Statistics Dialog (Right):

- Section: Descriptive Statistics
 - Minimum and maximum
 - Quartiles (Median, 25th and 75th percentile)
 - Percentile 90.0
 - Mean, SD, SEM
 - Coefficient of variation
 - Geometric mean
 - Skewness and kurtosis
 - Column sum
- Section: Confidence intervals
 - CI of the mean
 - CI of geometric mean
 - CI of median
- Section: Test if the values come from a Gaussian distribution (highlighted with a red box)
 - D'Agostino-Pearson omnibus normality test (recommended)
 - Shapiro-Wilk normality test (highlighted with a red circle)
 - Kolmogorov-Smirnov test with Dallal-Wilkinson-Lilliefors P value (not recommended)
- Section: Inferences
 - One-sample t test. Are column means significantly different than a hypothetical value? Hypothetical value: 0.0
 - Wilcoxon signed-rank test. Compare column medians to a hypothetical value.
- Section: Calculations
 - Subcolumns:
 - Compute the mean of the subcolumns for each row, and then calculate the column statistic for each row
 - Compute column statistics for each subcolumn separately
 - Show: 4 significant digits
 - Make these choices be the default for future analyses.

La distribution



Normality and Lognormality Tests		A	B
Tabular results		A	B
		Y	Y
Test for normal distribution			
Shapiro-Wilk test			
W		0.9720	0.6993
P value		0.6543	<0.0001
Passed normality test (alpha=0.05)?		Yes	No
P value summary		ns	****

* Minimum n=7

Normalité

=

Comparaison de moyennes
Test paramétrique

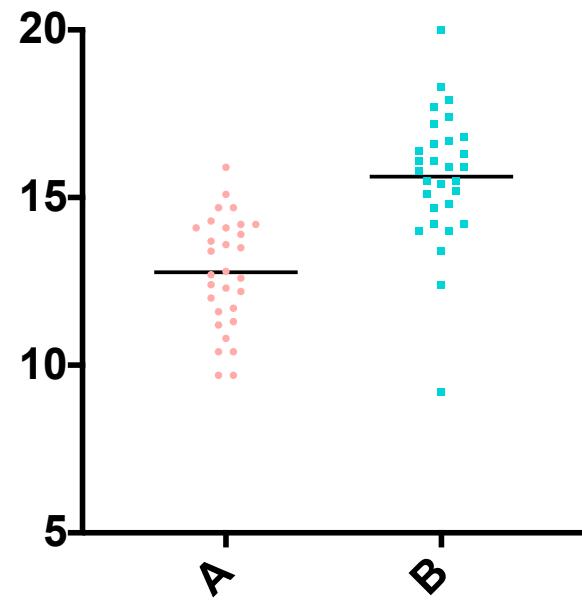
Non Normalité

=

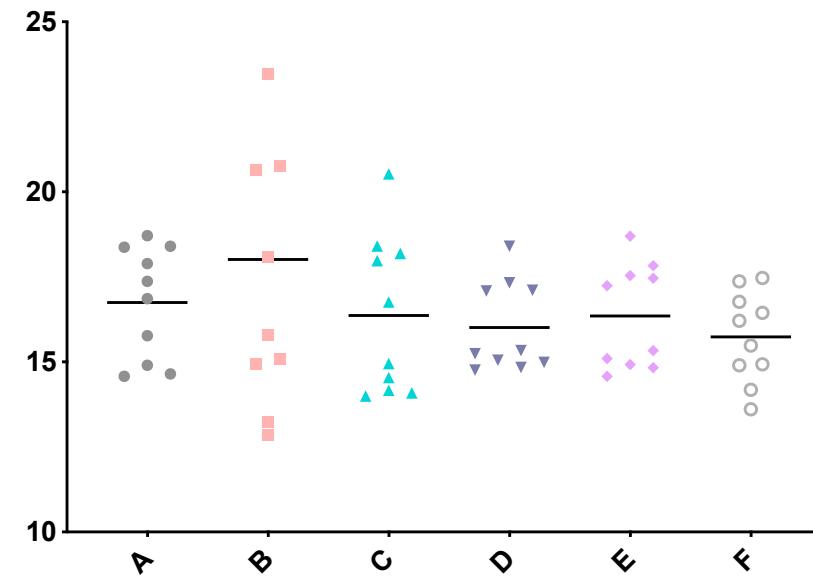
Comparaison de médIANES
Test non paramétrique

Tests de comparaison de moyennes

2 moyennes



A partir de 3 moyennes



T test
(test de student)

ANOVA

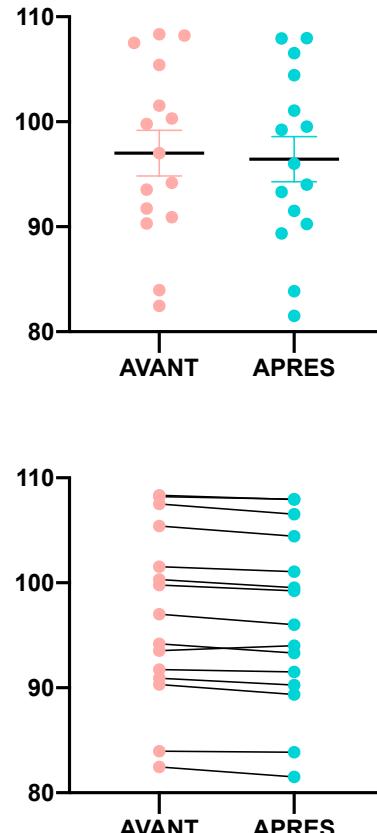
Tests de comparaison de moyennes

Données appariées – pairées :
Valeurs prises 2 fois sur les mêmes individus

Exemple :

Comparaison de la taille d'une tumeur avant/après traitement

AVANT TRAITEMENT	APRES TRAITEMENT
82	82
84	84
90	89
91	90
92	92
94	94
94	93
97	96
100	99
100	100
102	101
105	104
108	107
108	108
108	108
97	96



Unpaired t test

P value

P value summary

Significantly different ($P < 0.05$)?

0,8512
ns
No

Paired t test

P value

P value summary

Significantly different ($P < 0.05$)?

0,0001

Yes

Comparaison multiple

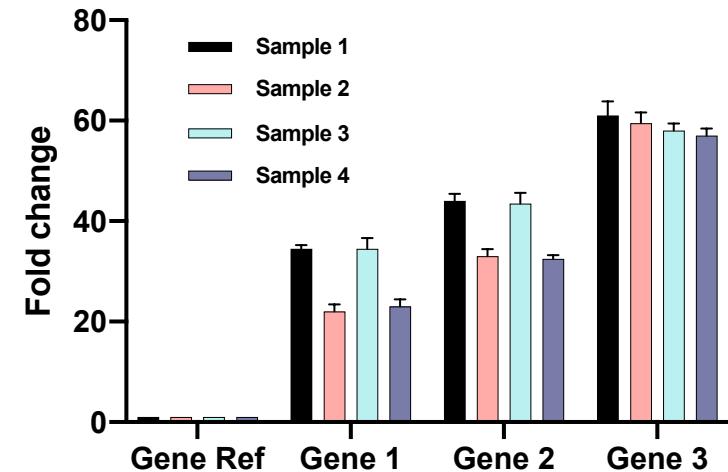
A partir de 3 moyennes à comparer : test ANOVA

Exemple :

Comparaison d'expression de gènes en qPCR

Table format: Grouped		Group A		Group B		Group C		Group D	
		sample 1	Sample 2	sample 3	Sample 4				
1	Gene Ref	1	1	1	1	1	1	1	1
2	Gene 1	34	35	23	21	33	36	22	24
3	Gene 2	43	45	34	32	45	42	33	32
4	Gene 3	59	63	61	58	57	59	56	58

2way ANOVA ANOVA results				
1	Table Analyzed	Data 1		
2				
3	Two-way ANOVA	Ordinary		
4	Alpha	0.05		
5				
6	Source of Variation	% of total variation	P value	P value summary
7	Interaction	1.742	<0.0001	****
				Significant?
				Yes



Au moins un des échantillons est différent des autres

Comparaison multiple

Quel échantillon est différent ?

RM Design RM Analysis Factor Names Multiple Comparisons Options Residuals

What kind of comparison?

Within each row, compare columns (simple effects within rows) ◀ ▶

échantillons

		Group A		Group B		Group C	
		Data Set-A		Data Set-B		Data Set-C	
		A:Y1	A:Y2	B:Y1	B:Y2	C:Y1	C:Y2
gènes	1	Mean		Mean		Mean	
	2	Mean		Mean		Mean	
	3	Mean		Mean		Mean	

The diagram shows a 3x6 grid of cells representing data sets. The columns are labeled Group A, Group B, and Group C, and the rows are labeled 1, 2, and 3. Each cell contains a 'Mean' value, which is highlighted with a colored oval: Row 1 has green ovals, Row 2 has blue ovals, and Row 3 has orange ovals. Horizontal arrows connect the means within each row, indicating comparisons between cell means within a group.

How many comparisons?

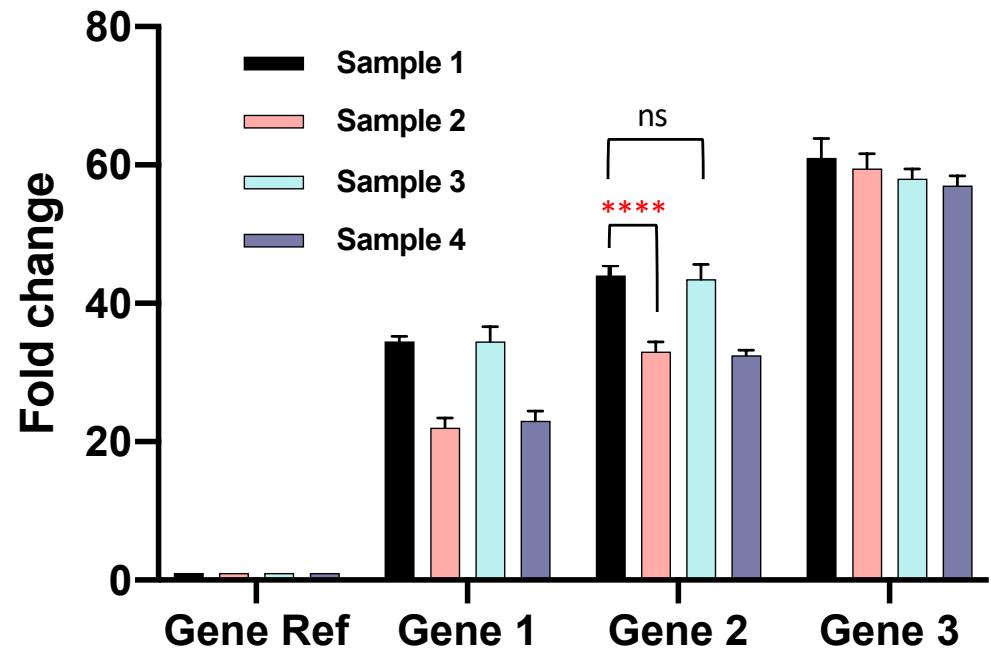
Compare each cell mean with every other cell mean on that row.
 Compare each cell mean with the control cell mean on that row.

Control column: Group A: Sample 1 ◀ ▶

Comparaison multiple

Quel échantillon est différent ?

Tukey's multiple comparisons test	Mean Diff.	Significant?	Summary	Adjusted P Value
Gene 1				
Sample 1 vs. Sample 2	12,5	Yes	****	<0,0001
Sample 1 vs. Sample 3	0	No	ns	>0,9999
Sample 1 vs. Sample 4	11,5	Yes	****	<0,0001
Sample 2 vs. Sample 3	-12,5	Yes	****	<0,0001
Sample 2 vs. Sample 4	-1	No	ns	0,9028
Sample 3 vs. Sample 4	11,5	Yes	****	<0,0001
Gene 2				
Sample 1 vs. Sample 2	11	Yes	****	<0,0001
Sample 1 vs. Sample 3	0,5	No	ns	0,9859
Sample 1 vs. Sample 4	11,5	Yes	****	<0,0001
Sample 2 vs. Sample 3	-10,5	Yes	****	<0,0001
Sample 2 vs. Sample 4	0,5	No	ns	0,9859
Sample 3 vs. Sample 4	11	Yes	****	<0,0001
Gene 3				
Sample 1 vs. Sample 2	1,5	No	ns	0,7396
Sample 1 vs. Sample 3	3	No	ns	0,2138
Sample 1 vs. Sample 4	4	No	ns	0,0649
Sample 2 vs. Sample 3	1,5	No	ns	0,7396
Sample 2 vs. Sample 4	2,5	No	ns	0,3545
Sample 3 vs. Sample 4	1	No	ns	0,9028



Comparaison multiple

Quel gène est différent ?

RM Design RM Analysis Factor Names Multiple Comparisons Options Residuals

What kind of comparison?

Within each column, compare rows (simple effects within columns) échantillons

gènes

	Group A		Group B		Group C	
	Data Set-A		Data Set-B		Data Set-C	
	A:Y1	A:Y2	B:Y1	B:Y2	C:Y1	C:Y2
1						
2						
3						

How many comparisons?

Compare each cell mean with every other cell mean on that column.

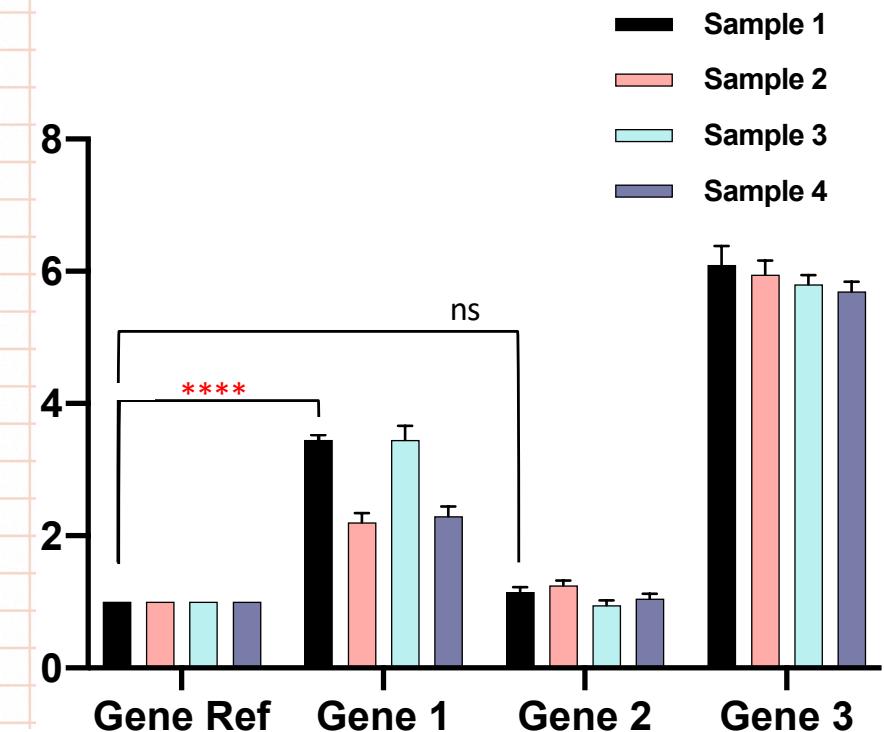
Compare each cell mean with the control cell mean on that column.

Control row: Row 1: Gene Ref

Comparaison multiple

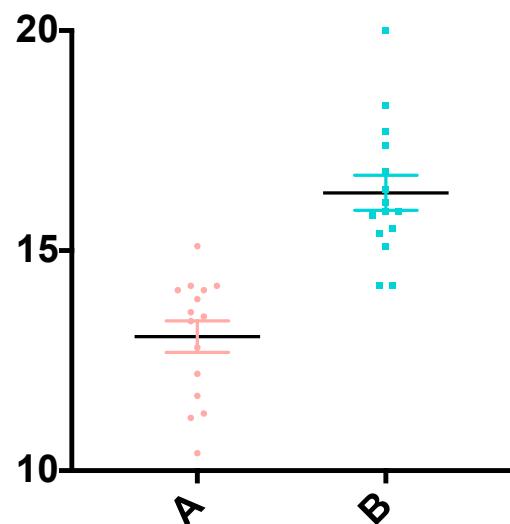
Quel gène est différent ?

Dunnett's multiple comparisons test	Mean Diff,	Significant?	Summary	Adjusted P Value
Sample 1				
Gene Ref vs. Gene 1	-2,45	Yes	****	<0,0001
Gene Ref vs. Gene 2	-0,15	No	ns	0,5403
Gene Ref vs. Gene 3	-5,1	Yes	****	<0,0001
Sample 2				
Gene Ref vs. Gene 1	-1,2	Yes	****	<0,0001
Gene Ref vs. Gene 2	-0,25	No	ns	0,176
Gene Ref vs. Gene 3	-4,95	Yes	****	<0,0001
Sample 3				
Gene Ref vs. Gene 1	-2,45	Yes	****	<0,0001
Gene Ref vs. Gene 2	0,05	No	ns	0,9633
Gene Ref vs. Gene 3	-4,8	Yes	****	<0,0001
Sample 4				
Gene Ref vs. Gene 1	-1,3	Yes	****	<0,0001
Gene Ref vs. Gene 2	-0,05	No	ns	0,9633
Gene Ref vs. Gene 3	-4,7	Yes	****	<0,0001



Interprétation du test

A	B
13,5	15,5
14,1	17,4
12,8	15,9
13,4	20
11,7	15,4
11,3	15,8
15,1	14,2
11,2	15,9
13,6	14,2
14,1	16,4
13,9	18,3
10,4	15,1
14,2	16,1
12,2	17,7
14,2	16,8
13	16



Test d'hypothèse nulle H_0
=

On suppose l'égalité

Unpaired t test	
P value	<0.0001
P value summary	****
Significantly different ($P < 0.05$)?	Yes

P-value :

Probabilité que les moyennes comparées soient égales

Risque α :

Seuil de la p-value - Risque que l'on accepte de prendre de se tromper

Souvent le risque α est de 5% :

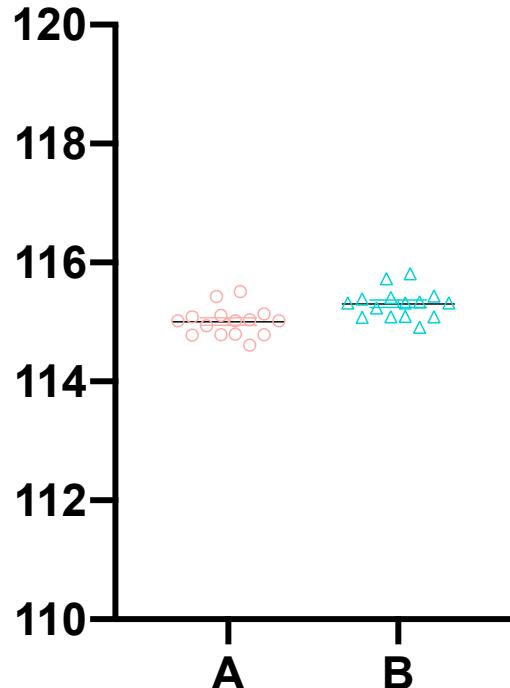
J'accepte de me tromper 5 fois sur 100 en affirmant que la différence n'est pas due au hasard

α va permettre d'interpréter la p-value et de prendre une décision :

Si p-value $< \alpha$: on rejette l'hypothèse que les moyennes soient égales

on parle de **différence significative**

Interprétation du test



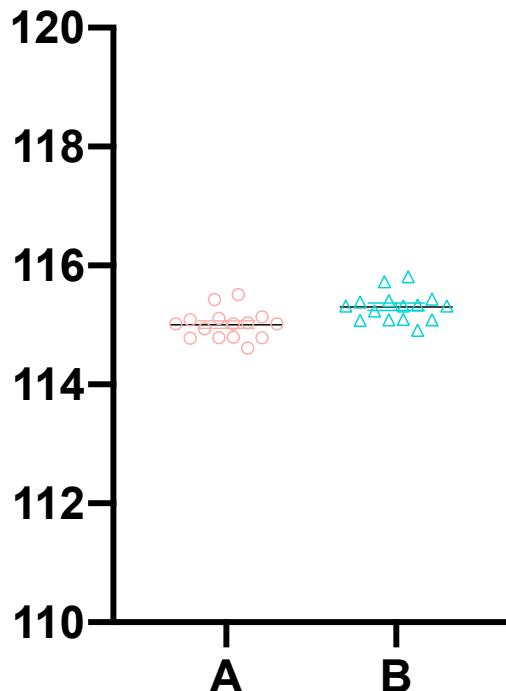
Unpaired t test	
P value	<0.0001
P value summary	****
Significantly different ($P < 0.05$)?	Yes

Nous pouvons conclure que les moyennes sont très différentes

Nous pouvons conclure que les moyennes sont très probablement différentes

How big is the difference ?

Interprétation du test



Unpaired t test	
P value	<0.0001
P value summary	****
Significantly different ($P < 0.05$)?	Yes

How big is the difference?	
Mean of column A	13.05
Mean of column B	16.31
Difference between means ($B - A$) \pm SEM	3.267 ± 0.5343
95% confidence interval	2.172 to 4.361
R squared (eta squared)	0.5717

SEM = Standard Error of Mean
IC à 95% = Intervalle de confiance

Interprétation du test

SEM = Standard Error of Mean

Souvent confondu avec l'Ecart type
(qui représente la variabilité/dispersion des données)

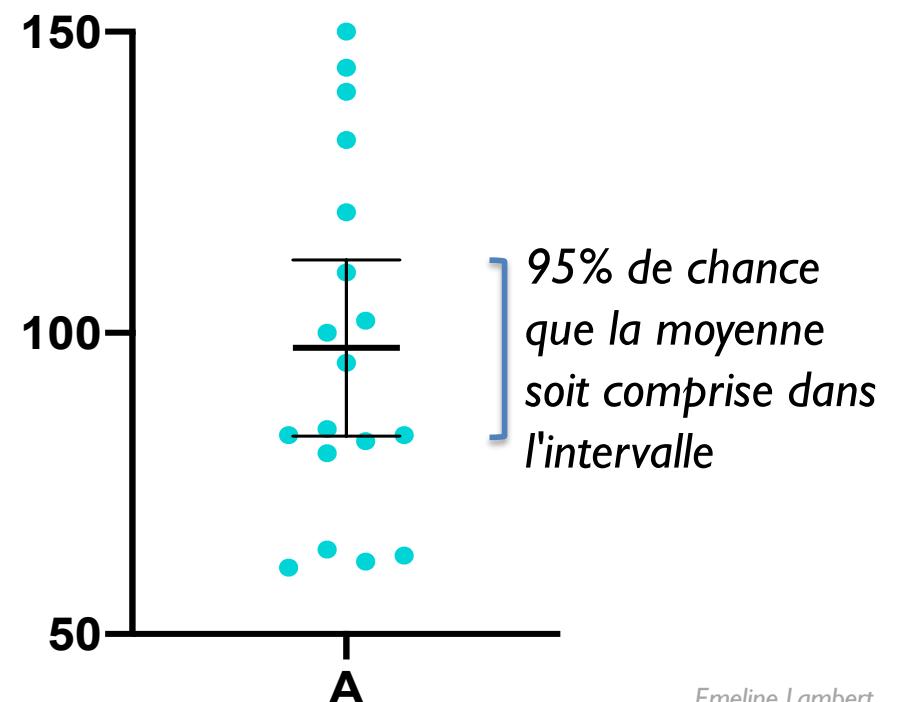
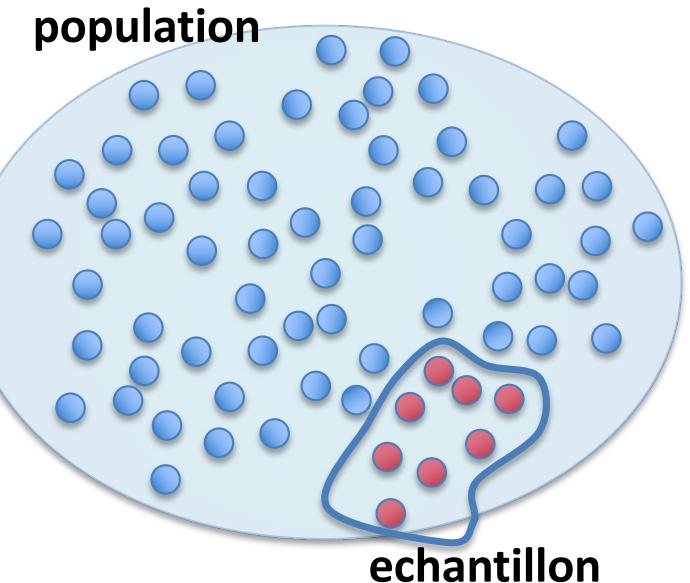
$$SEM = \frac{\text{ecart - type}}{\sqrt{\text{effectif}}}$$

IC 95% = Intervalle de Confiance à 95%

IC 95% = Moyenne +/- 2 SEM

Il y a 95% de chance que la moyenne de population soit comprise dans l'intervalle :

Moyenne d'échantillon +/- 2 SEM



Conclusion du test

Unpaired t test	
P value	<0.0001
P value summary	****
Significantly different ($P < 0.05$)?	Yes
One- or two-tailed P value?	Two-tailed
t, df	t=6.114, df=28
How big is the difference?	
Mean of column A	13.05
Mean of column B	16.31
Difference between means (B - A)	3.267 ± 0.5343
95% confidence interval	2.172 to 4.361

Je peux conclure qu'il y a très certainement une différence

Mais je ne peux pas conclure si cette différence est grande ou pas

Je peux conclure que la différence entre les moyennes est sans doute comprise dans l'intervalle de confiance

Arbre des tests

