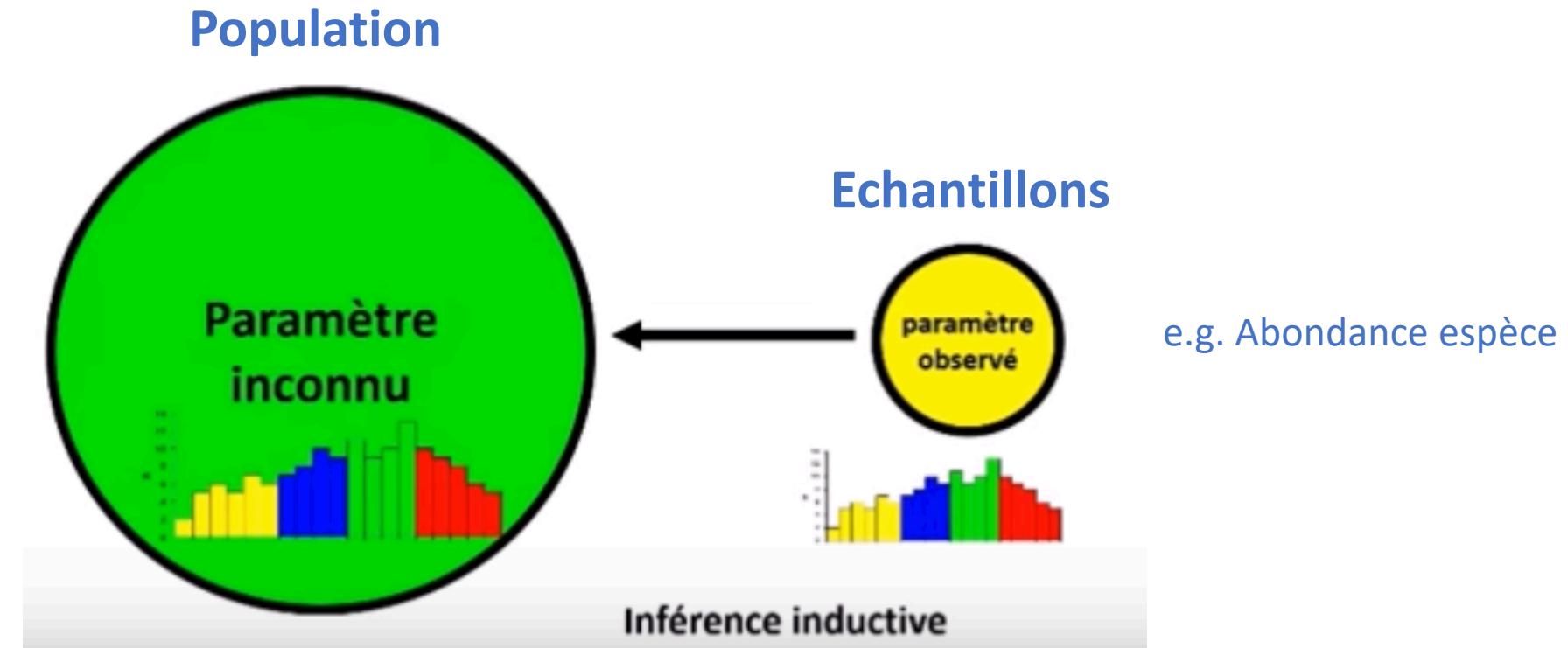


Problématique : Qu'une seule chose à retenir!

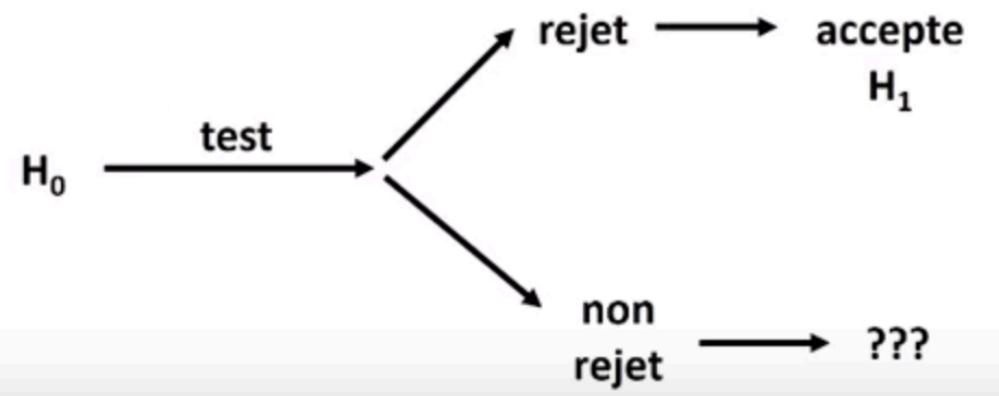
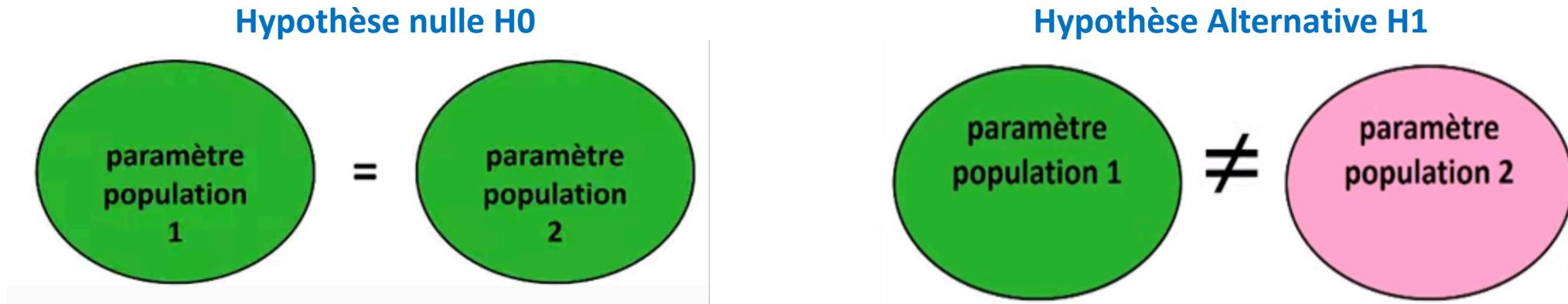
→ Hypothèses/inférences sur une **population** à partir de ce que l'on voit sur des **échantillons**



Pb des hypothèses/inférences: Soumises aux erreurs!!
Risque → Cas des résultats d'un test statistique (H_0)

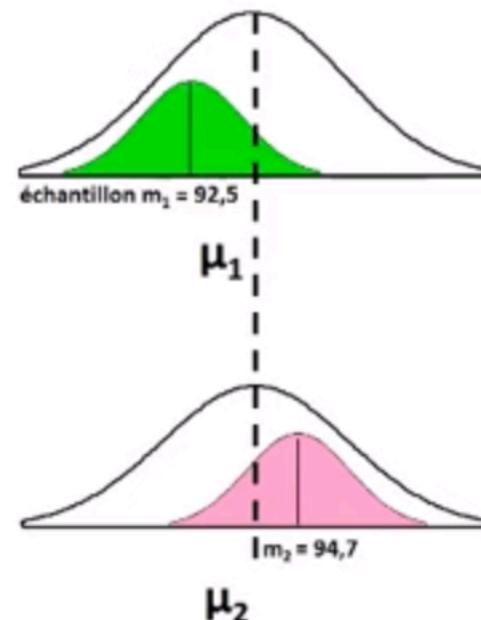
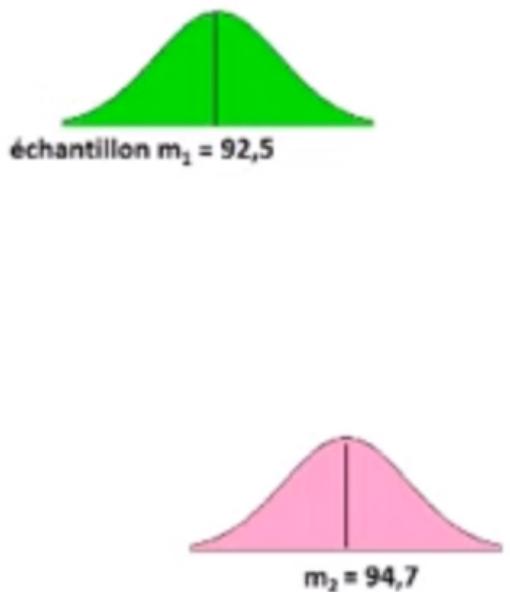
Comparaison de moyennes et Hypothèses

Question: Y a-t-il une vraie différence ... ou est ce le fruit du hasard?



Comparaison de moyennes et Hypothèses

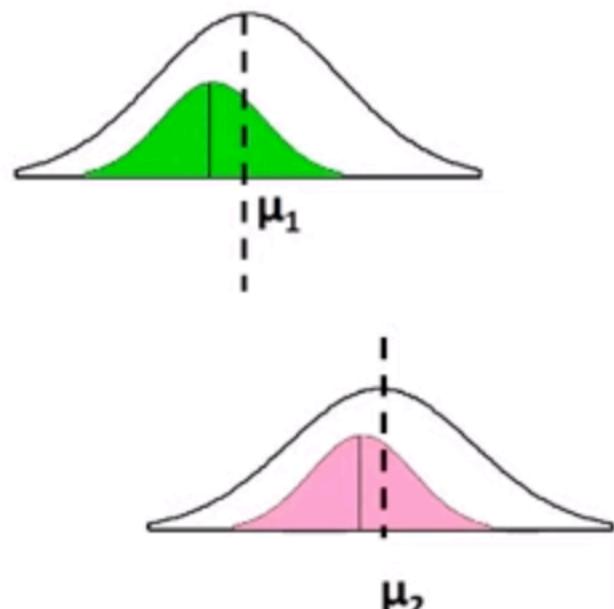
Si H₀ est vraie ... pas de différence



$$H_0: \mu_1 = \mu_2$$

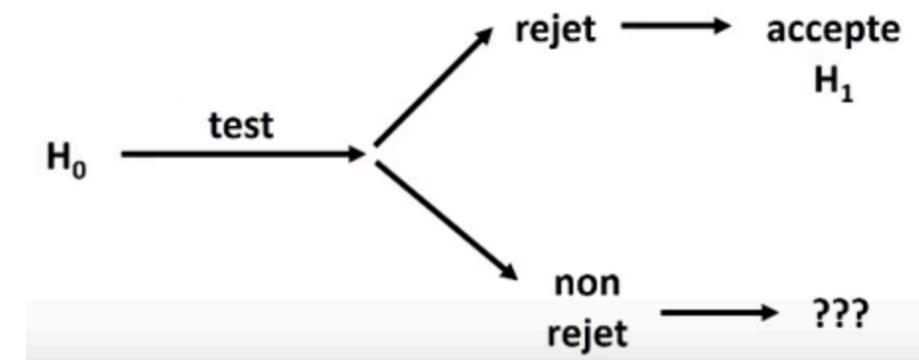
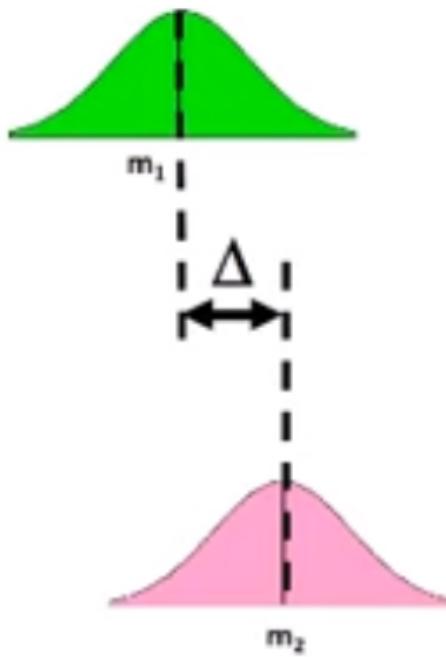
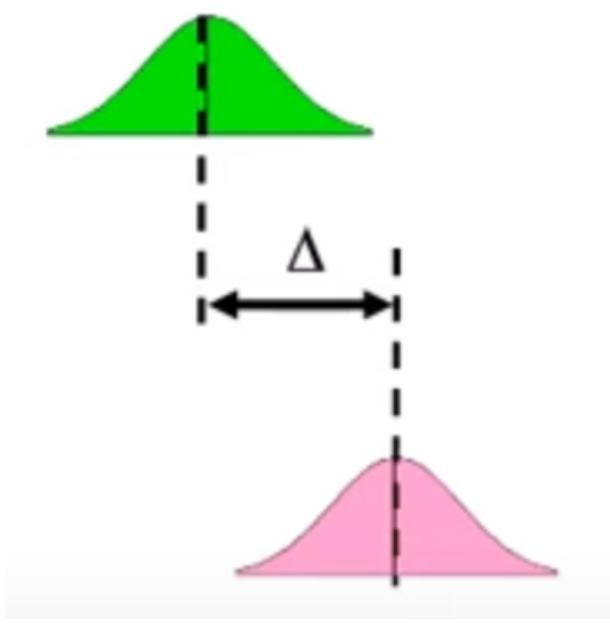
Même distribution
→ Fluctuation échantillonnage

Si H₀ rejeté, H₁ acceptée



$$H_1: \mu_1 \neq \mu_2$$

Deux distributions différentes!

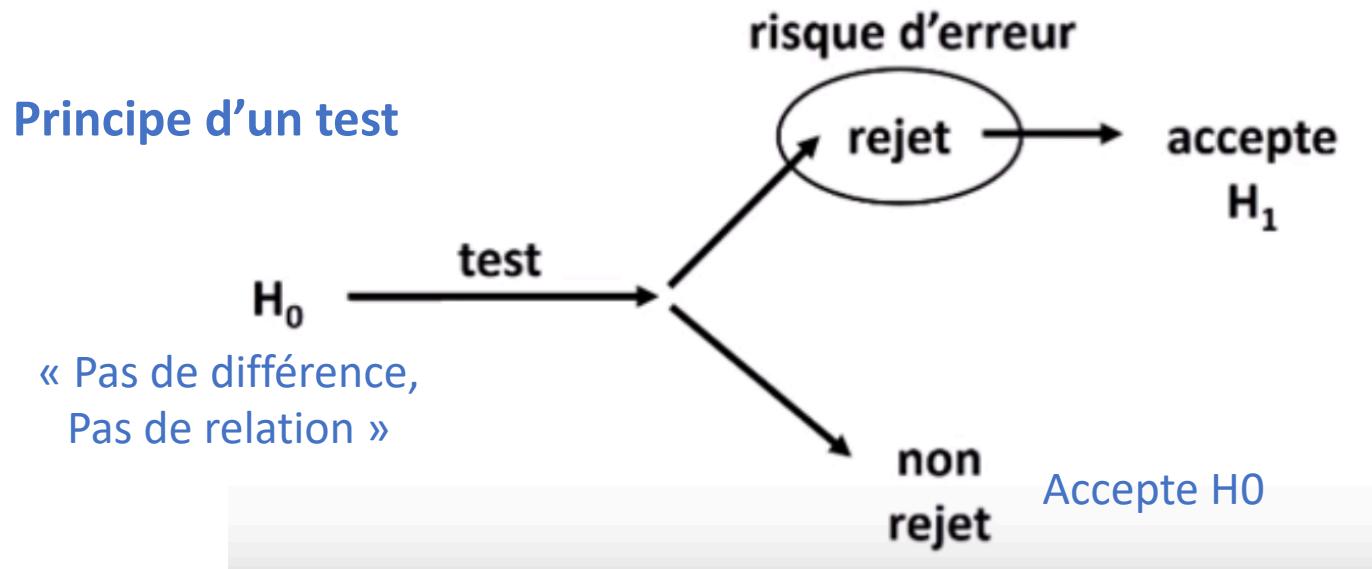


Si H_0 vraie : $\Delta = m_1 - m_2 \approx 0$

Pb des hypothèses/inférences: Soumises aux erreurs!!
 Risque → Cas des résultats d'un test statistique (H_0)

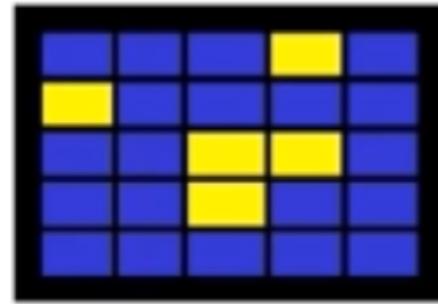
Le risque d'erreur de première espèce: α

- C'est une probabilité entre 0 et 1, ou 0% et 100%
- C'est lorsqu'on affirme une différence alors qu'il n'y en a pas (=Faux positif)!!

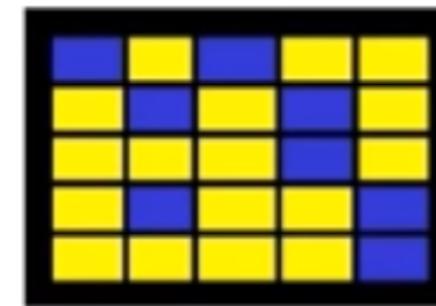


α = Risque de rejeter H_0 , si H_0 est vrai

N= 25 carreaux
→ 80% Bleu

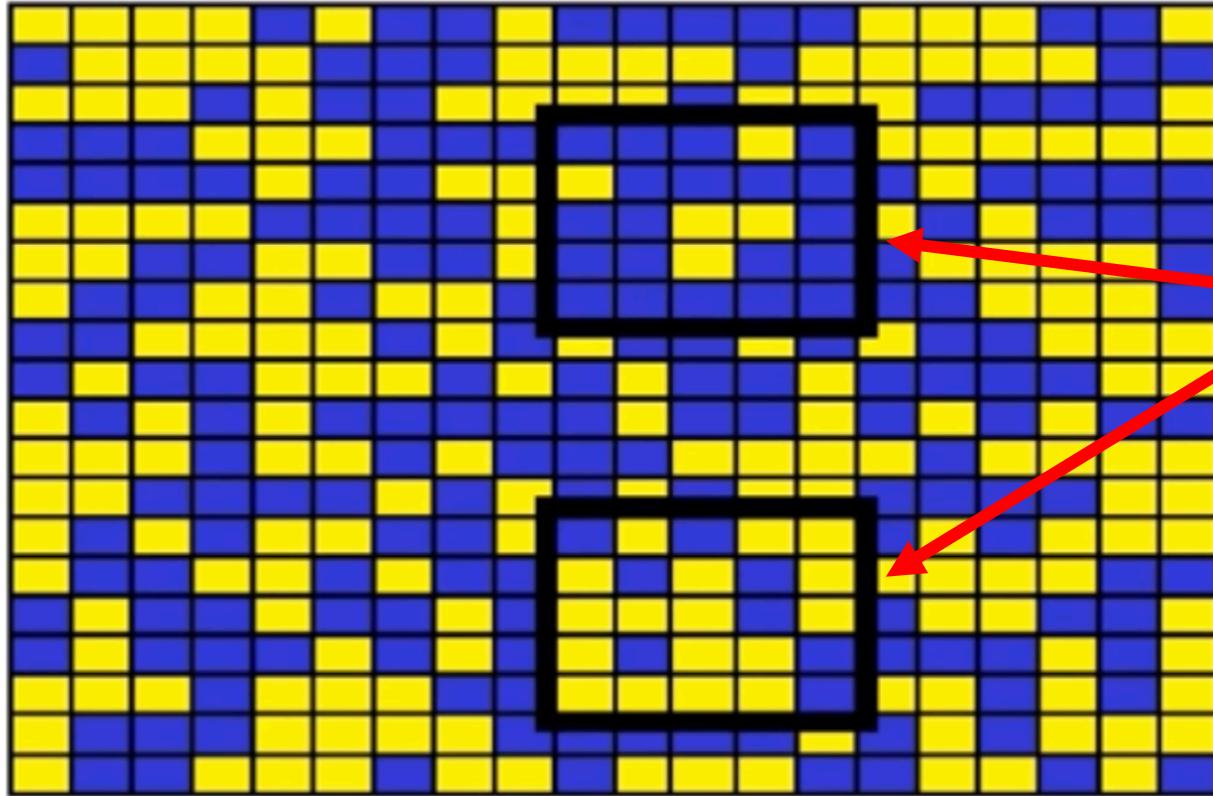


N= 25 carreaux
→ 32% Bleu



Les deux échantillons proviennent t-il du même dallage (même distribution)?

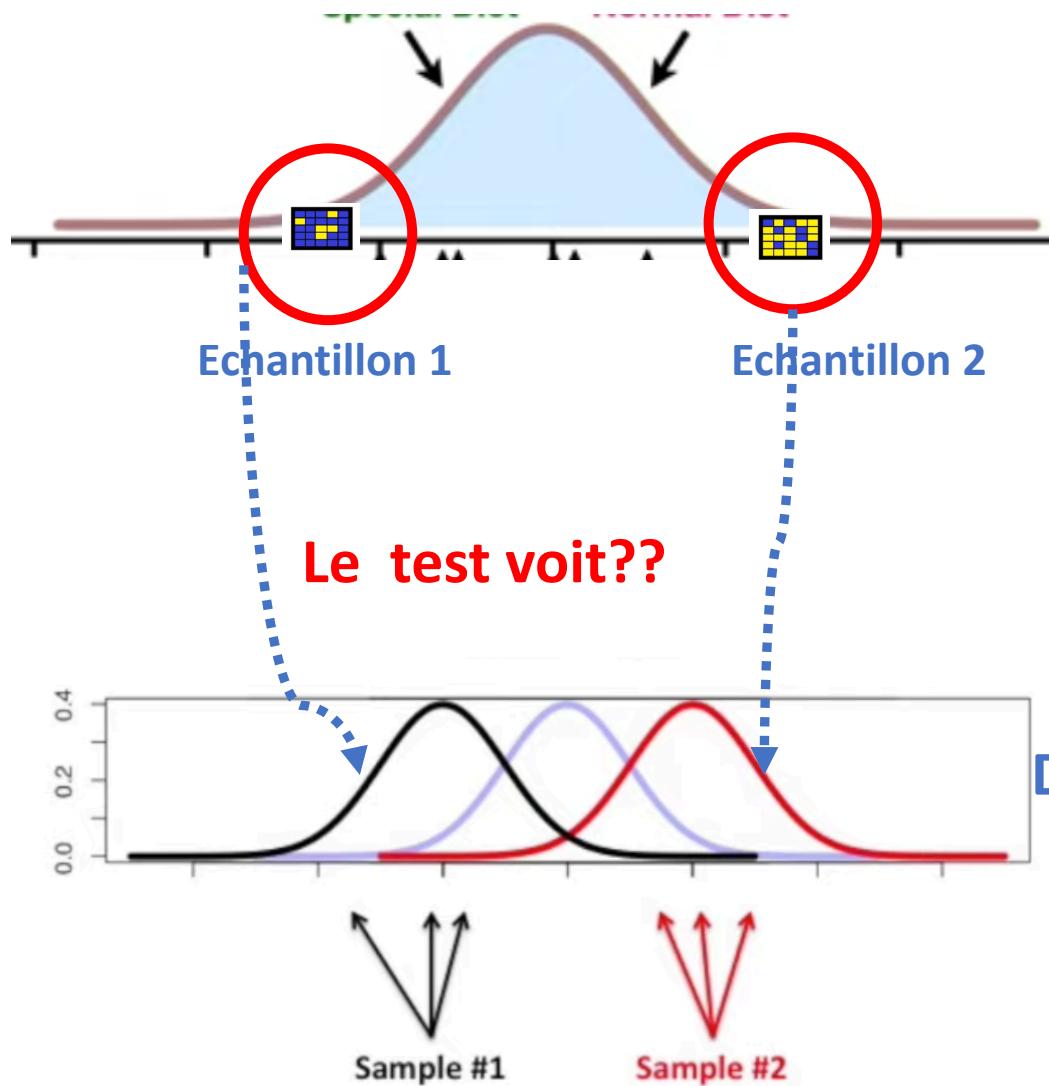
Provient du même dallage (50% Bleu, 50 % jaune)!!



Risque faible de tomber sur ce type d'échantillons (rare)

Conclure sur la base de nos échantillons qu'ils provenaient de **deux populations différentes** -> Erreur de type I

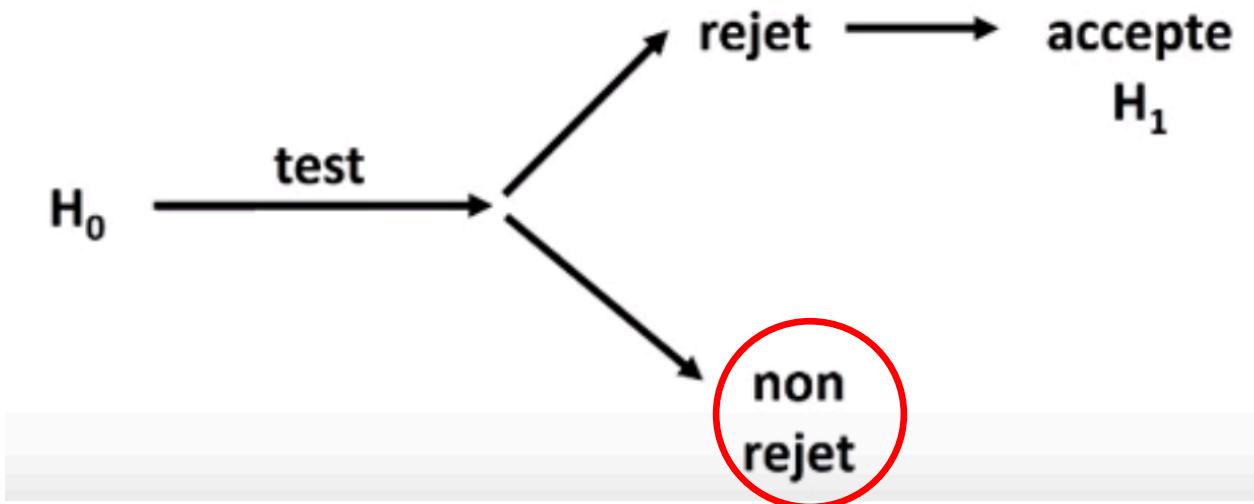
Données proviennent d'une même distribution mais...



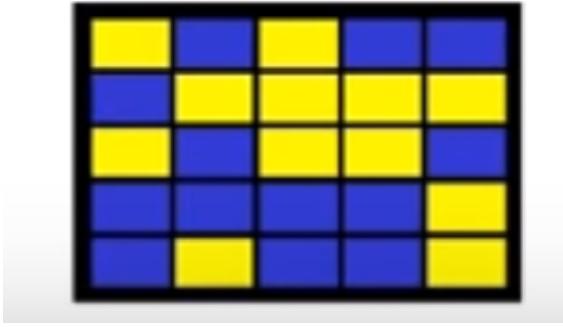
- Risque α est choisi AVANT le test
- α souvent positionné à 5% (rejeter à tort H₀)
- En sciences, le "presque aucune chance" se traduit par dans moins de 5% des cas où H₀ est vraie

Le risque d'erreur de deuxième espèce: β

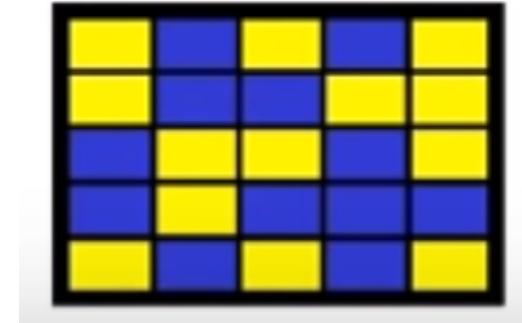
Erreur de type II : On ne conclut pas à une différence alors qu'il y en a 1 (=« Faux Négatif »)
→ Probabilité de ne pas rejeter H_0 , si H_1 est vrai



On ne sait pas calculer le risque β



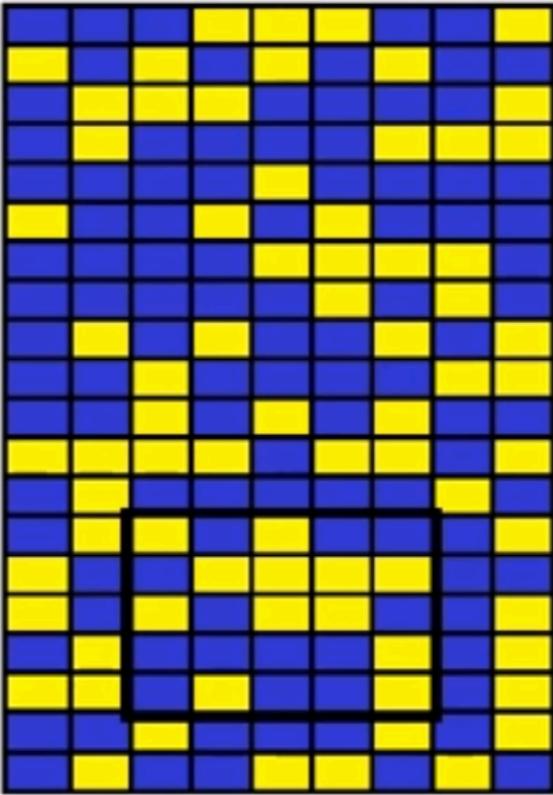
48% de bleu



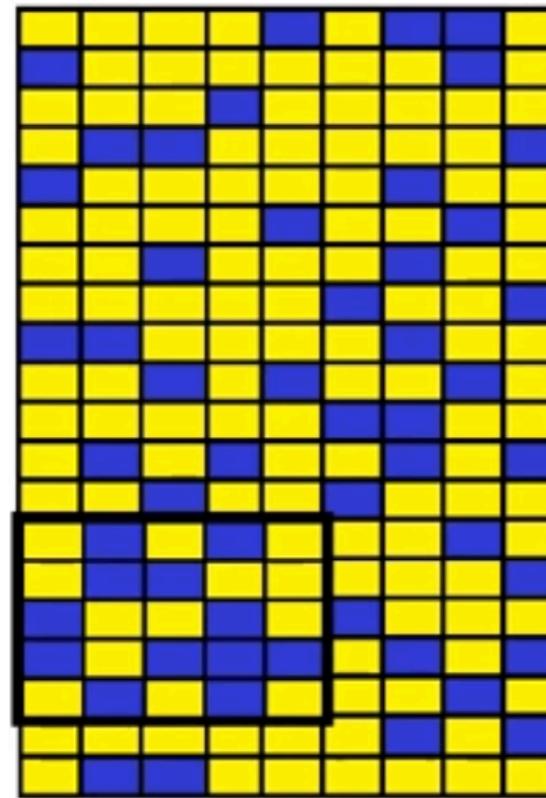
52% de bleu

Ces deux échantillons proviennent ils de deux dallages
(populations) différents ou non?

60% de bleu

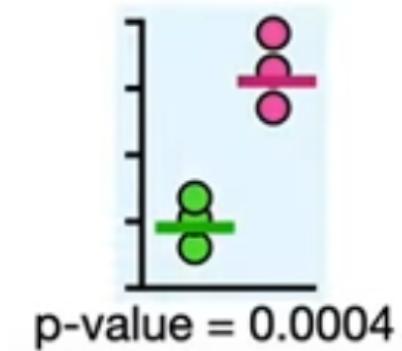
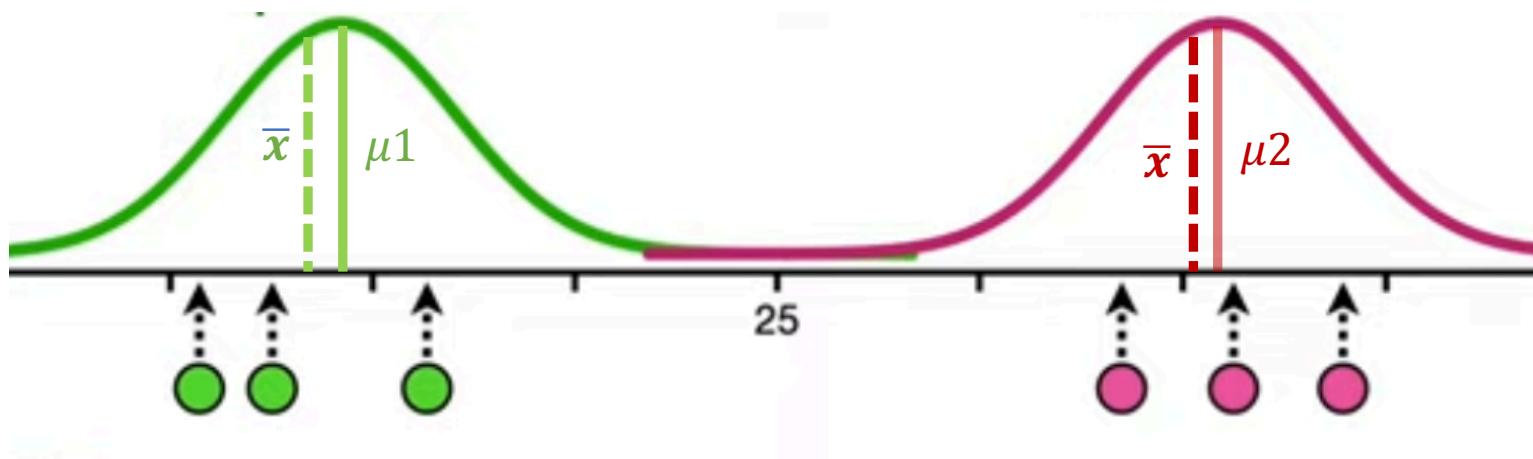


30% de bleu



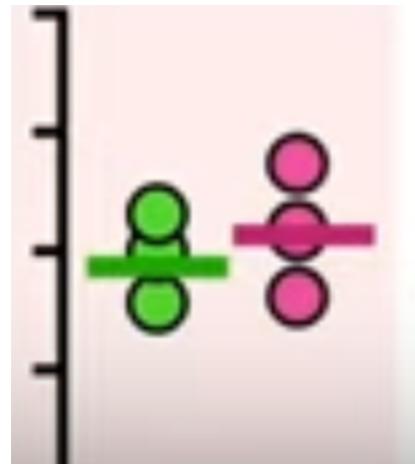
- 2 dallages différents = 2 populations différentes, H0 devrait être rejetée
Mais ça n'aurait pas été le cas lors du test avec notre échantillonnage...

Scientifiquement ... Cas échantillonnage représentatif de la population

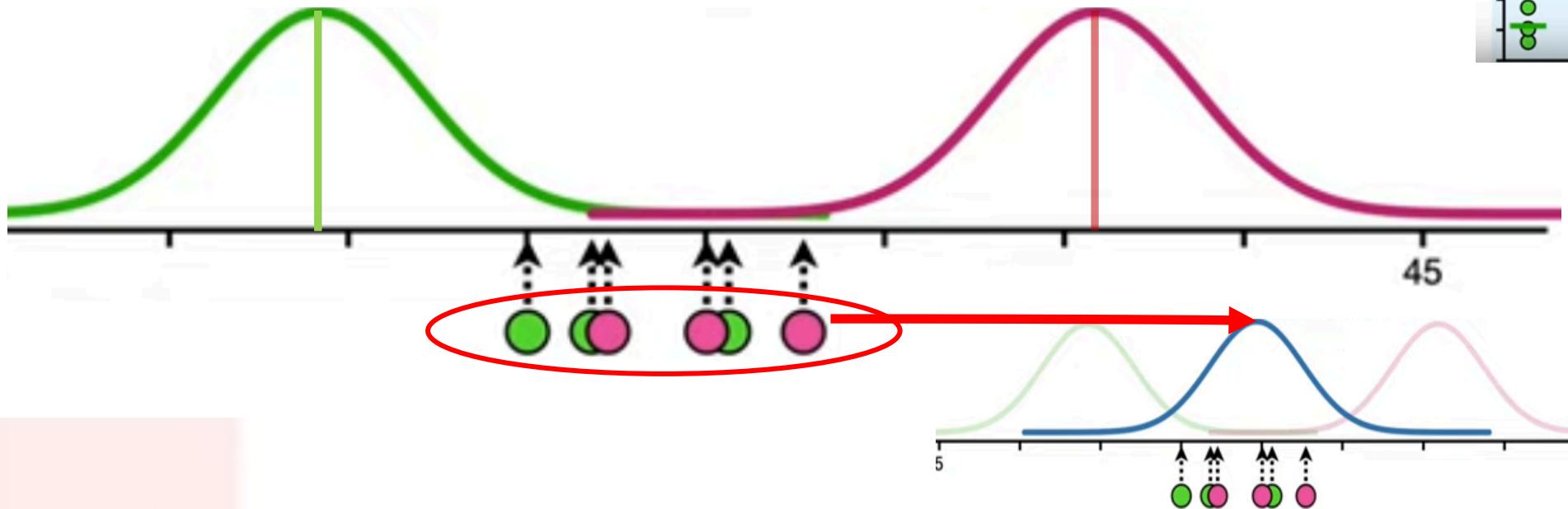


- H₀ « correctement » rejetée
- = les données n'appartiennent pas à la même distribution
- DEUX populations différentes

Mais dans certains cas...

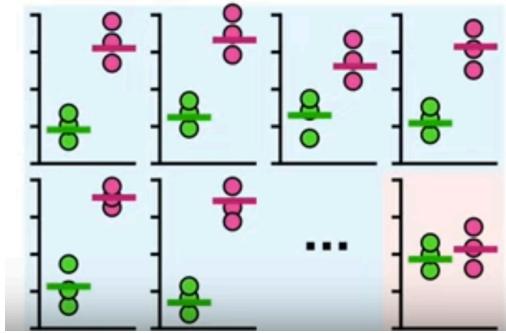


p=0.23!!!



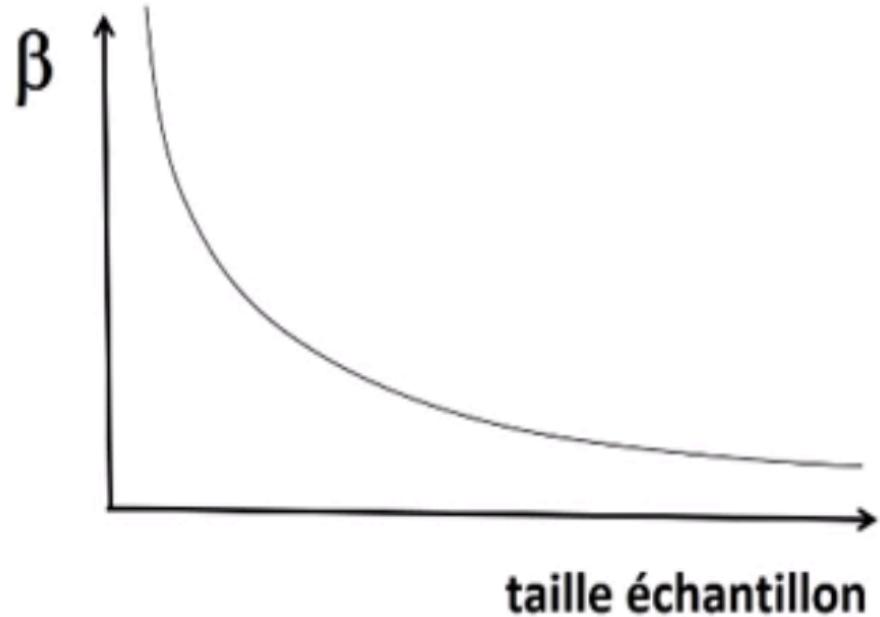
→ Même si deux distributions différentes (réalité pop) ... le test (vos données) pense qu'elles proviennent de la **MEME distribution**

→ Impossible de rejeter correctement H0 ...



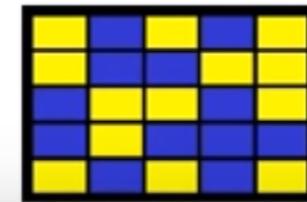
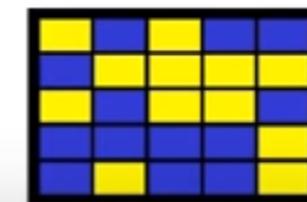
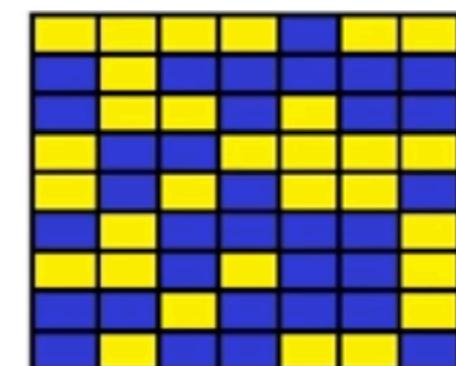
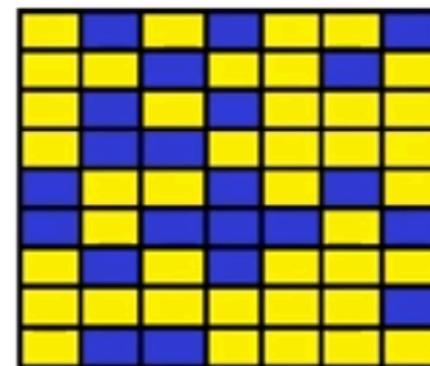
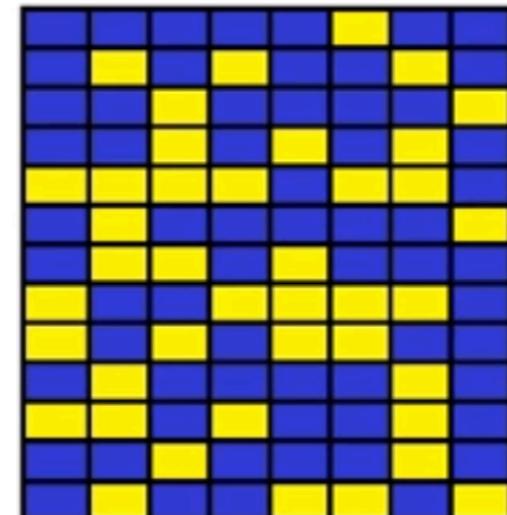
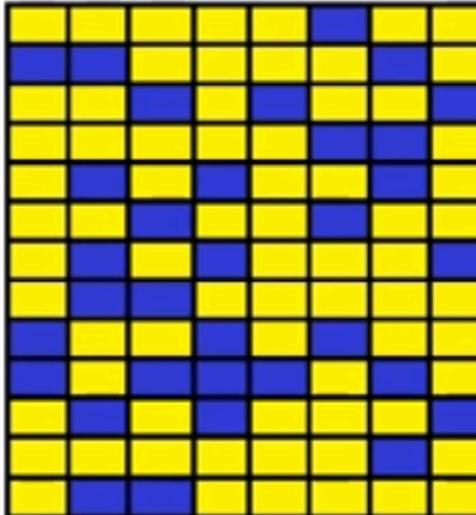
Relation fondamentale

$$\text{Power} = 1 - \beta$$

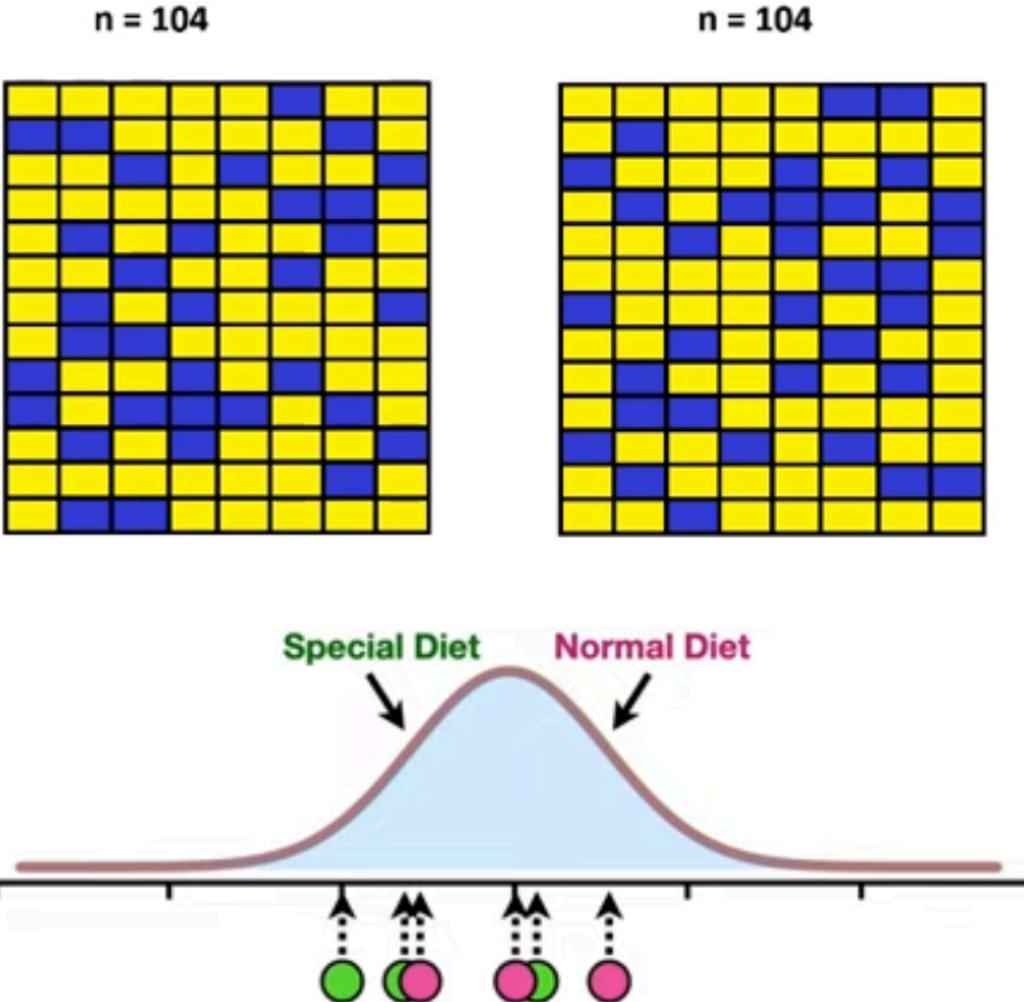


Plus la taille augmente plus les différences apparaissent!
La puissance du test augmente!

Power/Puissance test : Probabilité de rejeter correctement l'hypothèse H₀
= Capacité d'un test stat de détecter les différences ou relation



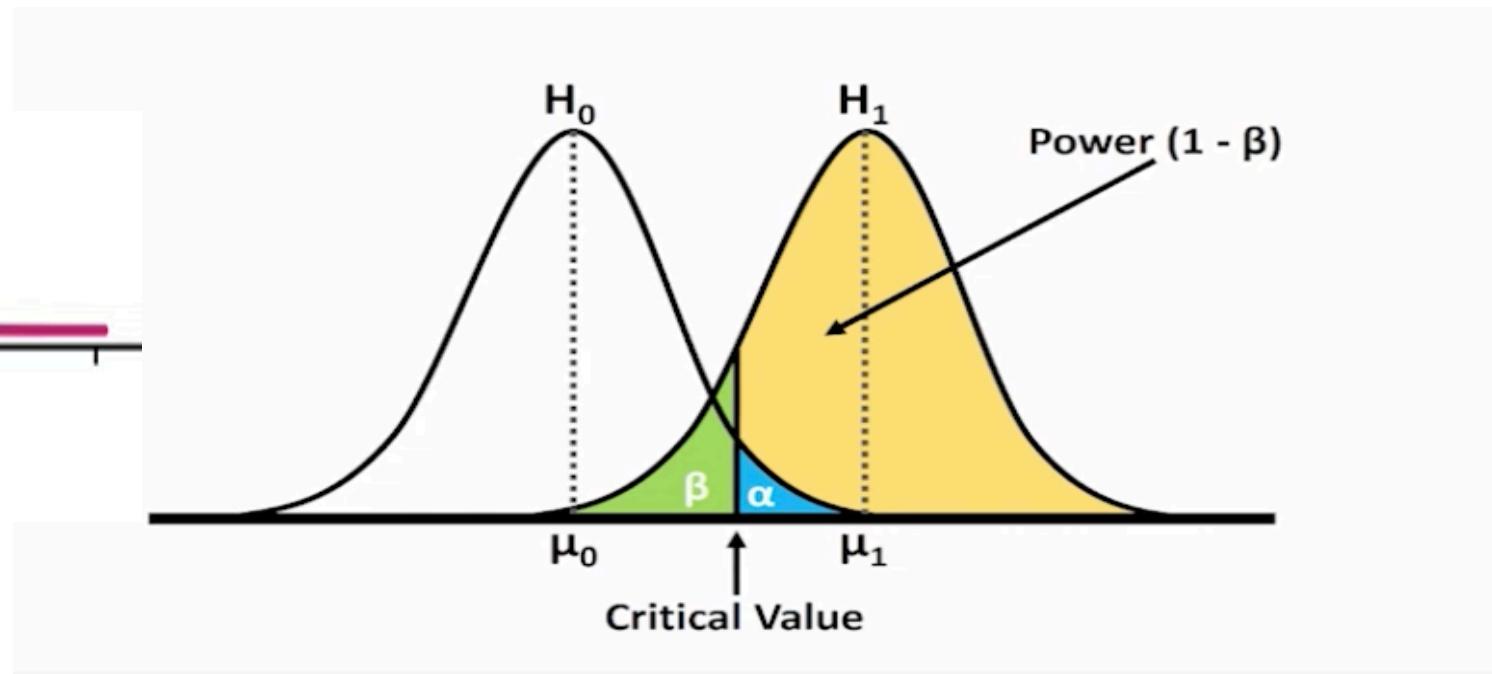
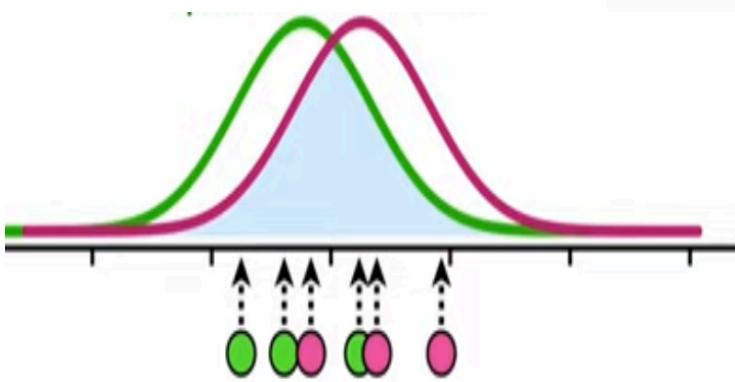
Attention !!!



Bilan

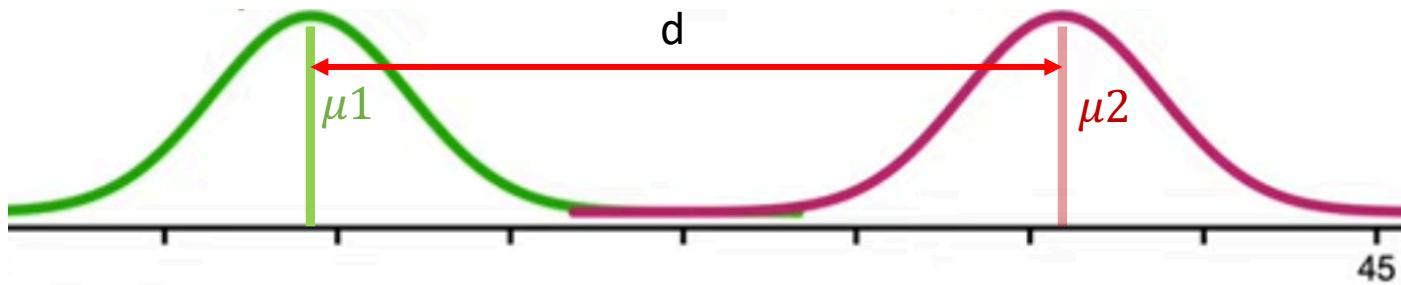
Population

	H_0 vraie	H_1 vraie
accepter H_0	OK	erreur type 2 β Faux Négatif
rejeter H_0	erreur type 1 α Faux positif	OK



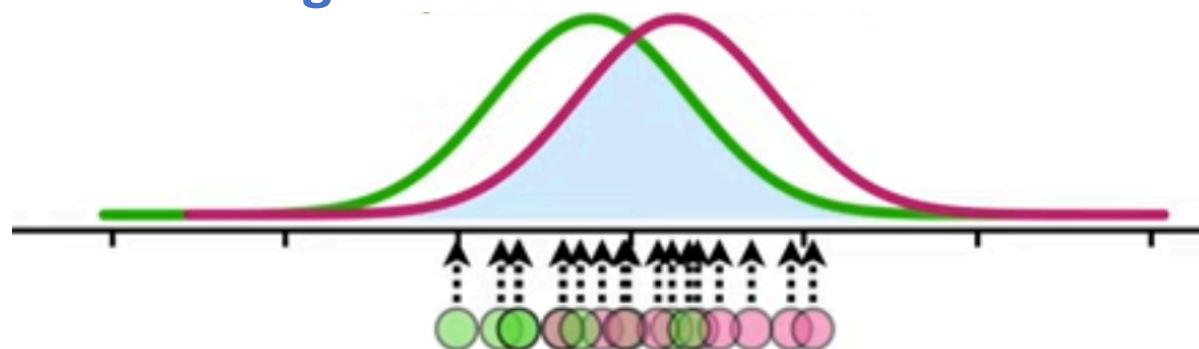
Puissance est fonction de la taille d'effet, nombre d'échantillons...

- Effect Size



Taille d'effet = mesure la force de la différence ou relation (r)
Elle est ce qu'elle est...

- Augmenter l'échantillonnage



Analyses statistiques pour la diversité Alpha

- **Analyses univariées** cherchent à quantifier l'association brute entre une variable à expliquer (**réponse**) et une variable explicative (**facteur**).

- Le choix du test statistique sera fonction du type de variable!

Cas: 1 variable quantitative et 1 qualitative!

→ Est-ce que les variations de la **richesse en espèce** (variable réponse) peuvent être expliquées par la variable explicative (facteur) **Traitement**?

→ Comparaison de **moyennes** entre groupes

- Utilisation de test paramétrique, non paramétrique??
- Choix du test??
- Combien de groupes??
- Faut il faire un Post hoc test ??



Tests Paramétriques (Loi normale)

- **T-test (paired or unpaired):** Compare of the means from **2 sample groups** for one variable
- **One way Anova (variance analysis) :** compare the means of **three or more sample groups** for one variable (e.g. Age, Sex, Region, ...)

NB: Two way Anova (two variables), Manova (multiple variables)

Tests non-paramétriques

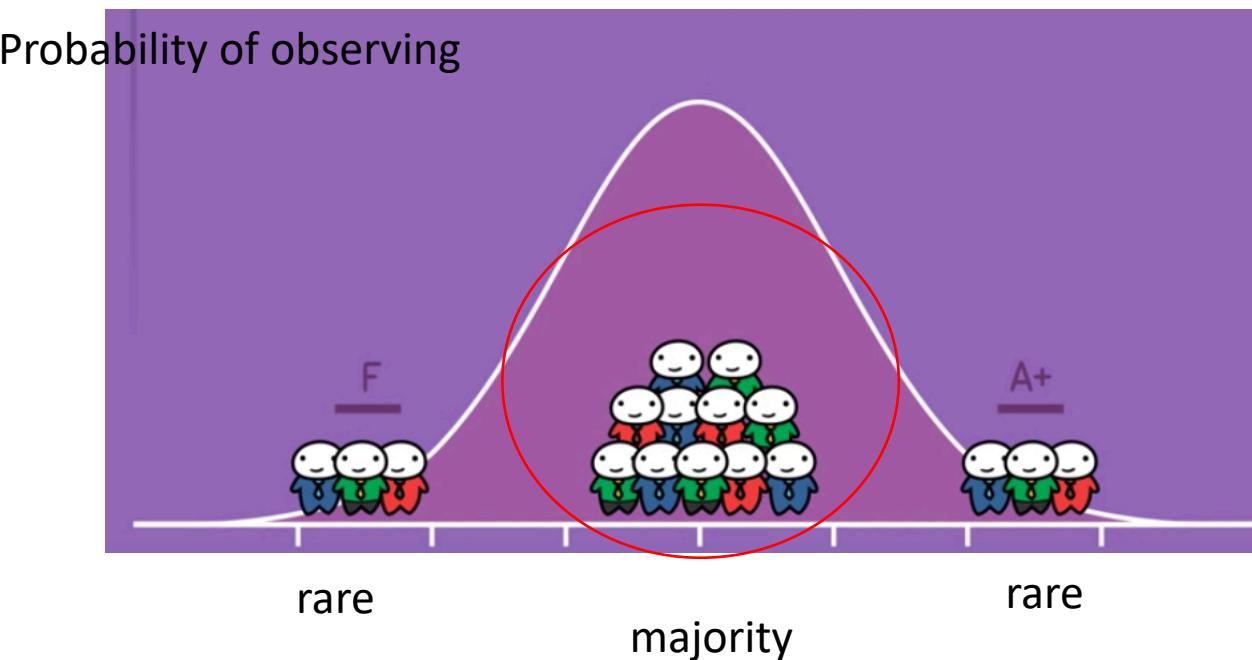
No assumptions are made for the distribution of data:
Distribution-free tests, they are alternative to parametric test

- Wilcoxon Rank test : samples are paired/unpaired, 2 sample groups
- Mann-Withney test: Independent samples, 2 sample groups
- Kruskal wallis test : Independant samples, Three or more groups
- Friedman test : paired samples, three or more groups

NB: Rank, permutations tests...

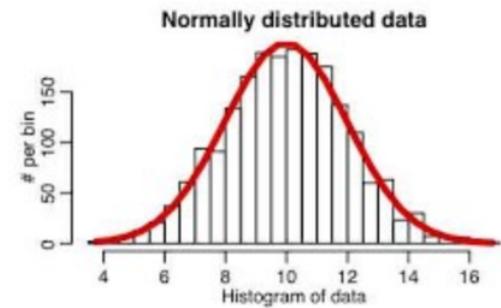
Tests paramétriques & Hypothèses

- Paramétriques plus puissants
- Conditions...
 - Indépendance des samples
 - Suit une loi → Loi normale, distribution continue



Appliquer un test paramétrique... check-list!

- Vérifier la **normalité** des données (nos cas): Shapiro Test & QQ-plots
!!Shapiro: H0: « données normalement distribuées »



- Vérifier **l'homogénéité** de la variance : F-test (2 groups), Bartlett's & Levene's tests
!!H0: « pas de différence »

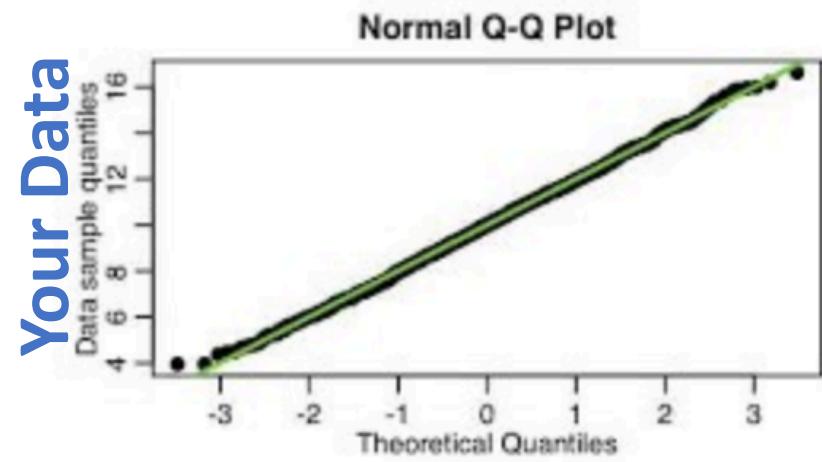
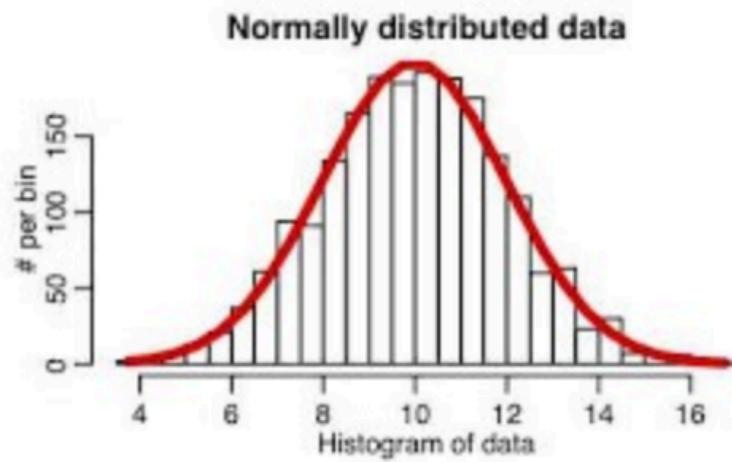
$$S^2 = 169$$

$$S^2 = 289$$



Q-Q plot normal: Comparer sa distribution à la distribution normale...

Est-ce que mes données suivent une loi normale?

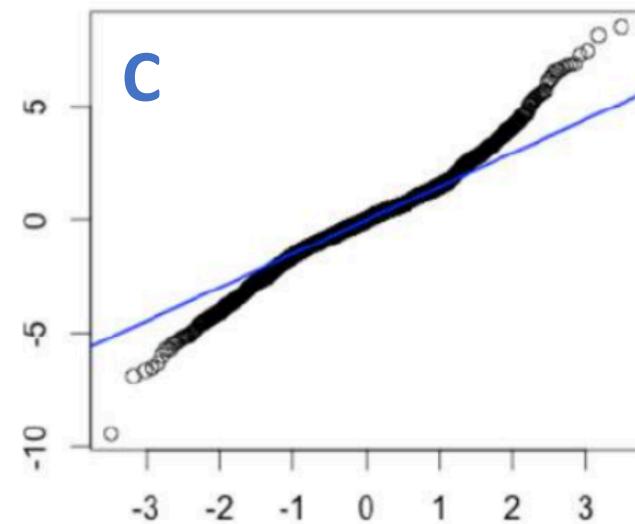
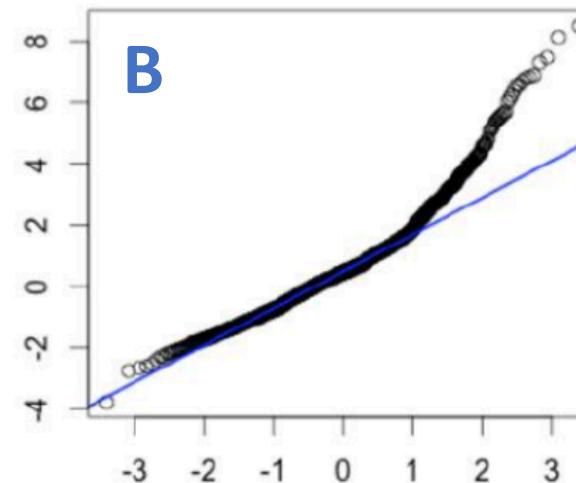
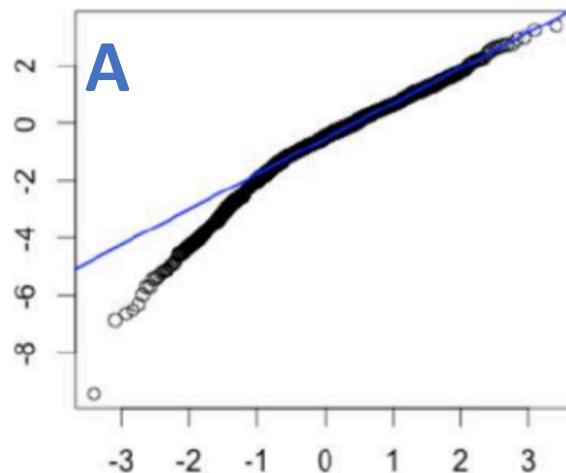


Conclusion?

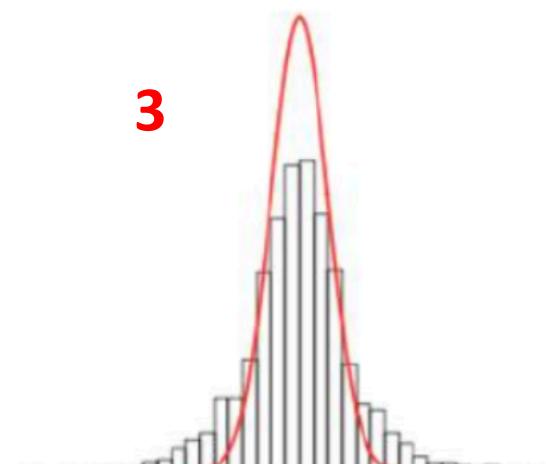
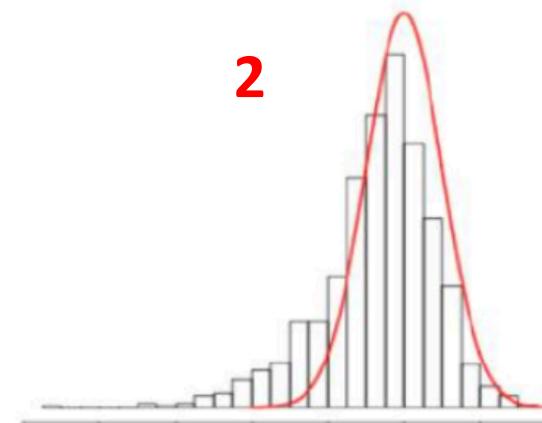
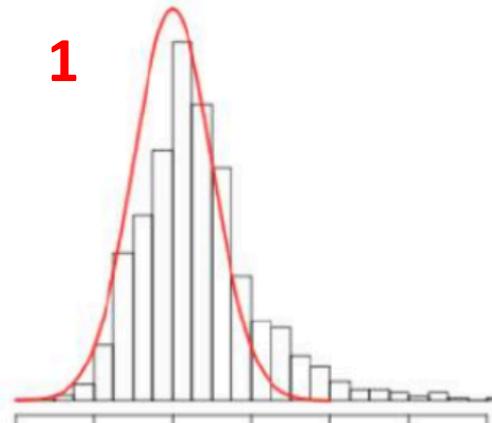
Normal Data ($\mu = 0, SD=1$)

La droite du QQ-Plot indique la position que doivent avoir les points s'ils obéissent exactement à la distribution normale

Quelles sont les distributions correspondantes à ces QQ-plots ?



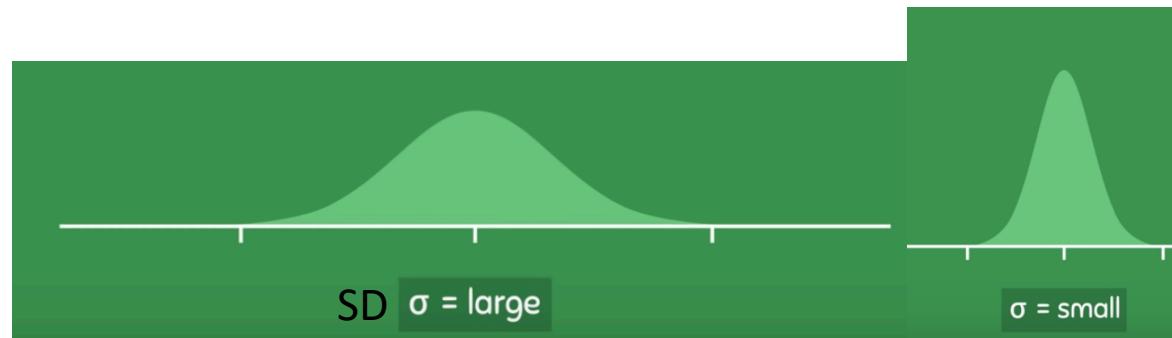
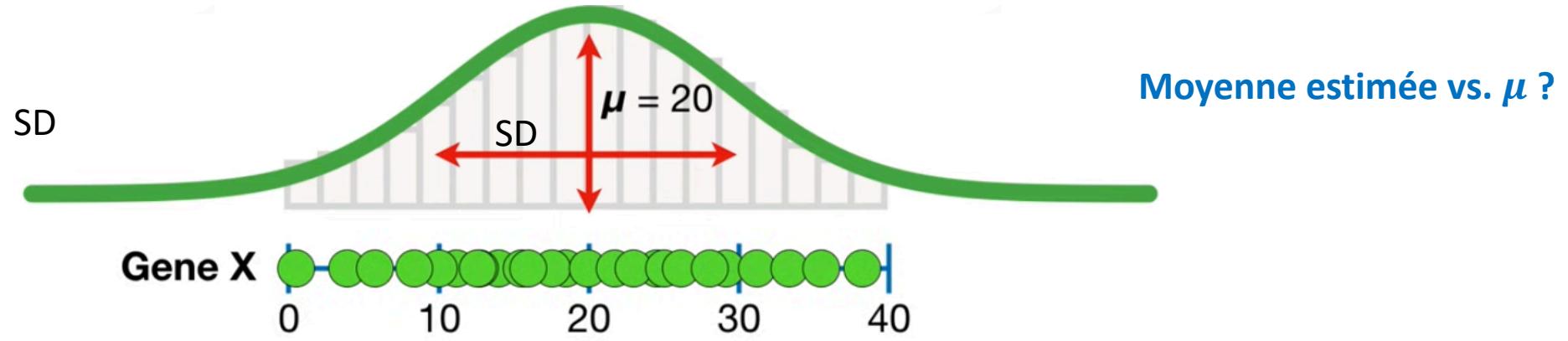
?????????????



Caractéristiques d'une distribution normale

Symétrique

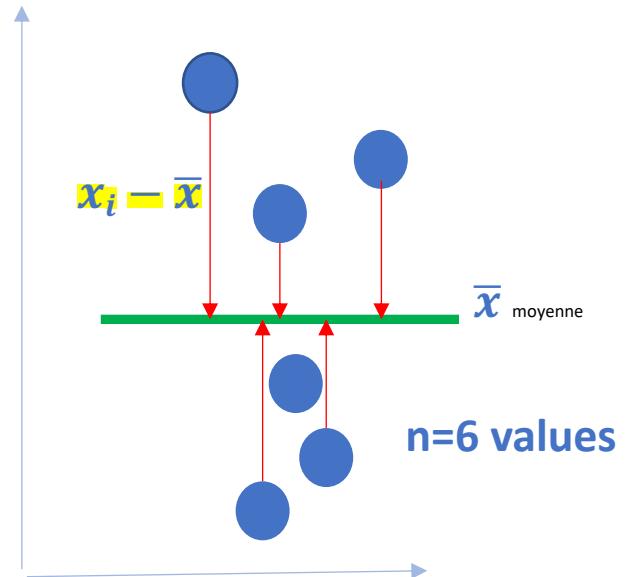
- Centré sur la moyenne/mean
- Dispersion autour de la moyenne: Standard deviation (SD)
 - 95% data sont dans -/+ 2 SD



Variance= S^2/σ^2

- Variance mesure le degré de dispersion d'un ensemble de données autour de la moyenne
- Moyenne arithmétique des carrés des écarts à la moyenne! 😞
- Exprimée en Unité carré

$$S^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$$



Ecart-type (Standard Deviation)= S/σ

$$S = \sqrt{S^2}$$

L'avantage de l'écart-type est de s'exprimer dans la même unité que la série de données

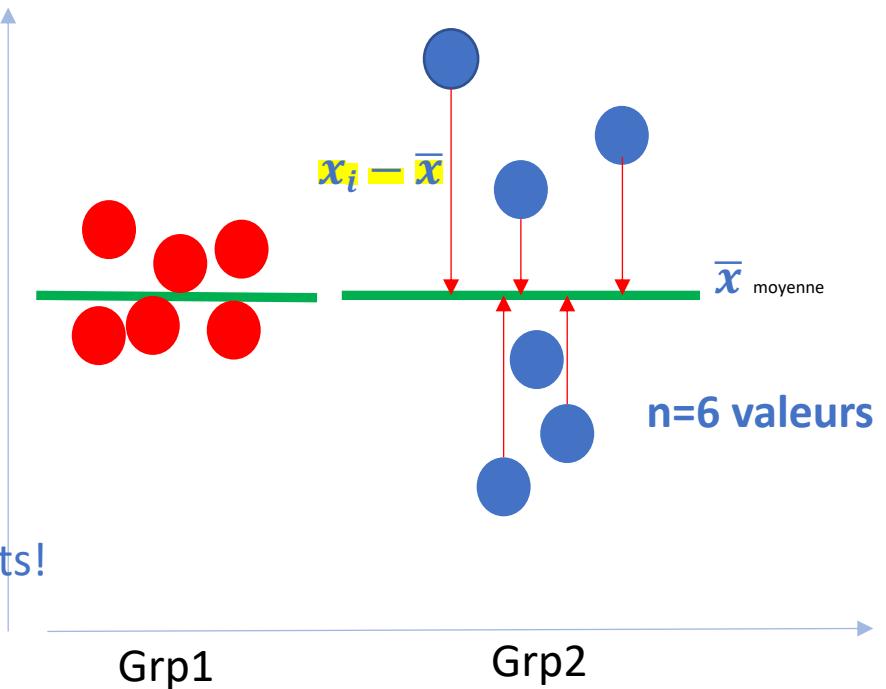
$$S^2 = \sum_{i=1}^n \frac{(x_i - \bar{x})^2}{n-1} = \frac{\text{Sum of Squares (SS)}}{n-1}$$

SS sera bien plus grande dans l'échantillon

Exemple de Résultats stats utilisant la variance

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
groupe	3	13.03	4.343	0.211	0.887
Residuals	14	288.75	20.625		

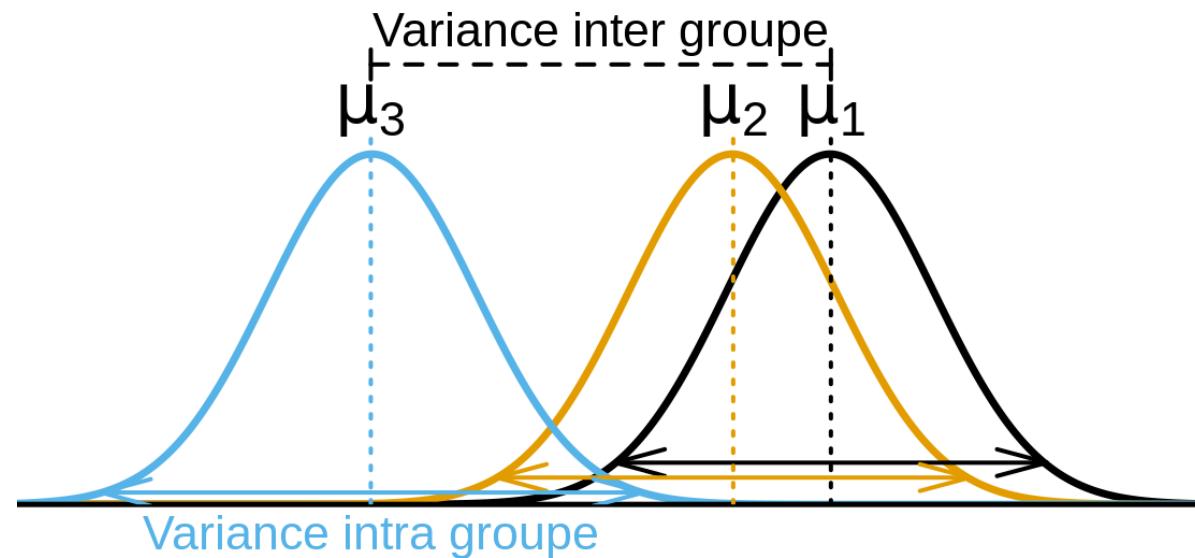
- **Sum of Squares (= SS, Sum Sq)** dans la description de vos résultats stats!
→ Numérateur de la Variance!!
- **Mean Square (= Mean Sq= Toute la formule = VARIANCE!!!)**



ANOVA: ANalysis Of VAriance (One way Anova= Univariée)

(Minimum 3 groupes)

- Compare la variance des groupes à celle à l'intérieur des groupes (i.e. variance intra-groupe) pour une seule variable explicative (Qualitative!)



ANOVA: ANalysis Of VAriance (One way Anova= Univariée)

- Postulat = *Les variations observées entre les moyennes des différents groupes (e.g. espèces) sont si faibles qu'elles s'expliquent facilement par le hasard!!!*
- Evaluation : Compare la variance inter-groupes à celle à l'intérieur des groupes (i.e. variance intra-groupe) pour une seule variable/facteur (Qualitative!)
- Pourquoi L'ANOVA → variations à travers la grandeur Variance :
- Comment ? Décomposition de la variance totale :
 - Une partie attribuable au facteur/variable indépendante = variance inter-groupes
 - Une partie attribuable à l'expérimentale/autre chose = variance intra-groupes (fluctuation de l'échantillonnage)

- Statistique F = $\frac{\text{Variance Inter-groupes}}{\text{Variance Intra-groupes}}$

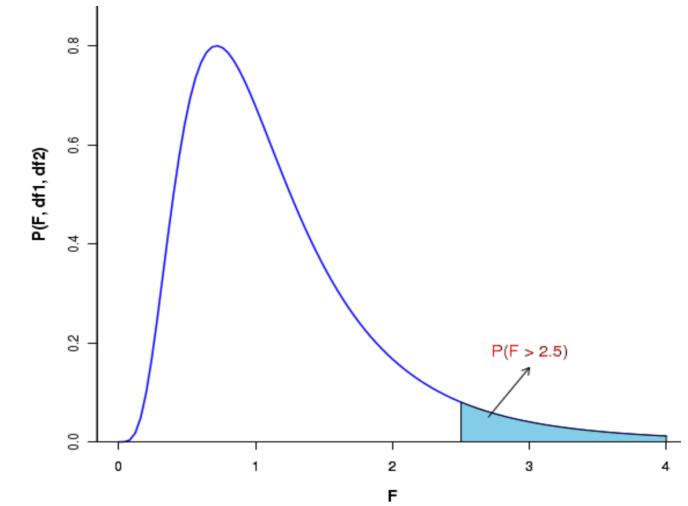
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
groupe	3	13.03	4.343	0.211	0.887
Residuals	14	288.75	20.625		

L'idée : si le facteur a vraiment un effet, la part des variations qu'on peut lui attribuer = Variance inter-groupes sera significativement plus élevée que la part des variations qu'on ne peut pas lui attribuer = Variance Intra-groupes!

Statistique F suit une loi dites de Fisher-Snedecor : = Distribution F utilisée pour test des variances, distribution des variances n'étant pas normale.

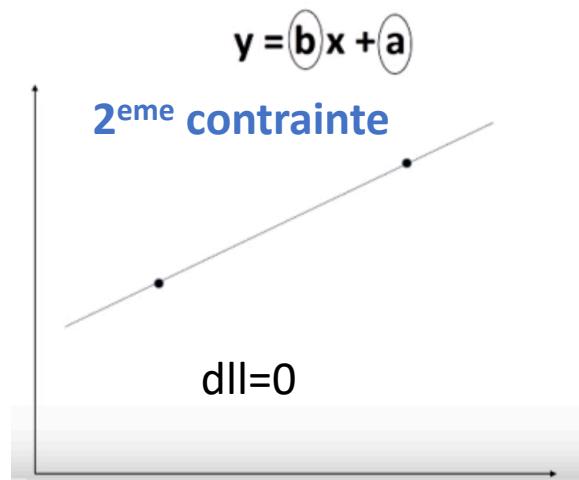
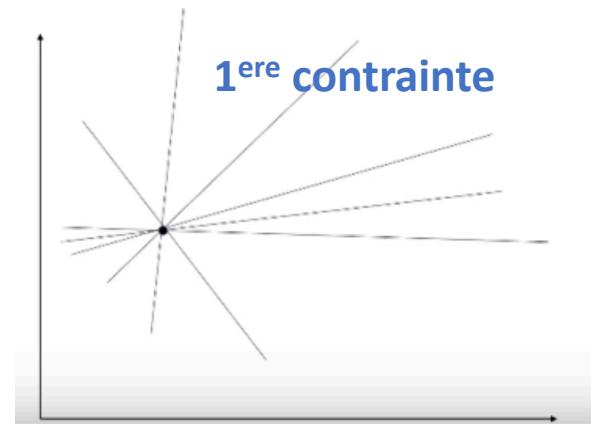
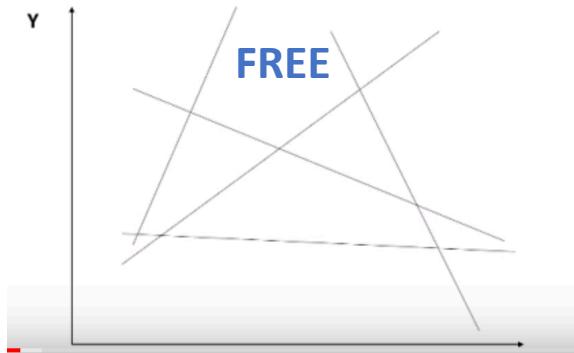
- Mise en relation d'une valeur observée de F , avec la probabilité a priori de rencontrer une telle valeur ($>$ ou $=$) par hasard:
 → probabilité donnée par la loi = p-value!

	Dénominateur S^2	Numérateur S^2	S^2			
groupe		Df Sum Sq Mean Sq		0.211	Pr(>F)	
	3 13.03	4.343		0.887		
Residuals	14 288.75	20.625				
	⋮	⋮				



Degré de liberté

Selon le test (t-test, F-test etc)
La formule du calcul dll est connue



Cas ANOVA: Test F

variances	ddl	F
entre k groupes	v_k	$k-1$
résiduelle	v_r	$N - k$

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
groupe	3	13.03	4.343	0.211	0.887
Residuals	14	288.75	20.625		

Modalités	Effectifs
A	
B	
C	
D	
Total	56

Modalités	Effectifs
A	22
B	11
C	15
D	
Total	56

Dll=?

Régression Linéaire Simple (=GLM)

Modèle cherchant à établir une relation linéaire entre une variable, dite expliquée/dépendante (Y), et une autre dite explicative/indépendante (X).

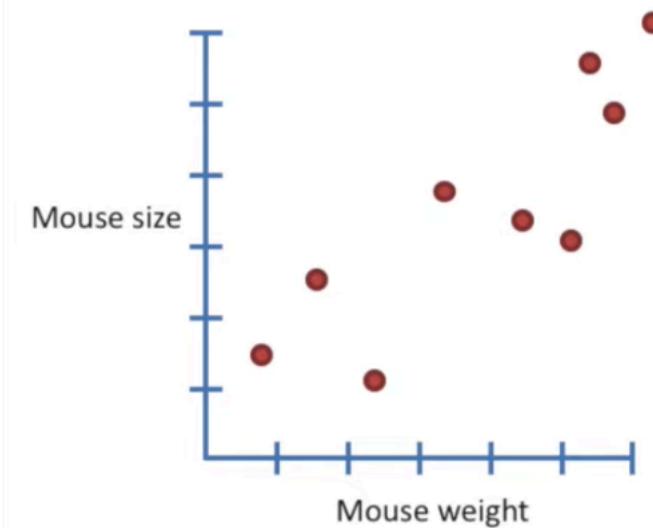
- Expliquer et prédire!
- Modélise une relation de type linéaire ($Y=aX+b$)

Comment ?

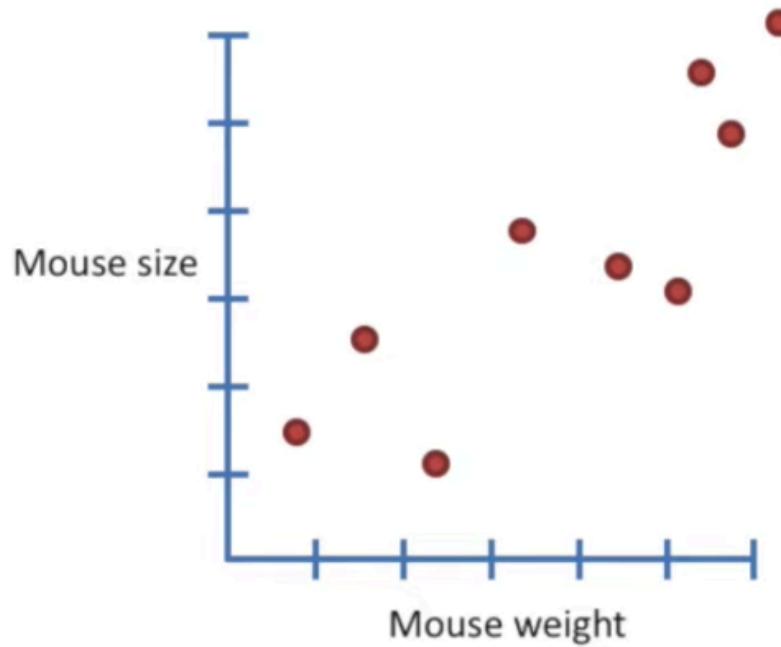
- Méthode des moindres carrés
- Alternative: Inférence Bayésienne, Maximum de vraisemblance

Hypothèse – prérequis

- H0: pas de relation!!
- Distribution Normale
- Homogénéité des variances

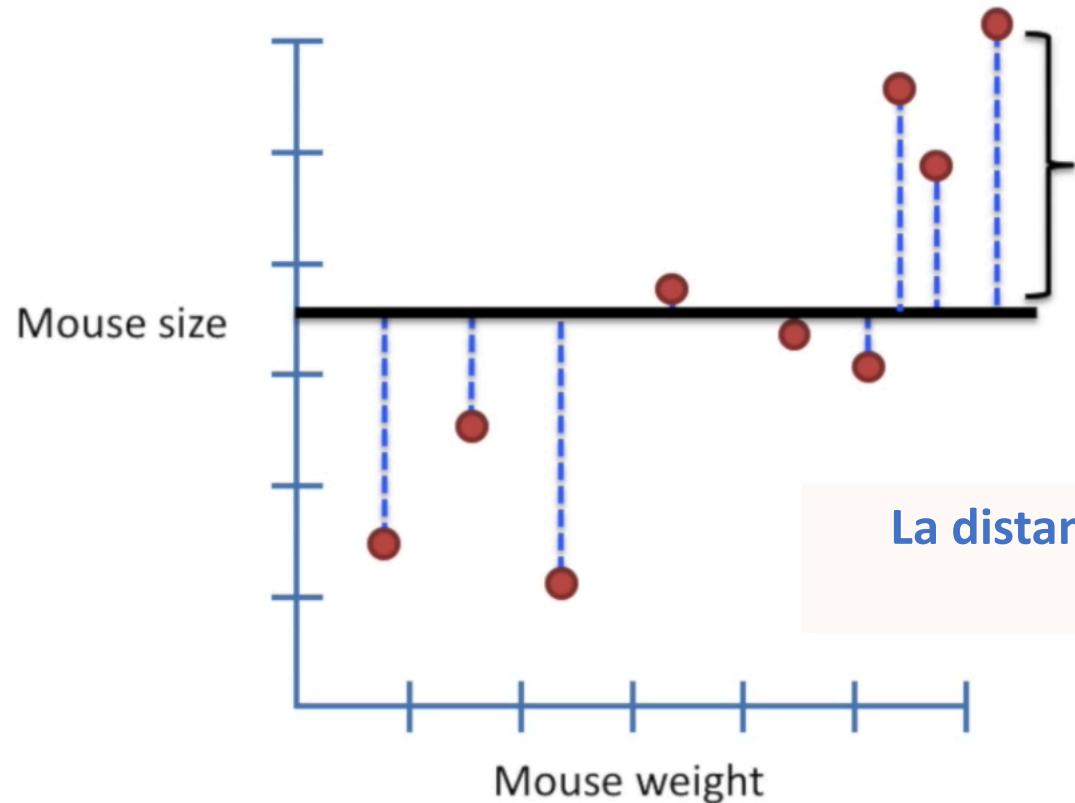


Mesures du poids (Weight) et de la taille (Size) de souris



- Existe-t-il une **relation entre Poids et Taille** permettant de rassembler les points autour d'une ligne droite dans le nuage de points ?
→ Déterminer une **fonction affine**
- Est-ce que « le poids » peut prédire « la taille » correctement?
→ Qualité de la régression: **coefficient détermination R^2**
- Est-ce que la relation entre le poids et la taille est due au hasard!
→ **Calcul de la p-value**

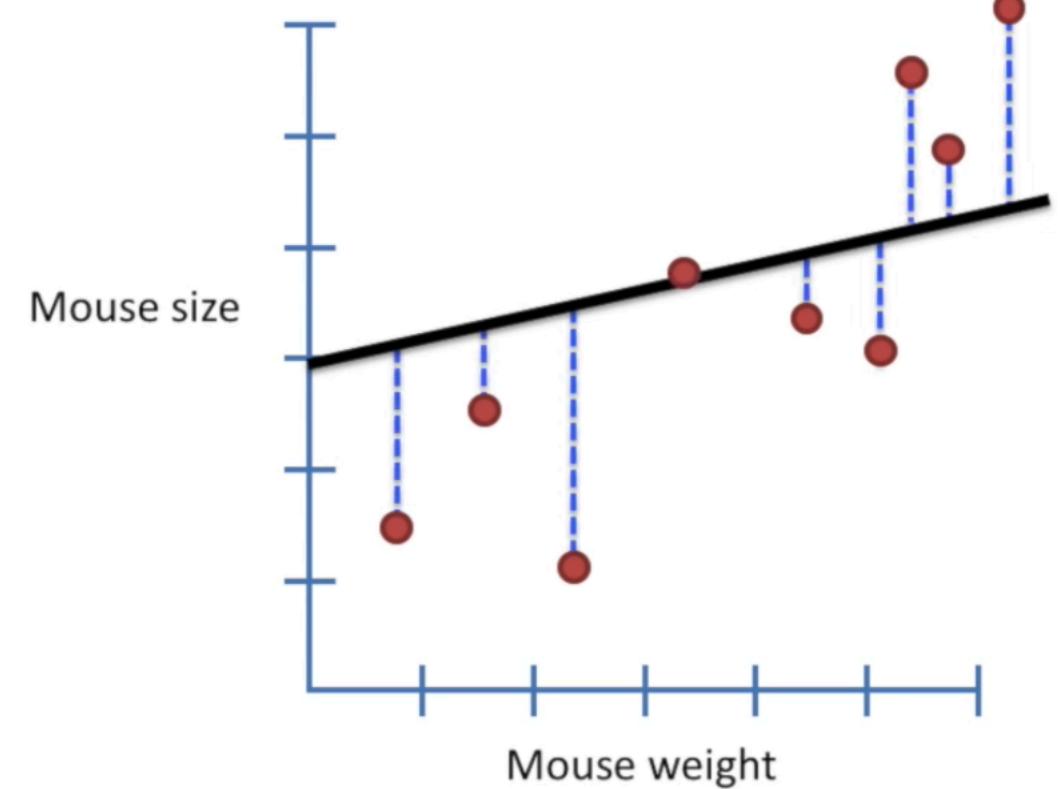
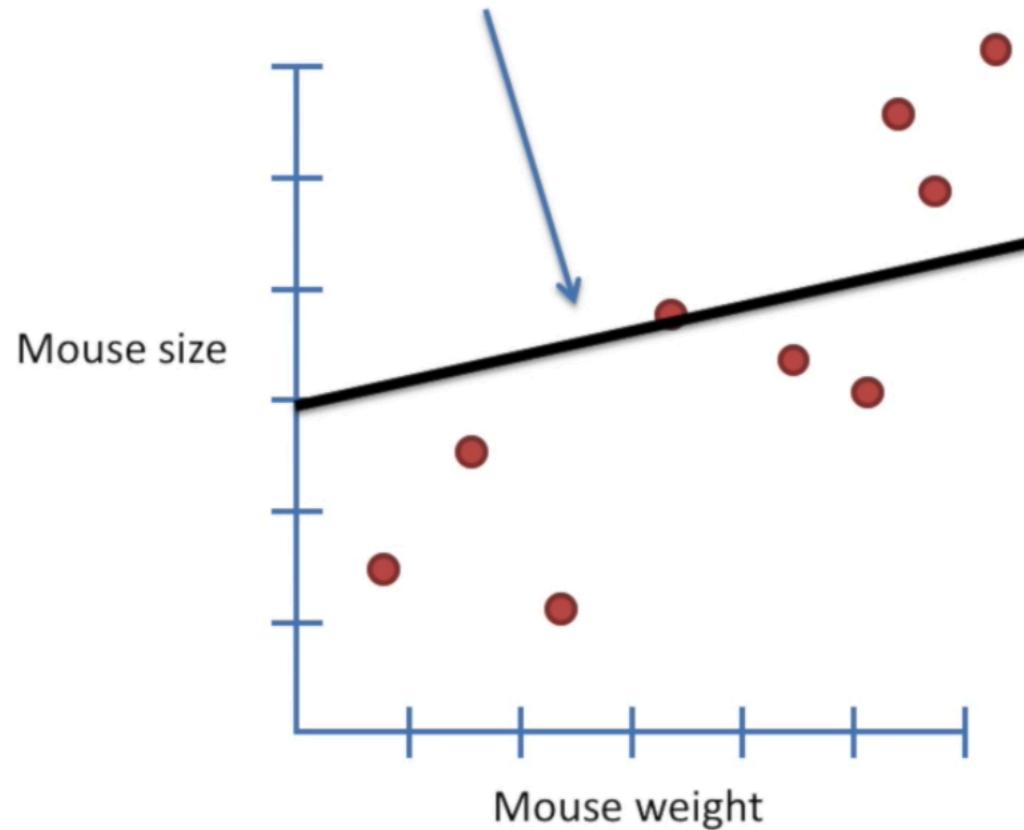
Approche de l'ajustement



- Tracer une droite perpendiculaire à l'axe « Size »
- Mesurer la distance des points à la ligne (projection verticale)
- Faire la somme des carrés des distances (S1)

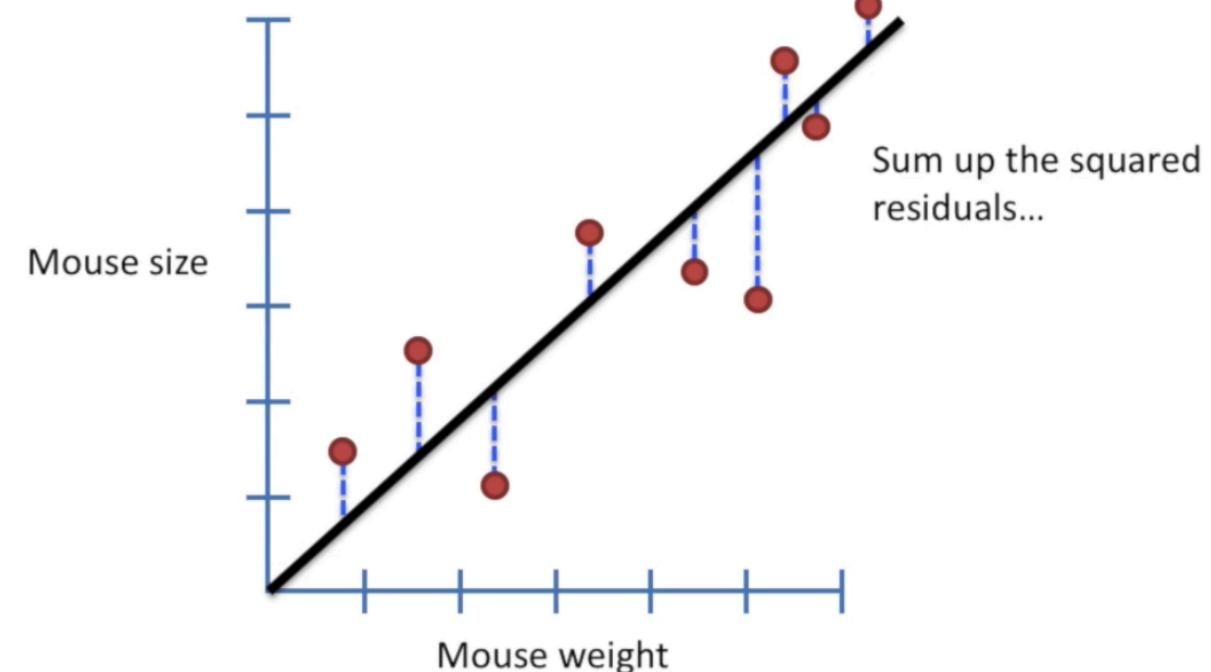
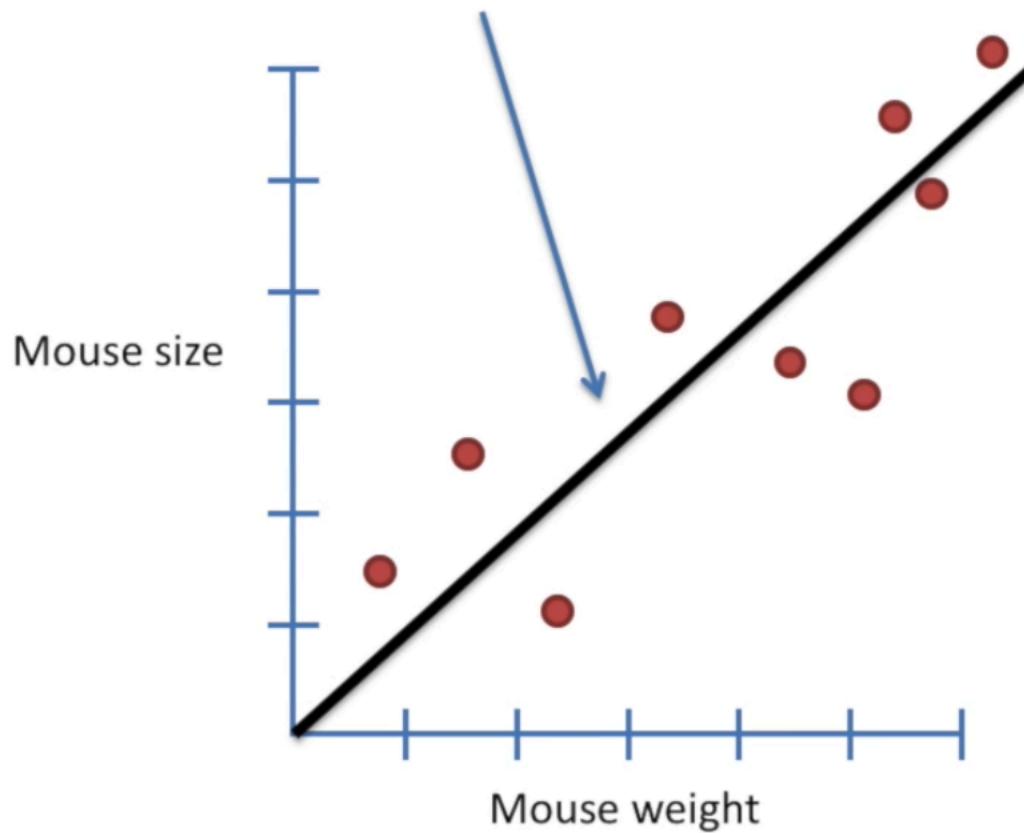
La distance « point-droite » est appelée
« RESIDUAL »

Third, rotate the line a little bit...



→ Somme des écarts carrés des « residuals » : S^2

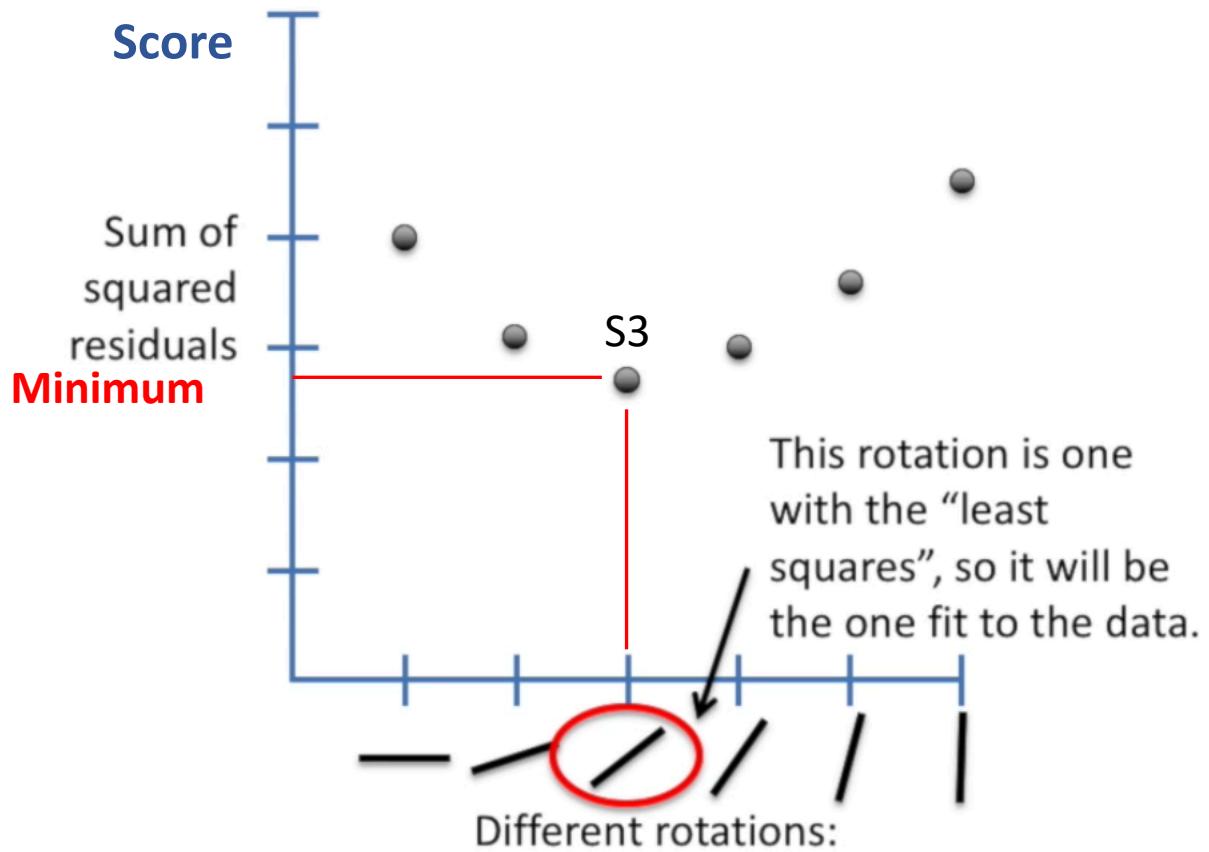
Rotate the line a little bit more...



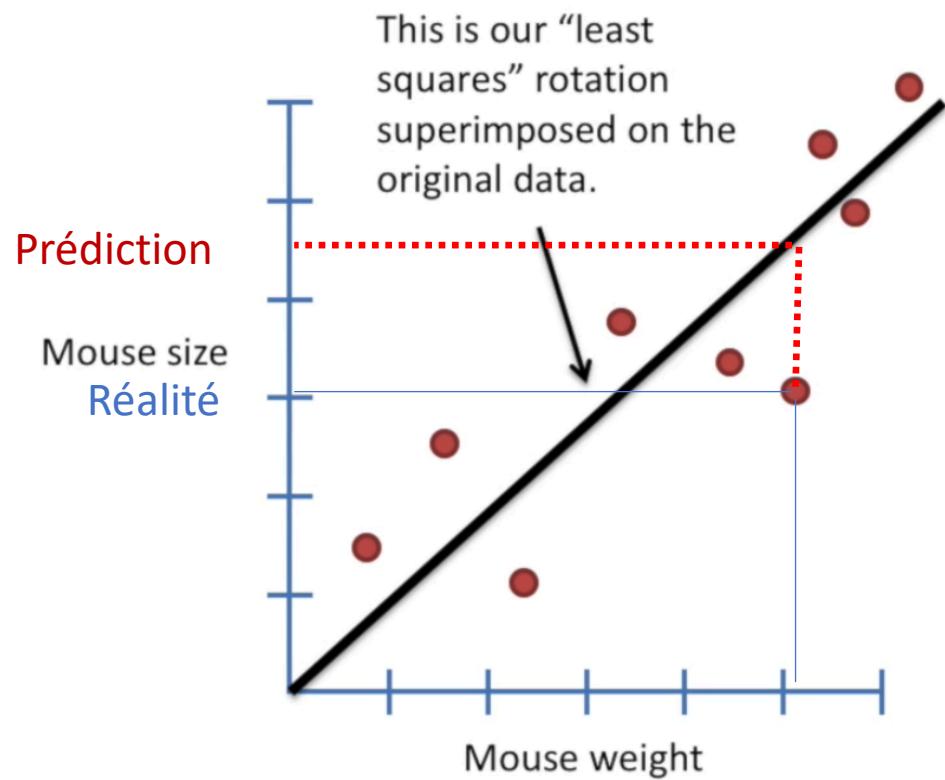
Again & again, Rotation, Somme des
écart carrés des résiduels (S3, S4, etc)

...

Bilan : Sommes des écarts carrés des « résiduels » pour chaque rotation



Meilleure rotation (position de la droite) est celle qui minimise le score de la somme des écarts carrés des résiduels!!!!
→ Moindres carrés



$$y = 0.1 + 0.78x$$

Dépendance au « Mouse weight »

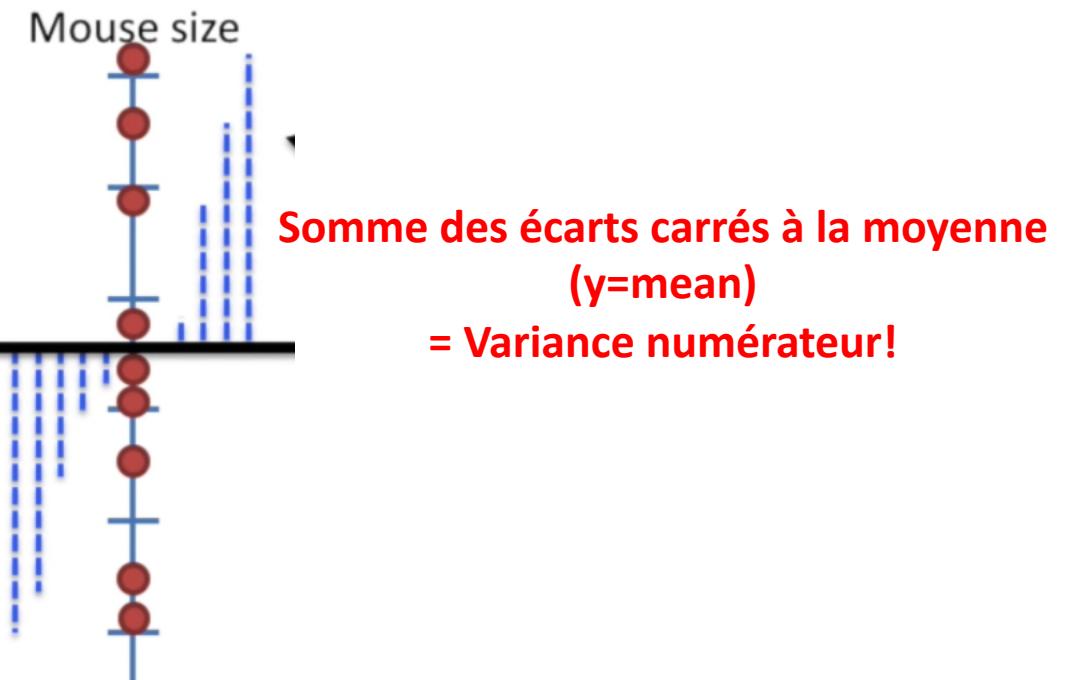
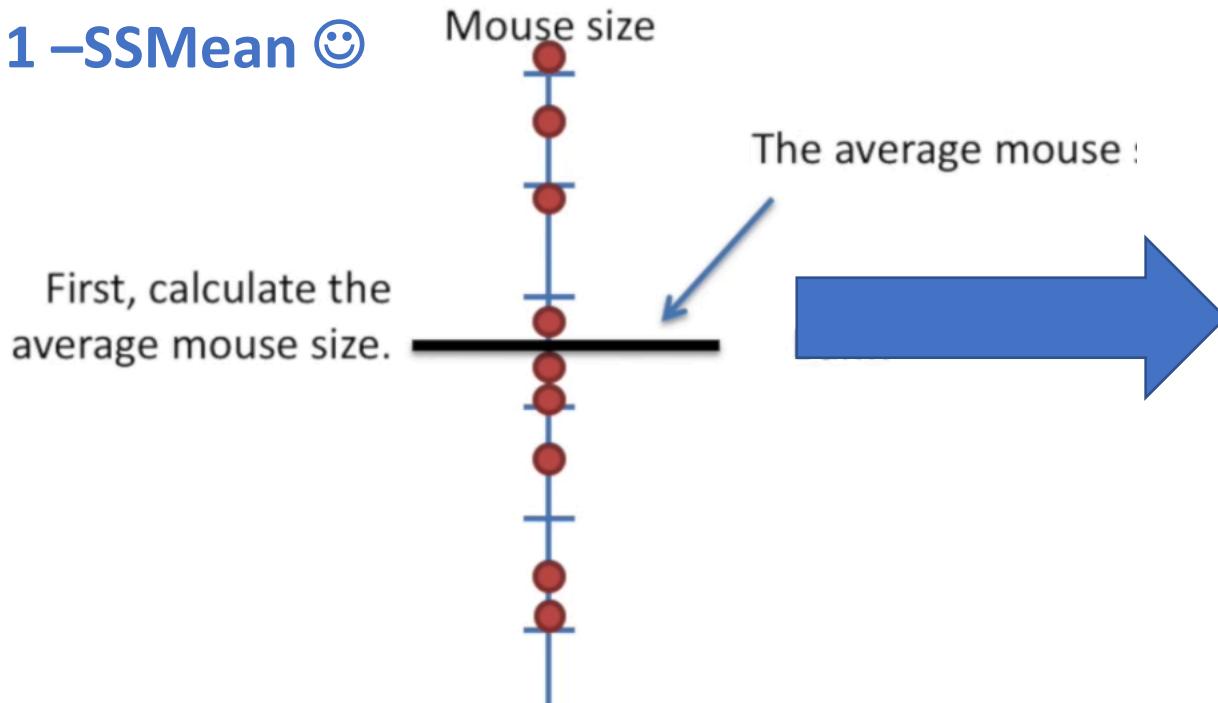
- Evaluation du modèle obtenu!!!!

→ A quel point l'estimation du « Size Mouse » sera correcte avec une nouvelle donnée «mouse weight »?

→ Coefficient R² permet de répondre à la question: how good is the model to predict Mouse size taking into account Mouse weight!!

R² Coefficient de détermination

1 –SSMean ☺



Sum of the Squares around the Mean = SS(M)

$$SS(\text{mean}) = (\text{data} - \text{mean})^2$$

$$\text{Var}(\text{mean}) = \frac{SS(\text{mean})}{n}$$

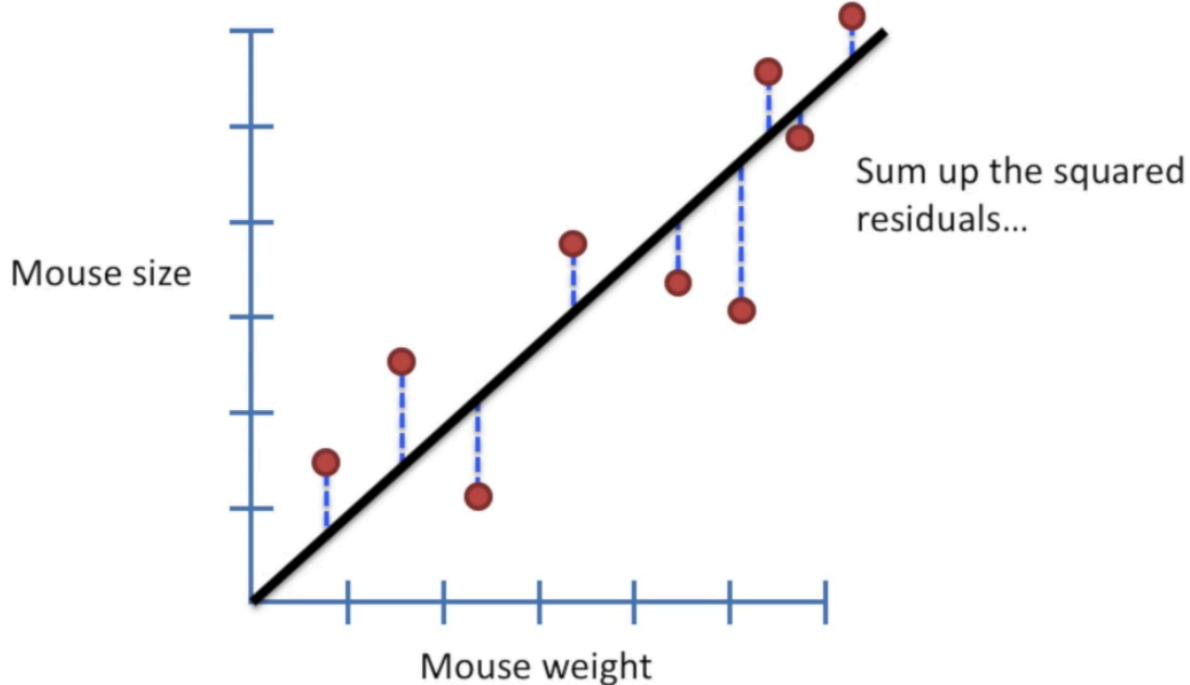
= Modèle sans prédicteur = Variation Totale

→ Pas de prise en compte de l'effet du paramètre weight

R²

Rappel « best fit » (après plusieurs rotations)

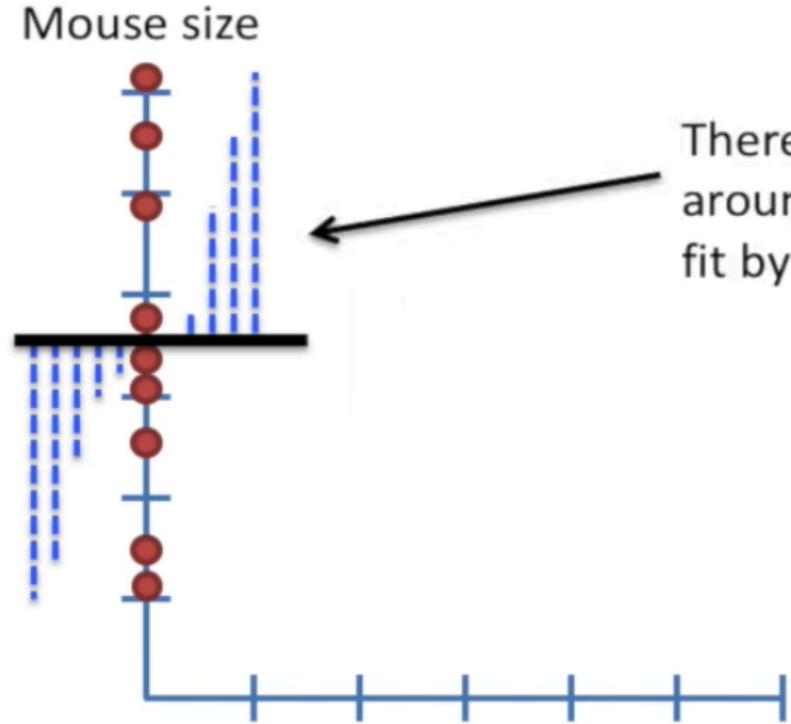
2 –SSfit



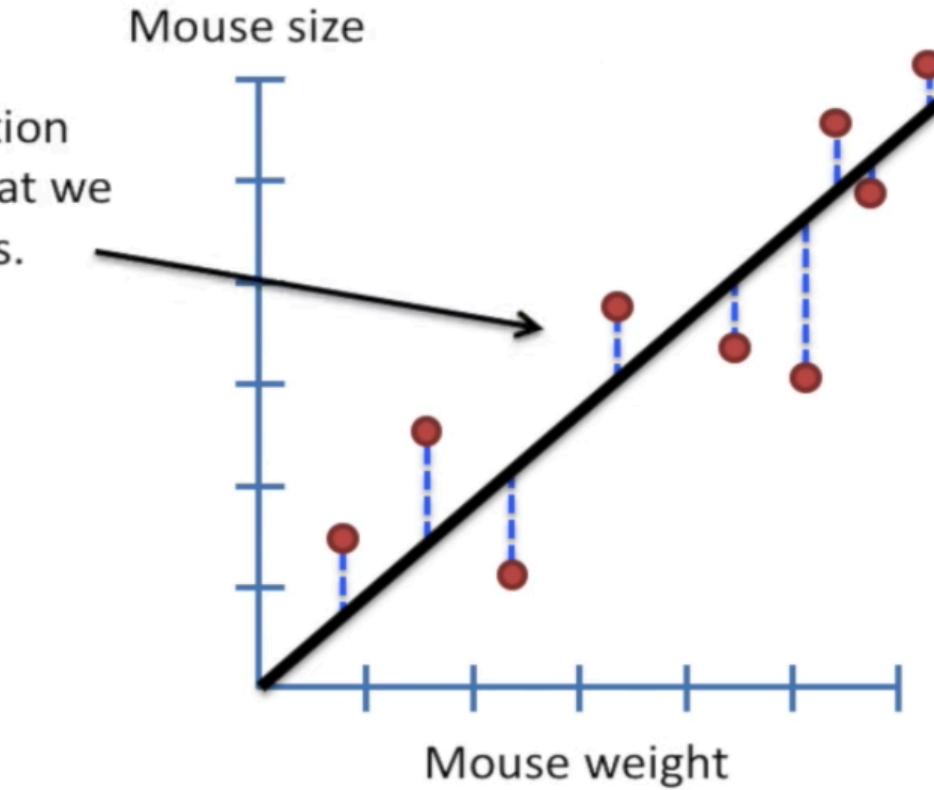
SSfit = Somme des écarts carrés à la droite de régression

$$SSfit = (data - \text{droite reg})^2$$

$$\text{Var(fit)} = \frac{(data - \text{line})^2}{n}$$



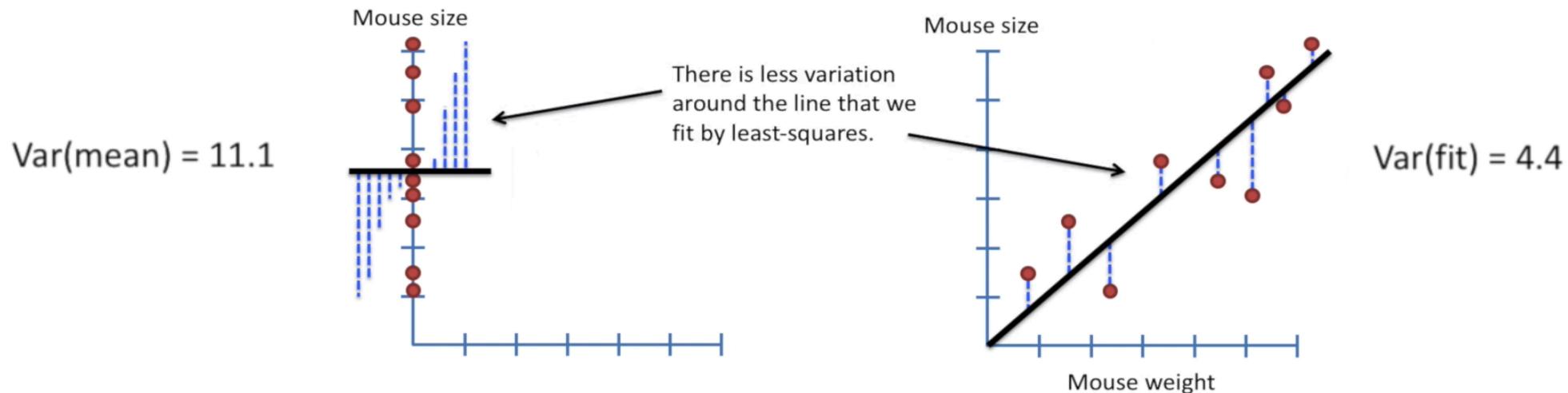
There is less variation around the line that we fit by least-squares.



- Prenant en compte la variable « poids souris », moins de variations ($SS_{fit} < SS_{Mean}$)!
- La variation « taille souris » peut être expliquée en prenant en compte le « poids souris »!
→ Mais dans quelle mesure?

R² = % de variation la variable réponse expliquée par un modèle linéaire (variable poids)

$$R^2 = \frac{\text{Var(mean)} - \text{Var}(fit)}{\text{Var(mean)}}$$



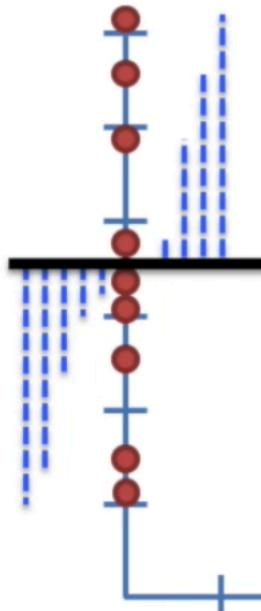
$$R^2 = \frac{11.1 - 4.4}{11.1} = 0.6 = 60\%$$

$$R^2 = \frac{\text{Variation expliquée}}{\text{Variation totale}}$$

→ Le modèle établit explique 60% de la variabilité/variance de la variable « Mouse size »
→ R² compris entre 0 et 1

TO be sure ...

$$\text{Var(mean)} = 11.1$$

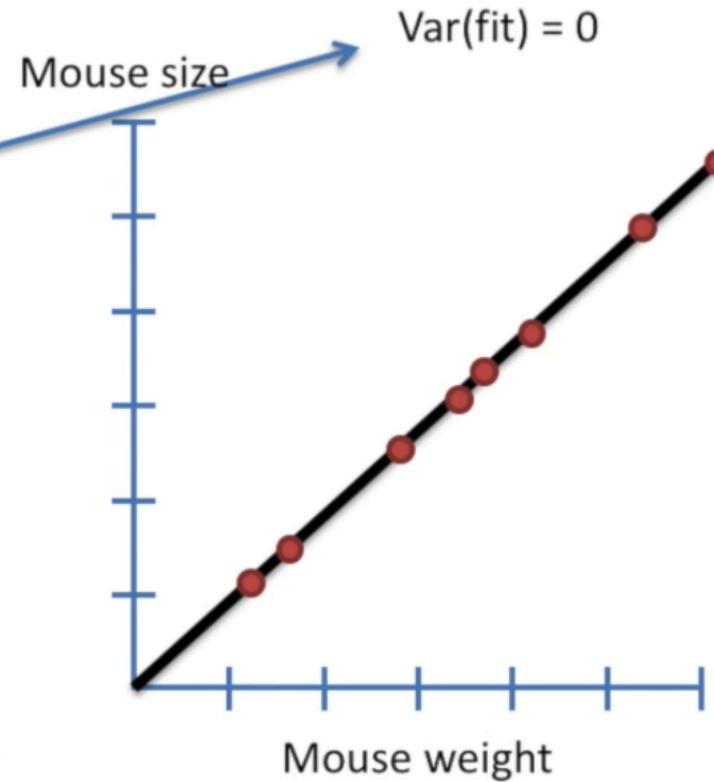


$$R^2 = \frac{\text{Var(mean)} - \text{Var(fit)}}{\text{Var(mean)}}$$

$$R^2 = \frac{11.1 - 0}{11.1}$$

$$R^2 = 1 = 100\%$$

Pas de variation non expliquée par X
Pour chaque valeur X on a la valeur exact de Y



R² statistiquement significatif?

- Besoin d'une p-value...
- Variance ... donc p-value à partir du ratio F et de la distribution F

$$F = \frac{\text{The variation in mouse size explained by weight}}{\text{The variation in mouse size not explained by weight}}$$

Plus F est grand, meilleure est le modèle!
De F à la p-value...

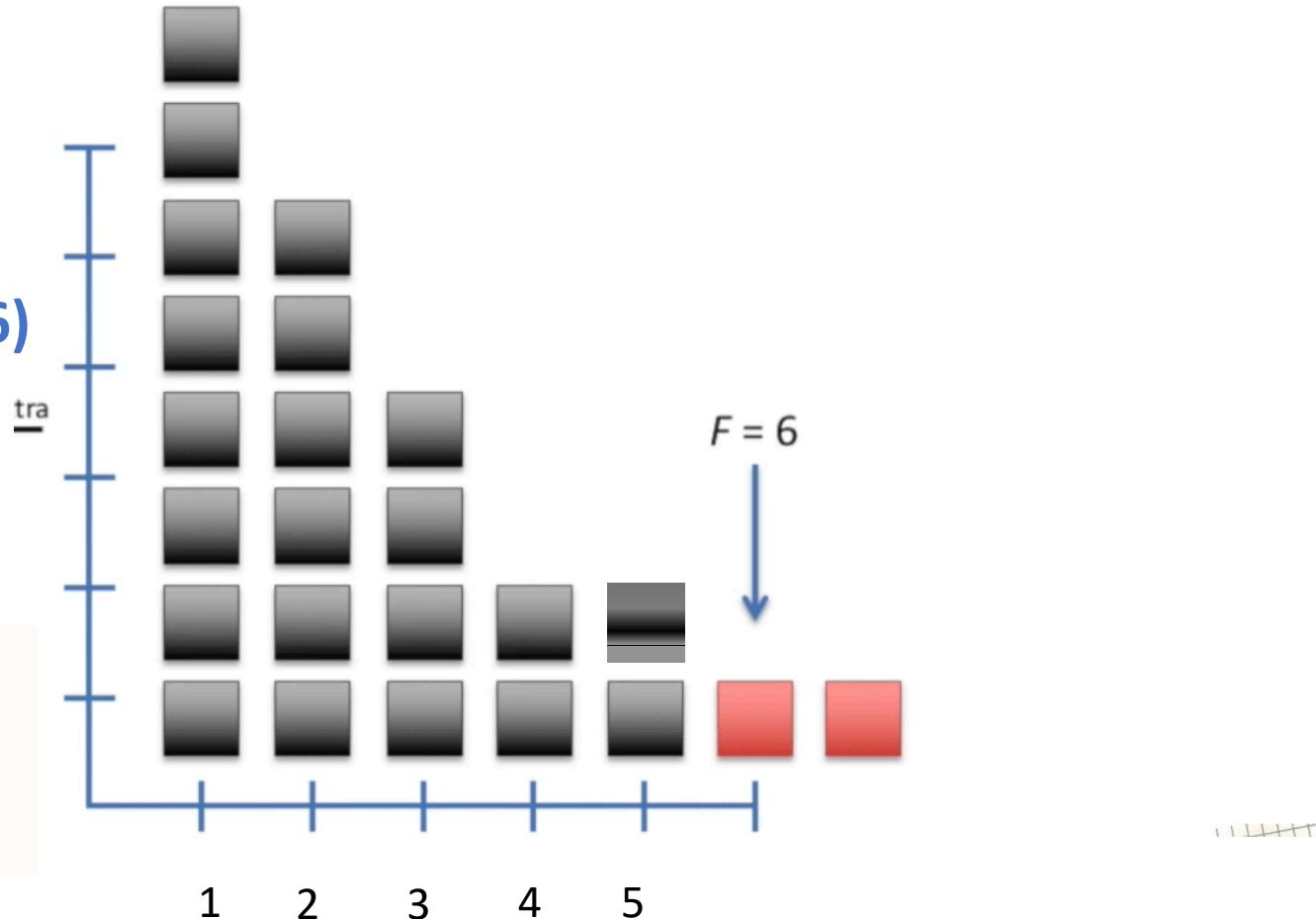
La p-value dérive de la distribution F

- Sous sets aléatoires à partir des données
- Calcul de F pour ces sous sets
- Calcul du F pour les données initiales ($F=6$)
-> génère la distribution F

P-value :

- Probabilité d'observer le résultat ($F=6$)
- Probabilité d'observer quelque chose d'équivalent
- Probabilité d'observer quelque chose de plus extrême

$$p\text{-value} = P_{F6} + P_{\text{equal}} + P_{\text{more extreme}}$$

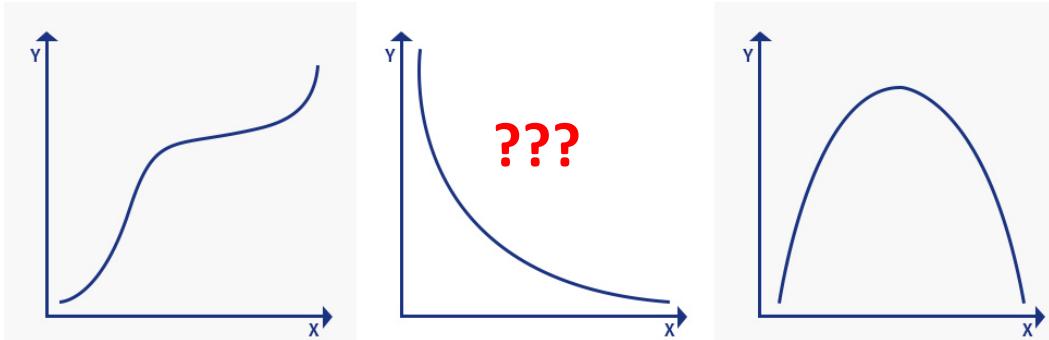


What is the p-value score? 😕

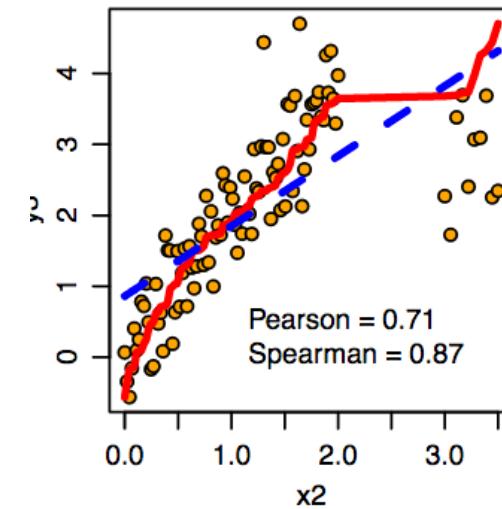
Association: coefficient de corrélation r

Intensité & direction de l'association entre deux variables (i.e. param)

- Relation Linéaire stricte : Pearson (**r**, paramétrique) = covariance
- Relation Monotone : Spearman (**Rho**, non paramétrique, basée sur les rangs)
- Relation Monotone : Kendall (**Tau**, non paramétrique) : Alternative à Spearman (petit échantillon)



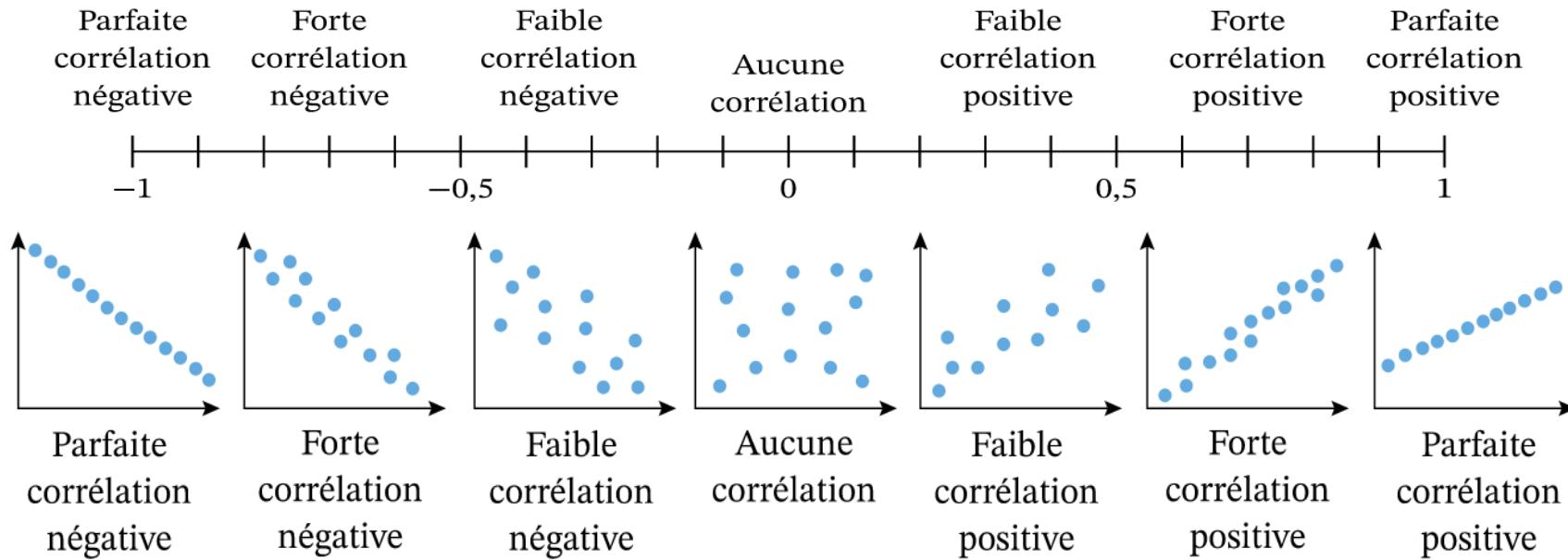
Inspecter ces données n'est jamais inutile...



Coefficient r varie entre -1 et 1

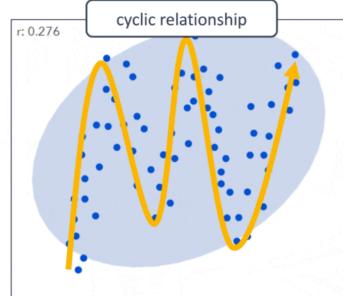
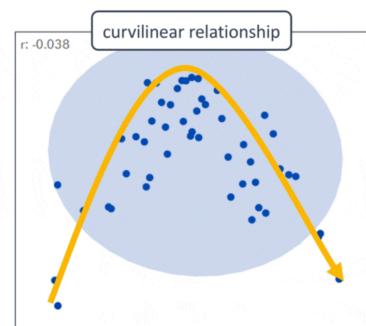
- Corrélation positive: Les valeurs des deux variables tendent à augmenter ensemble
- Corrélation négative: Les valeurs d'une variable tend à augmenter et les valeurs de l'autre variable diminuent
- Zéro : pas d'association **LINEAIRE** (Pearson)

A titre Indicatif!!!



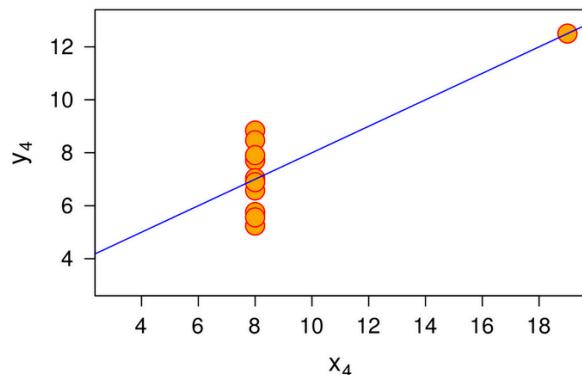
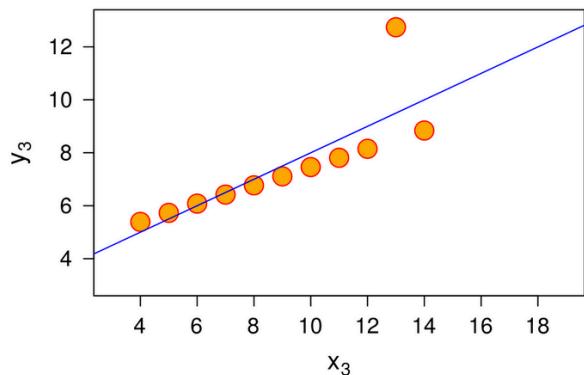
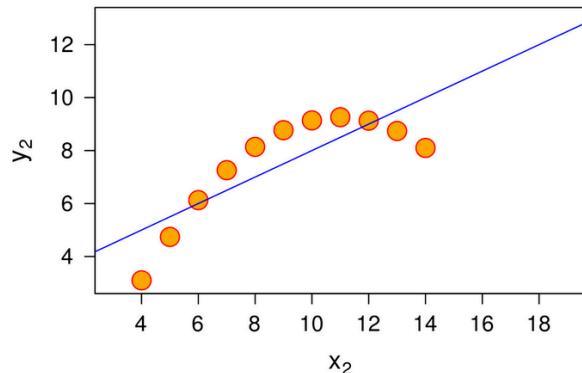
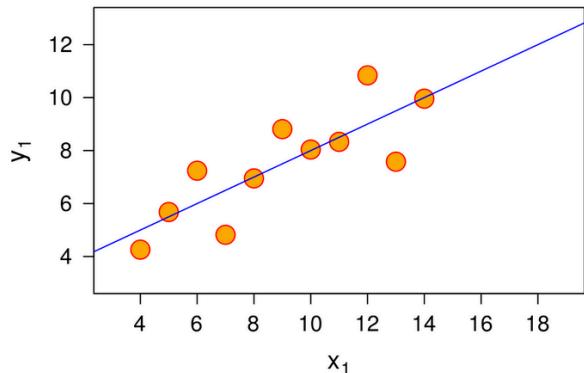
Car inspecter ses résultats n'est jamais inutile...

- r proche de Zéro : pas d'association??



Jamais inutile... Le quartet d'Anscombe...

- 4 jeux de données
- Mêmes propriétés stat



Propriété	Valeur
Moyenne des x	9,0
Variance des x	10,0
Moyenne des y	7,5
Variance des y	3,75

$r = ?$



- Loi de **distribution de r** sous l'hypothèse H_0 : Aucune liaison statistique entre X et Y
→ Accès aux **p-values**

Relation entre r et R^2

Le coefficient de **corrélation de Pearson r** peut avoir une interprétation conjointe à la régression linéaire simple **son carré étant la variance expliquée par la régression (R^2)**

Si $r = 0.5 \rightarrow R^2 = 0.25 \rightarrow 25\%$ de la variance de Y expliquée par X... ☹

R^2 : part de la variance de la variable dépendante qui provient de celles des variables indépendantes

Multiple Testing Issue: increasing the risk...

Test is based on **probabilities**, so there is always a risk of drawing the **wrong conclusion!**

→ No hypothesis test is 100% reliable

Performing hypothesis testing:

- You have two hypotheses :
- H0: Null hypothesis = the reference hypothesis : No difference
- H1: Alternative hypothesis: There is a difference



- You encounter: Type I error : α = Risk alpha

$\alpha = 0.05$ Is the **probability** (significance threshold) to incorrectly **reject H0!**
In other words, an acceptable chance of a false positive!!

Differential abundance : Multiple testing!!

ONE TEST :

$$P_{\text{False Positive}} = P_{\text{error}} = \underline{\alpha} = 0.05$$

Complementary Prob

$$P_{\text{no_error}} = 1 - \underline{\alpha} = 0.95$$



time to
RISK

TWO TEST without making error : $P_{\text{no_error in two tests}} = (1 - \underline{\alpha}) * (1 - \underline{\alpha}) = (1 - \underline{\alpha})^2$

Complementary Prob

$$P_{\text{at_least_ONE_error in two tests}} = 1 - (1 - \underline{\alpha})^2$$

Generalization to n TESTS

$$P_{\text{at_least_ONE_error in } n \text{ tests}} = 1 - (1 - \underline{\alpha})^n$$

It's called the global $\underline{\alpha}$ risk

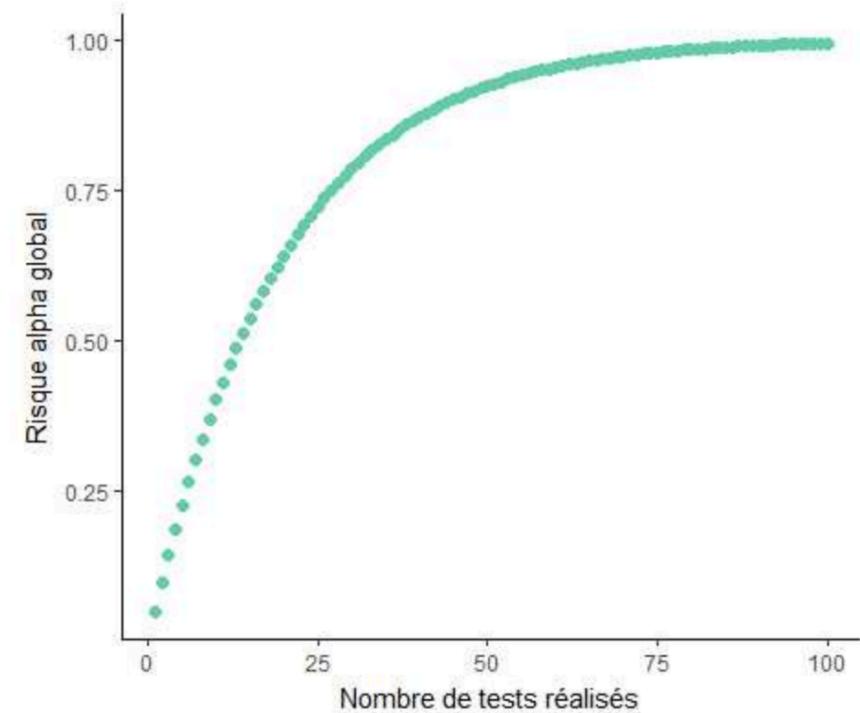
What does it means...



- You test **ONE** ASVs ($n=1$) for differential abundance: $1-(1-\alpha)^n = 1-(1-0.05)^1 = \textcolor{yellow}{0.05}$
- You test **3** ASVs ($n=3$): $1-(1-0.05)^3 = \textcolor{yellow}{0.14}$
- You test **100** ASVs ($n=100$): $1-(1-0.05)^{100} = \textcolor{yellow}{0.9941}$

The global risk α reach $0.9941=99.41\%!!!!$

→ 99% to wrongly reject the H0 at least
One times



Need to adjusted this phenomena by using p-value **adjusted!**

FDR : False Discovery Rate : Benjamini-Hochker

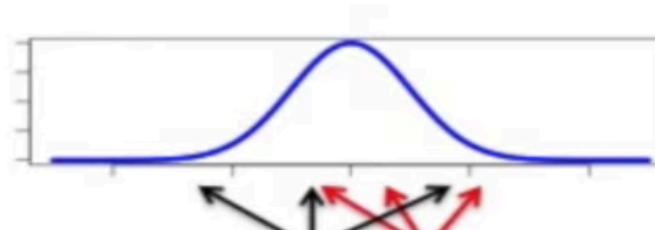
The idea : Discard bad data that looks good!!!

Benjamini-hocherk **adjusts p-values**
to limit the number of **false positives**
that are **reported as significant (pvalue < 0.05)**

Adjusts p-values
means that it makes them **larger!**

Using FDR cutoff < 0.05
means less than 5% of the significant results will be false positives

Mathematical approach FDR-Benjamini-Hochker



10 pairs of samples taken from the same distribution. (i.e. 10 genes that were not effected by the drug).

p-values: 0.91 0.11 0.71 0.31 0.51 0.41 0.61 0.21 0.81 0.01

Notice that one of the p-values is a false positive (that is to say, less than 0.05)

p-values

0.01	0.11	0.21	0.31	0.41	0.51	0.61	0.71	0.81	0.91
------	------	------	------	------	------	------	------	------	------

smallest largest

1- Ranking pvalue

Prepare space for adjusted pvalue

Let's make spaces for the FDR adjusted p-values.

p-values:	0.01	0.11	0.21	0.31	0.41	0.51	0.61	0.71	0.81	0.91
rank:	1	2	3	4	5	6	7	8	9	10

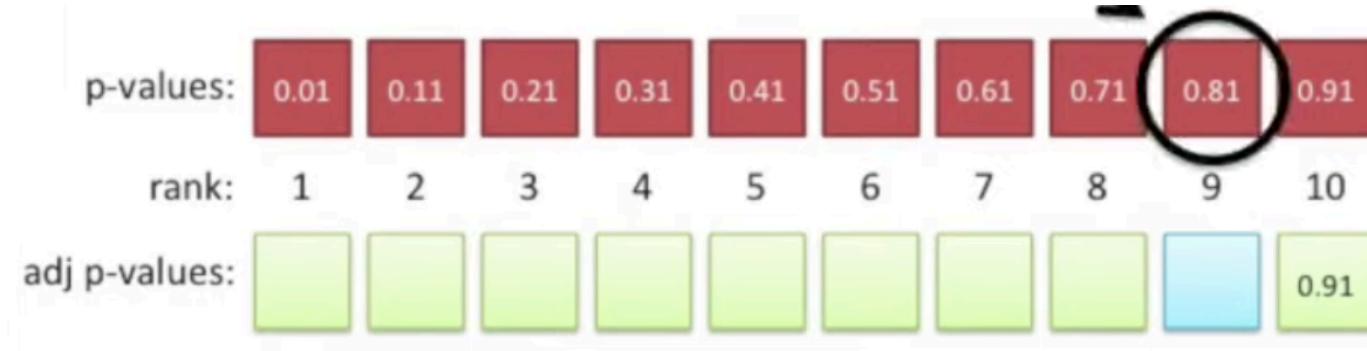
adj p-values:										
---------------	--	--	--	--	--	--	--	--	--	--

p-values:	0.01	0.11	0.21	0.31	0.41	0.51	0.61	0.71	0.81	0.91
rank:	1	2	3	4	5	6	7	8	9	10

adj p-values:										0.91
---------------	--	--	--	--	--	--	--	--	--	------

2- Largest adjusted pvalue and larger pvalue are same

Next adjusted pvalue

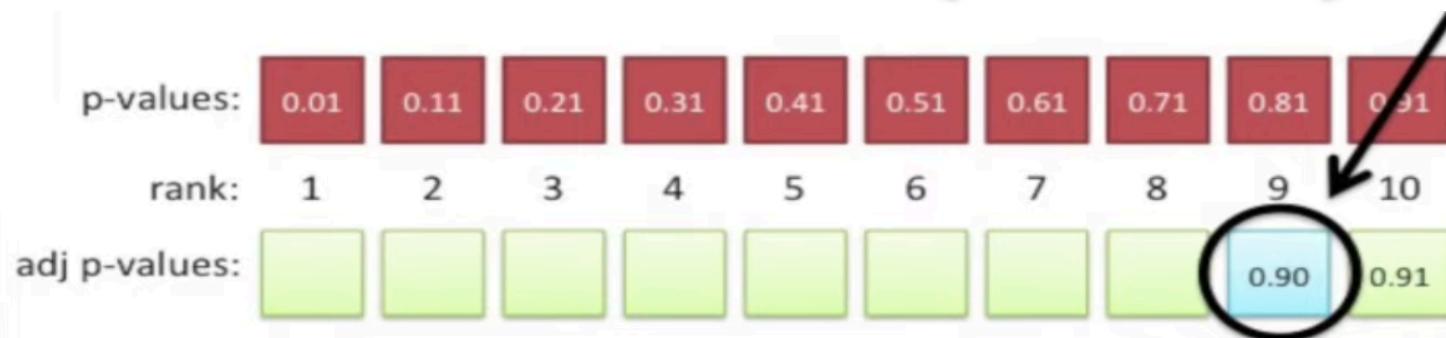


The smallest of the two options

$$\text{b: the current p-value} * \left\{ \frac{\text{total \# of p-values}}{\text{p-value rank}} \right\}$$

a: The previous adjusted p-value = 0.91

$$\text{b: } 0.81 * \left\{ \frac{10}{9} \right\} = 0.90$$



Finally...

The false positive p-value... is no longer significant.

