Test d'hypothèses
Corrélation et régression
sous forme d'analyses bivariées

25 & 26
Octobre 2023

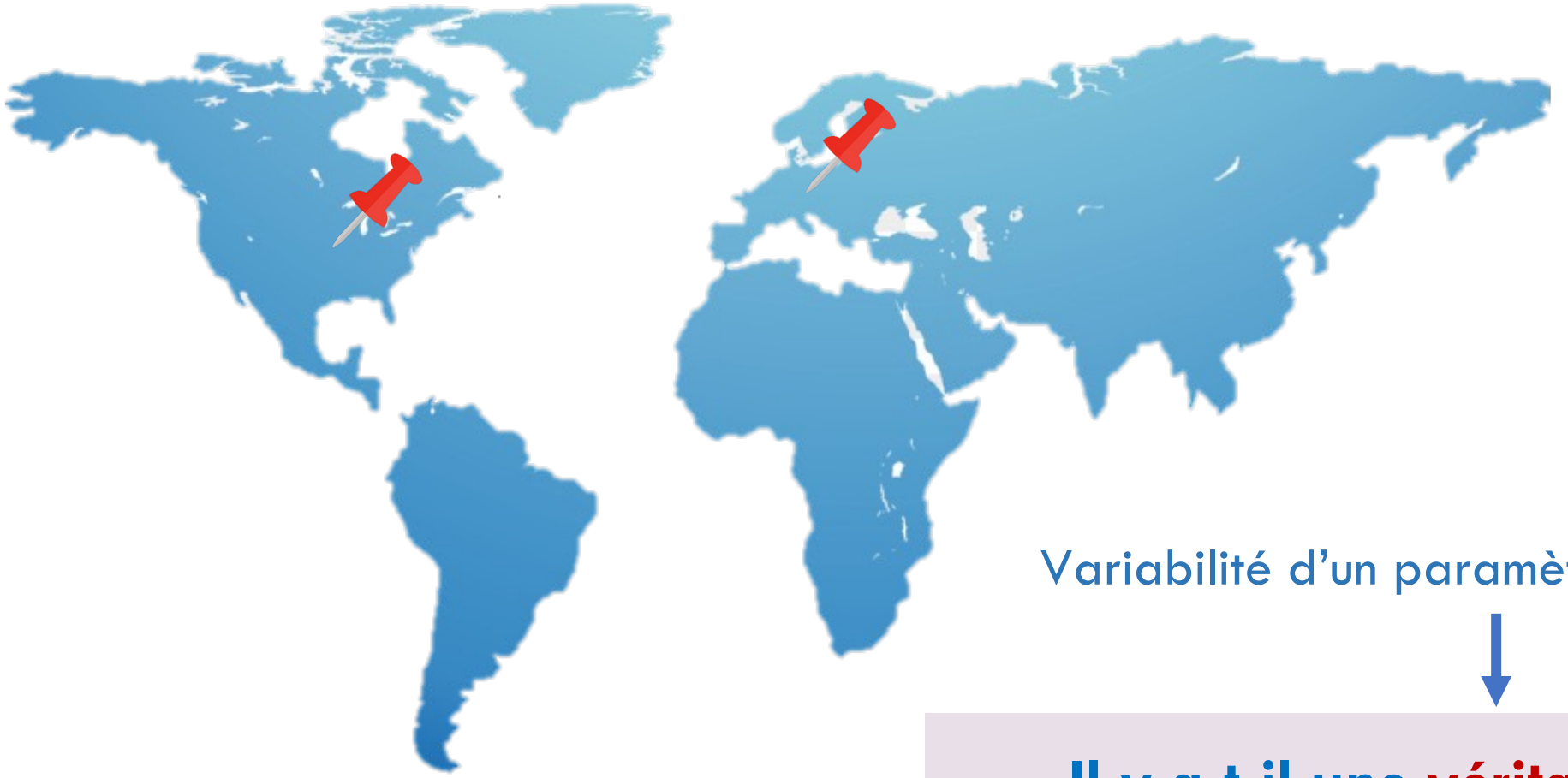
PFS IRD

Analyse de données biologiques sous R

Institut Méditerranéen d'Océanologie

Institut de Recherche pour le Développement
FRANCE

ISSBAT



Variabilité d'un paramètre entre les USA et l'UE

Il y a-t-il une véritable différence significative ou est-ce le fruit du hasard ?

Les statistiques peuvent répondre à notre question!!

Population VS échantillon

Population: ensemble d'individus ou d'objets de même nature (très grand ou infini)

→ On ne peut pas étudier une population en entier : en statistique, on étudie donc un nombre limité d'individus, une partie de la population : un échantillon.

→ On cherche à déduire des propriétés de la population à partir de l'échantillon

→ Si on veut étudier la variabilité d'une variable / caractéristique d'intérêt de la population, il faut avoir un échantillon représentatif (tiré de façon aléatoire)

Dans une population, on peut mesurer un caractère : une variable qui est le résultat d'un phénomène aléatoire

- Qualitative
- Quantitative (continue)

Une loi de probabilité décrit le comportement aléatoire d'un phénomène dépendant du hasard

Dans une population, on peut mesurer un caractère : une variable qui est le résultat d'un phénomène aléatoire

- Qualitative
- Quantitative (continue)

Une loi de probabilité décrit le comportement aléatoire d'un phénomène dépendant du hasard

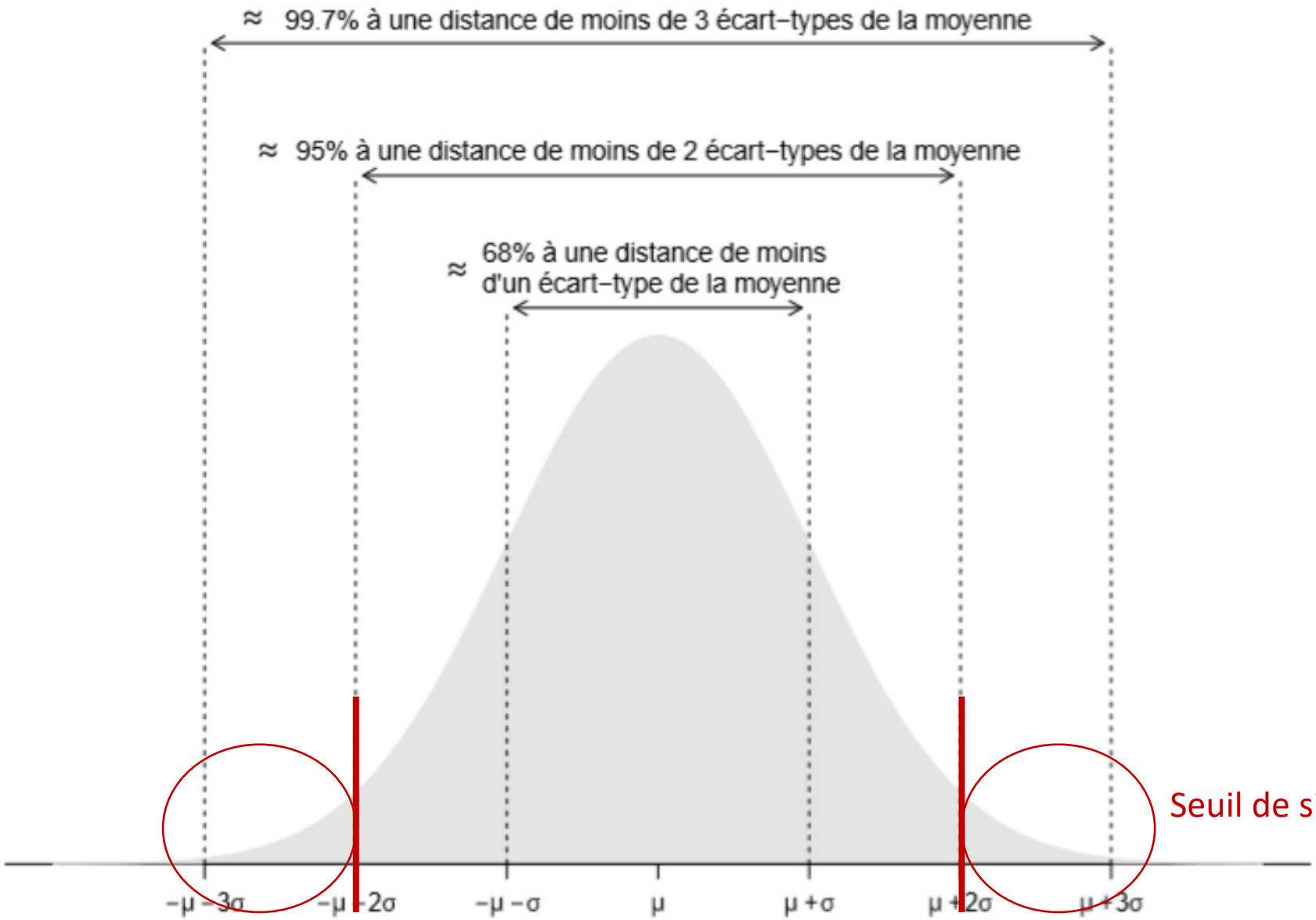
La loi normale

Si on dispose de 1000 échantillons issus d'une variable suivant une loi normale, et que l'on trace le nombre d'échantillons égaux à chaque valeur, on obtient une courbe en "cloche" / une distribution gaussienne :

$X \sim N(\mu,\sigma 2)$ avec μ and $\sigma 2$ les paramètres de cette distribution:

- μ: moyenne
- σ: écart type = dispersion des valeurs autour de la moyenne

Répartition des valeurs autour de la moyenne

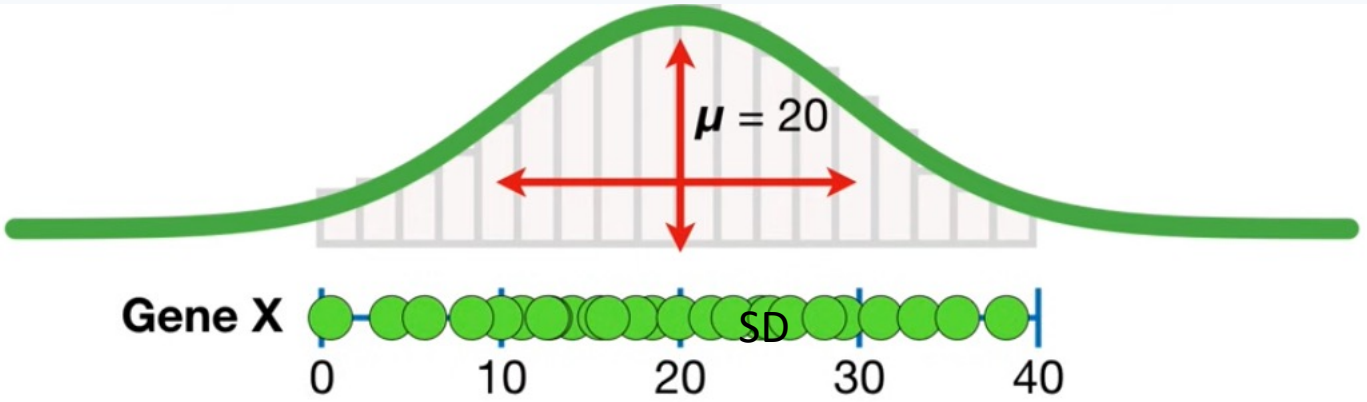


Seuil de significativité

Statistiques descriptives (analyses univariées)

Décrire, afficher et résumer vos données

- **Tendances générales** (moyenne, mediane…)
- **Dispersion** (variance, écart type)
- **Fréquence de distribution** (count, relative, cumulative)

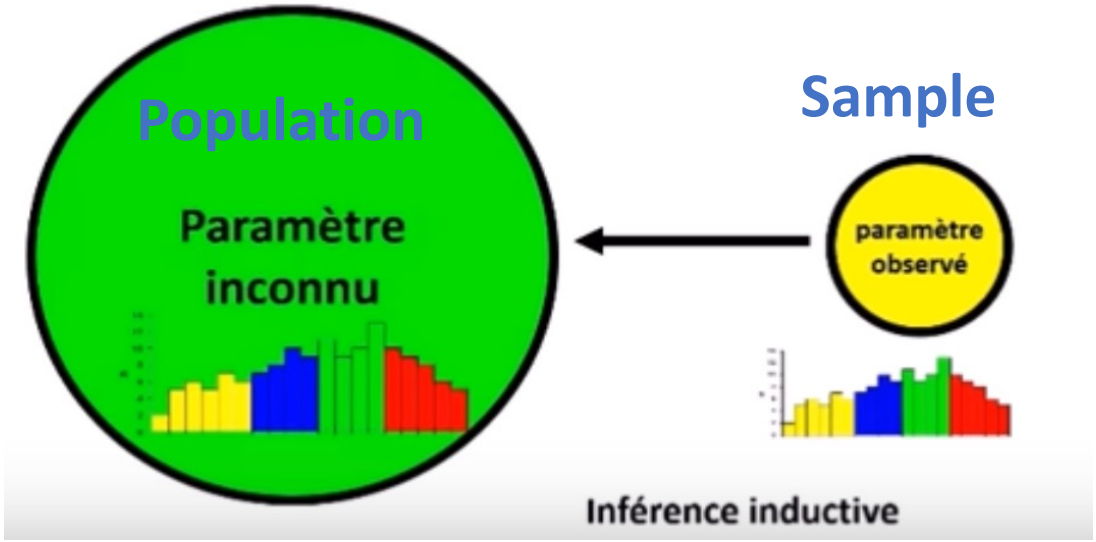


Identifier les caractéristiques de chaque variable de vos données

→ **Permet de formuler des hypothèses et guider vos analyses statistiques**

Statistiques inférentielles

Predictions - Generalisations



Make inferences about the population

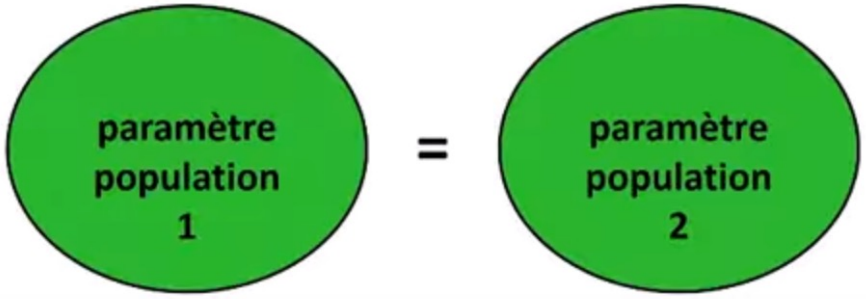
- Comment puis-je utiliser mon échantillon pour faire des prédictions sur la population ?
  = **Estimation**
- Comment prouver une théorie sur le comportement de mes données (comparaison) ?
- = **Hypothesis Testing**

Comparaison de moyennes et Hypothèses

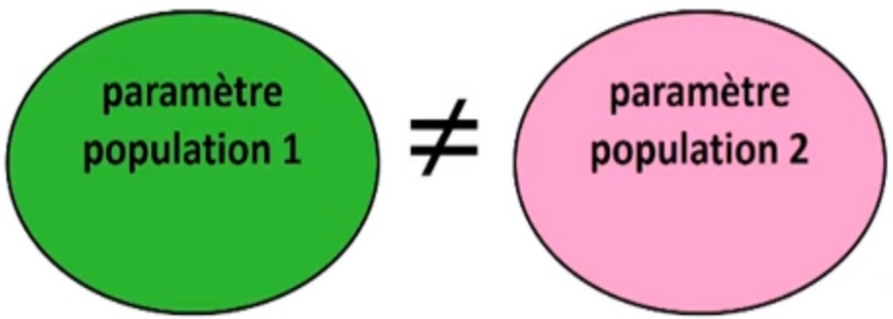
Essayer de valider une hypothèse relative à un paramètre de la population à partir d'un échantillon de comparaisons

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

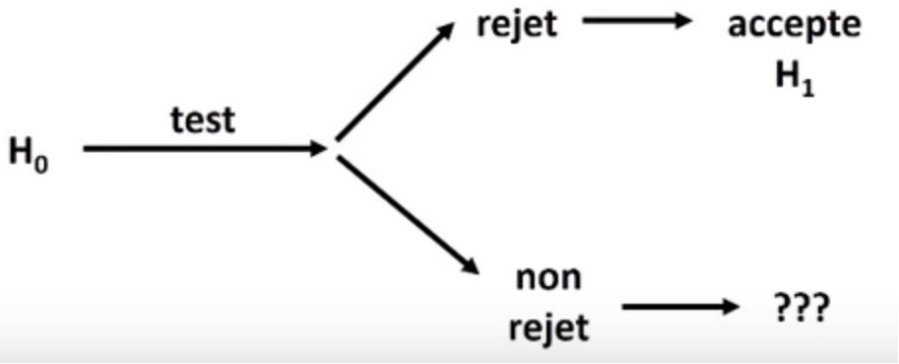
Null Hypothesis H0



Alternative hypothesis H1

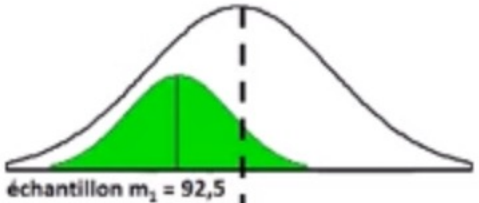


We are testing the null hypothesis!



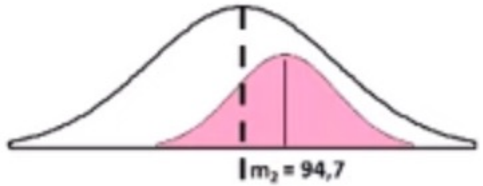
Comparaison de moyennes et Hypothèses

Si H0 vraie… pas de différences



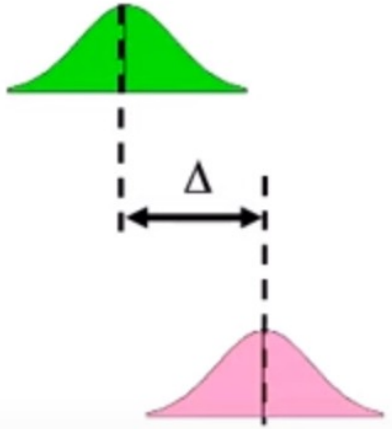
μ_1

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



μ_2

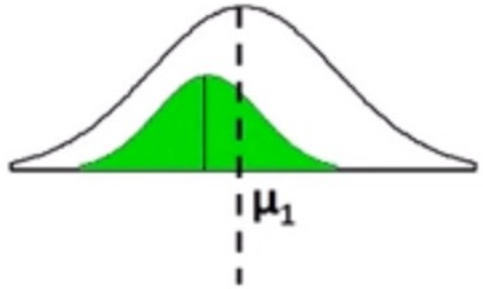
$H_0 : \mu_1 = \mu_2$



Même distribution

→ Fluctuation d'échantillonnage

Si H0 rejetée, H1 acceptée



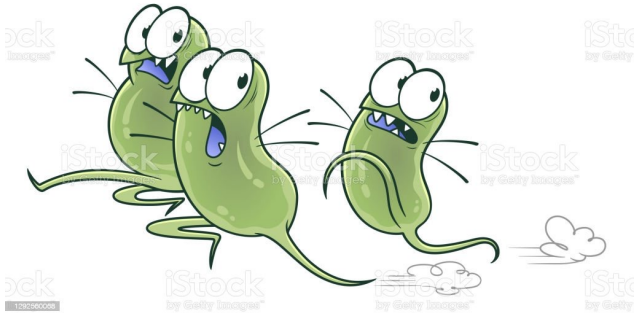
μ_1

μ_2

$H_1 : \mu_1 \neq \mu_2$

2 distributions différentes

Problème d'inférences : soumises à des erreurs !
Le risque est lié au résultat du test d'hypothèse
A cause de votre échantillonnage !

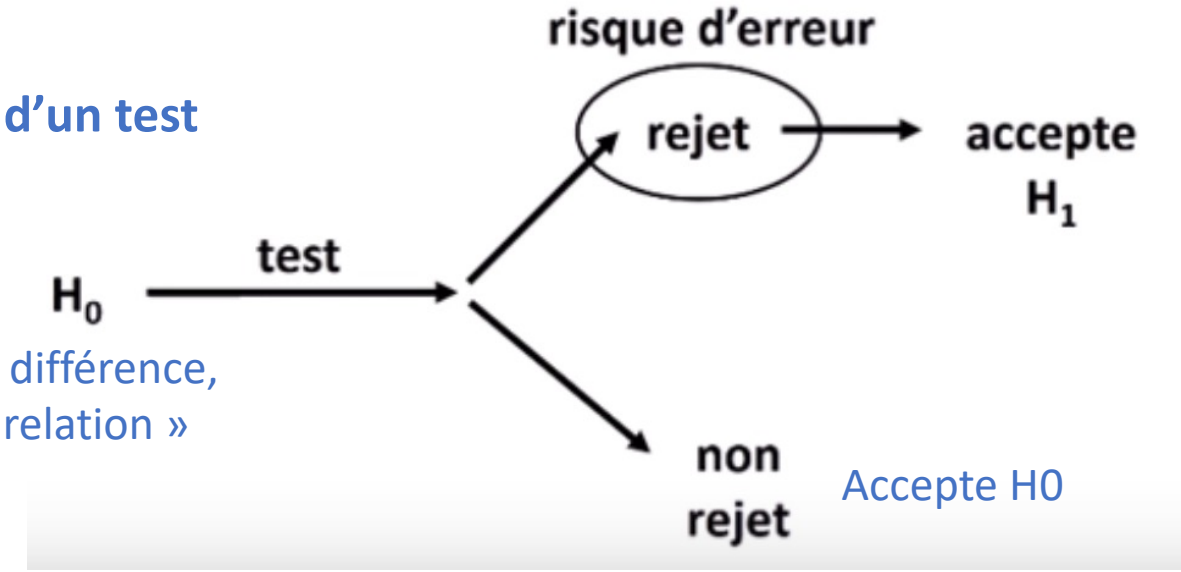


Le risque d'erreur de première espèce: α

- C'est une probabilité entre 0 et 1, ou 0% et 100%

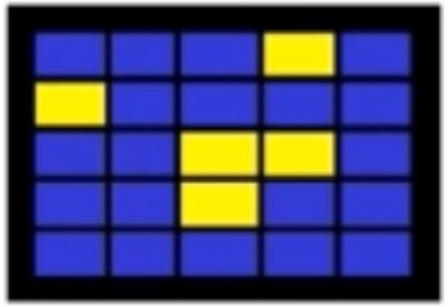
- C'est lorsqu'on affirme une différence alors qu'il n'y en a pas (=Faux positif)!!

Principe d'un test

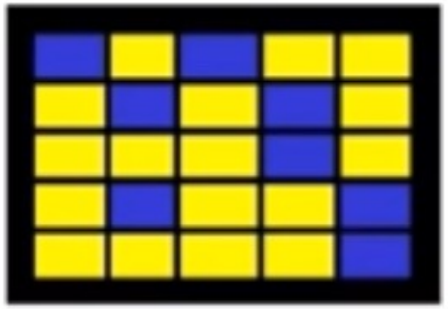


α= Risque de rejeter H0, si H0 est vrai

N= 25 carreaux
→ 80% Bleu



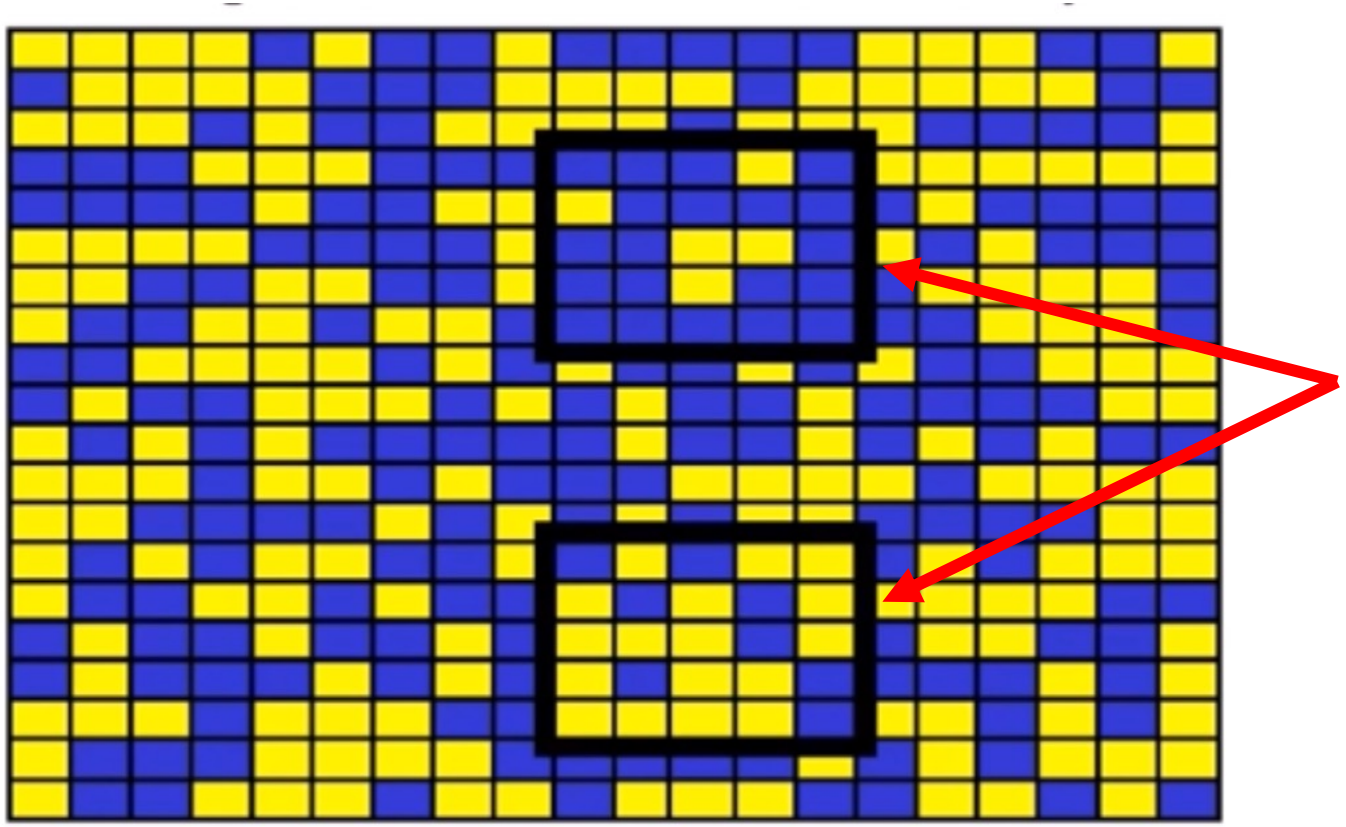
N= 25 carreaux
→ 32% Bleu



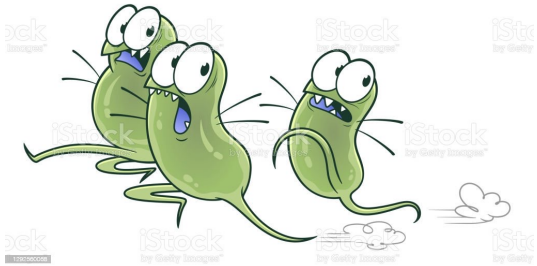
Les deux échantillons proviennent t-il du même dallage (même distribution)?

- Test stat non (rejet H0)…
- mais …

Provient du même dallage/population (50% bleu, 50 % jaune)!!



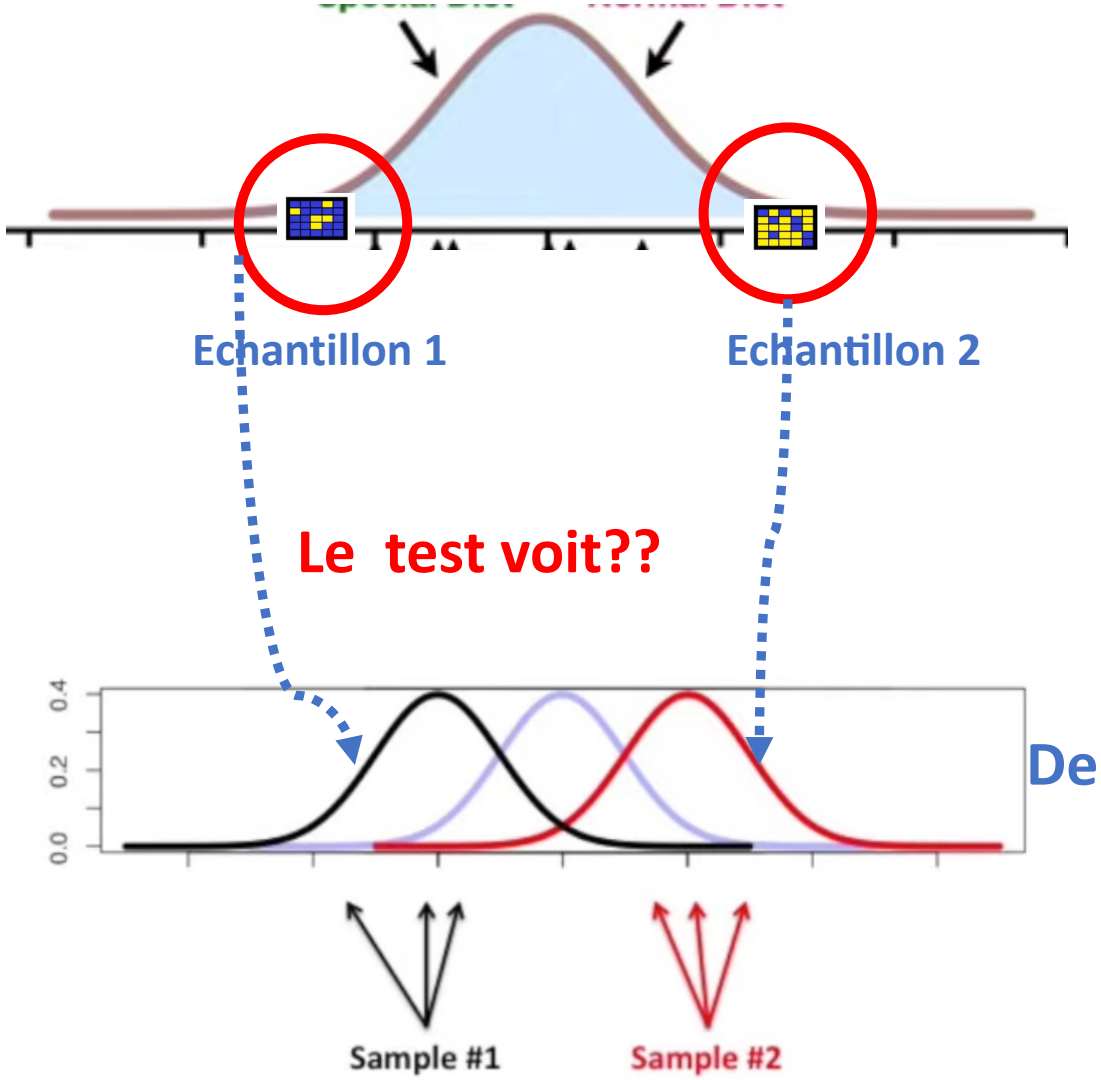
Risque faible de tomber sur ce type d'échantillons (rare)



Conclure sur la base de nos échantillons qu'ils provenaient de deux populations différentes

--> Erreur de type I

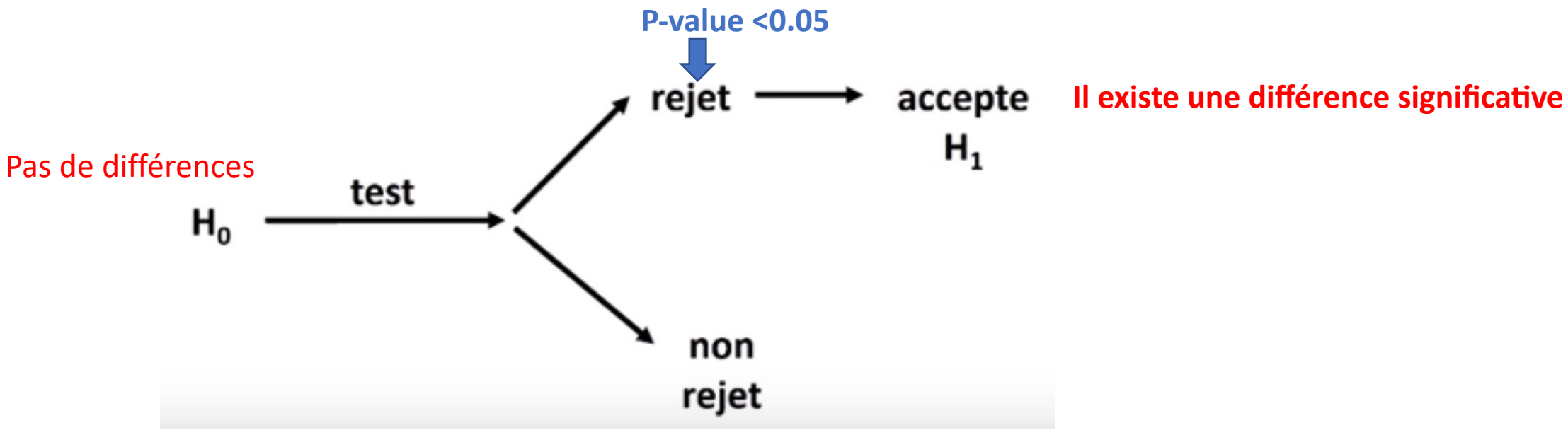
Données proviennent d'une même distribution mais…



Deux distributions différentes…

- Risque α est choisi AVANT le test : **Seuil de significativité**

- α souvent positionné à 5% (rejeter à tort H0)
- En sciences, le "presque aucune chance" se traduit par
  dans moins de 5% des cas ou Ho est vraie = **pvalue < 0.05**



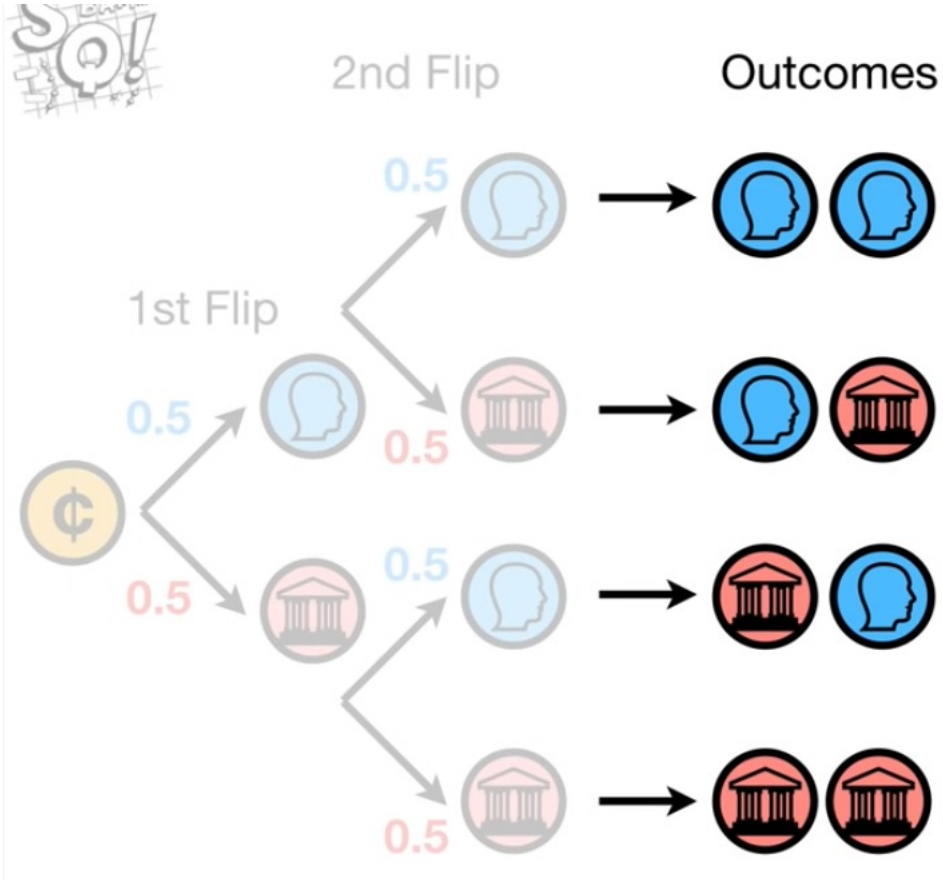
Concept de la p-value…

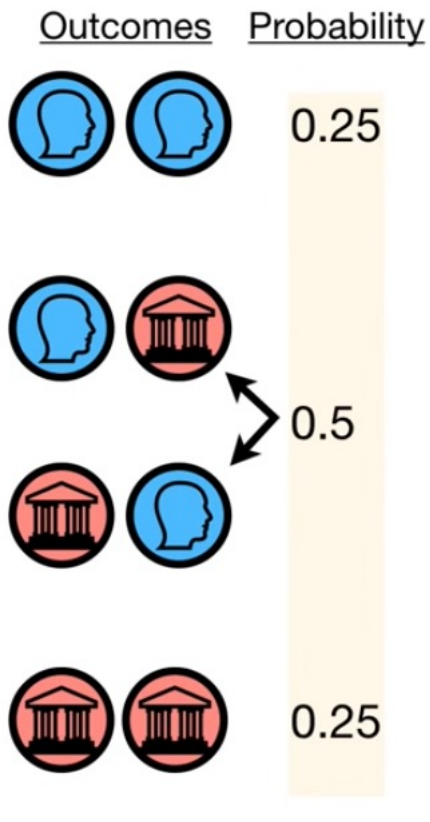
Ma pièce est spéciale : Pile deux fois de suite !

L'hypothèse nulle H0 : même si j'ai obtenu deux Pile d'affilée, ma pièce n'est pas différente d'une pièce normale !

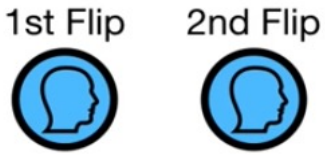
Une petite p-value nous dira de rejeter H0 (p-value <0.05)!

Testons cette hypothèse en calculant la p-value!





$$\frac{\text{The number of times we got } \textbf{2 Heads}.}{\text{The total number of outcomes.}}$$

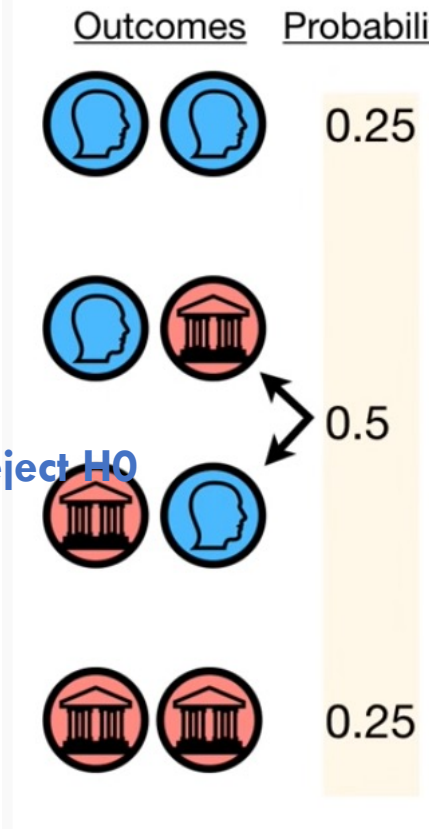


A ***p-value*** is composed of three parts:

1) The probability random chance would result in the observation.

2) The probability of observing something else that is equally rare.

3) The probability of observing something rarer or more extreme.



A small p-value will tell us to reject H0 (p-value <0.05)!

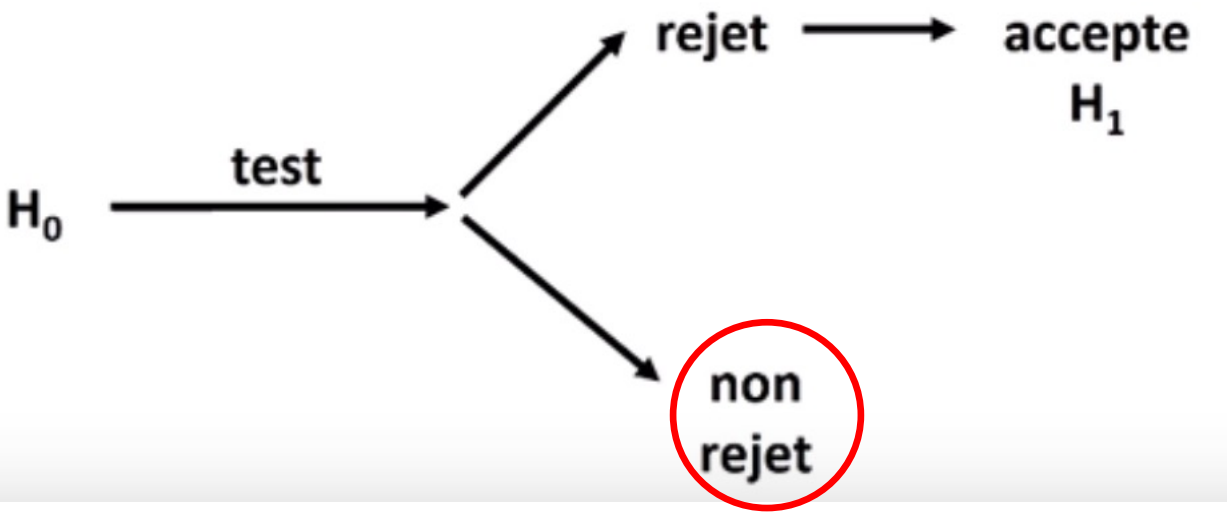
Nothing

P- value for 2 Heads (Sum of three parts)= 0.25+ 0.25 + 0 = 0.50!

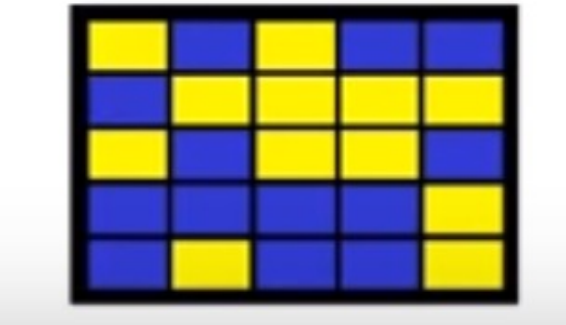
My coin is not special! p-value >>> 0.05!!!

Le risque d'erreur de deuxième espèce: β

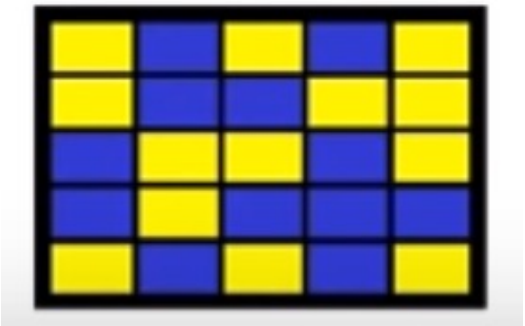
Erreur de type II : On ne conclut pas à une différence alors qu'il y en a 1 (=« Faux Négatif »)
→ Probabilité de ne pas rejeter H0, si H1 est vrai



On ne sait pas calculer le risque B



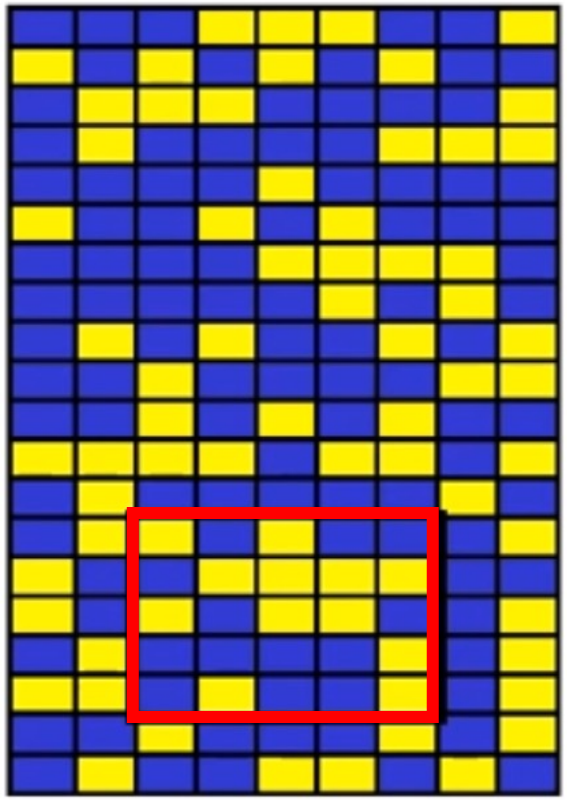
48% de bleu



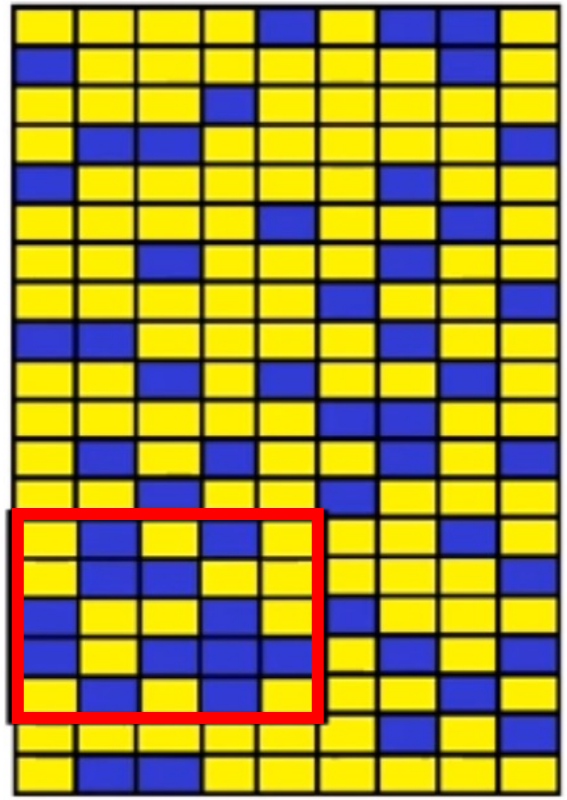
52% de bleu

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

60% de bleu



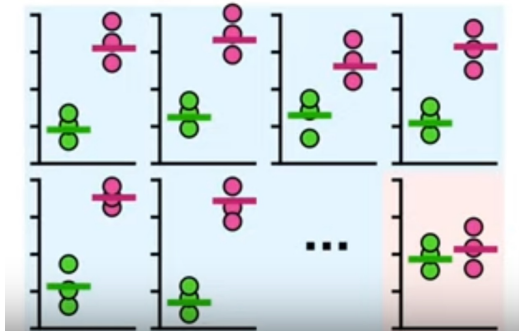
30% de bleu

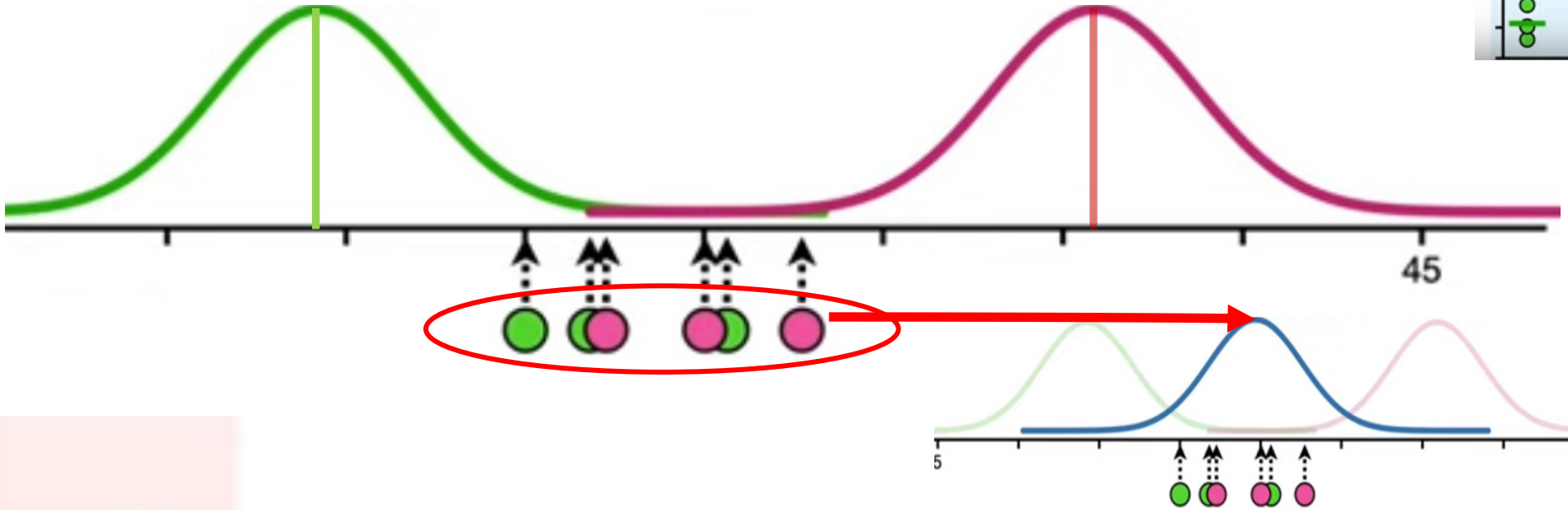


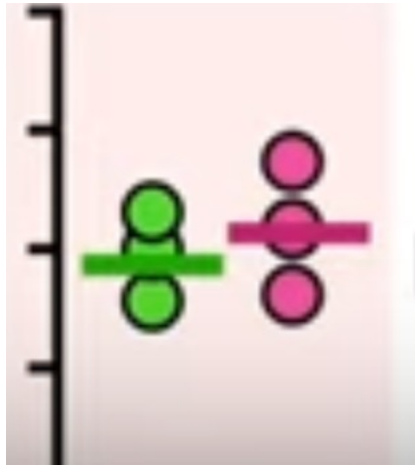
- 2 dallages différents = 2 populations différentes, H0 devrait être rejetée

Mais ca n’aurait pas été le cas lors du test avec notre échantillonnage…

Mais dans certains cas…





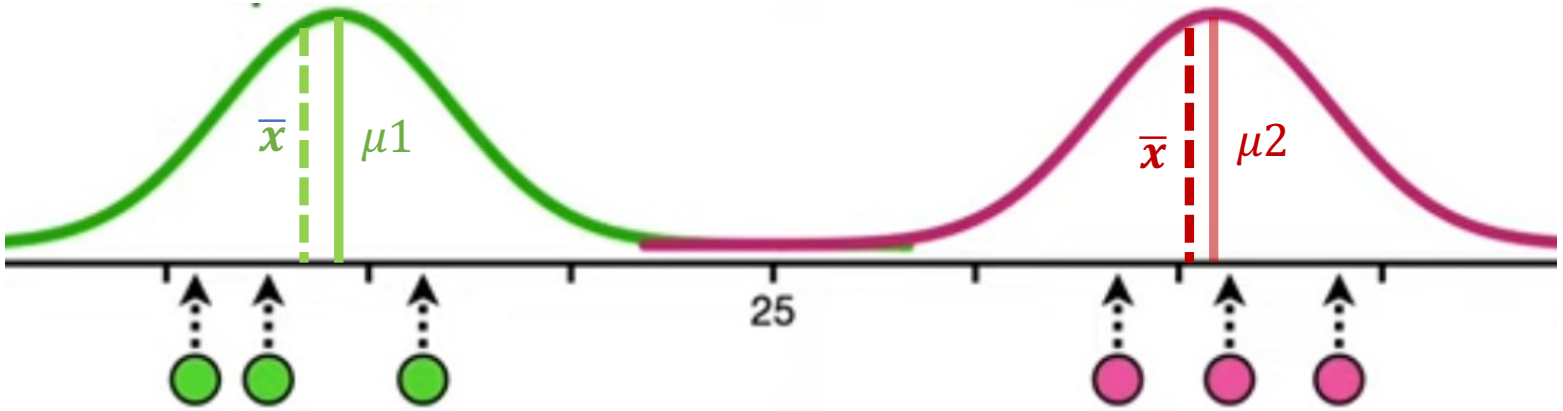


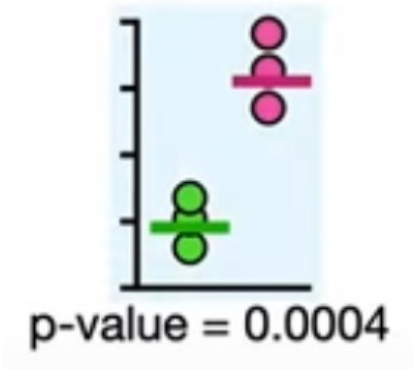
p=0.23!!!

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

→ Impossible de rejeter correctement H0 …

Scientifiquement … Cas échantillonnage représentatif de la population





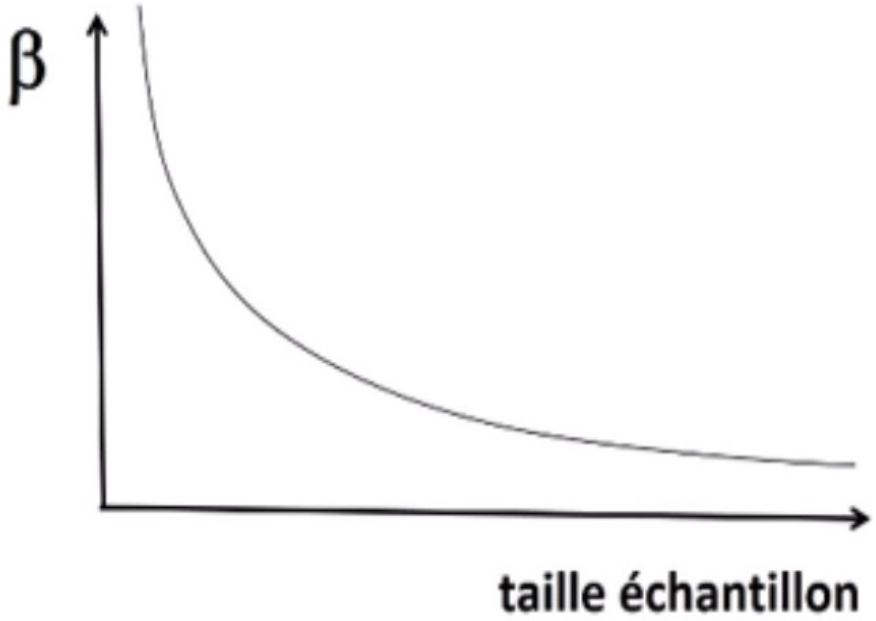
→ H0 « correctement » rejetée

→ = les données n'appartiennent pas à la même distribution

→ DEUX populations différentes

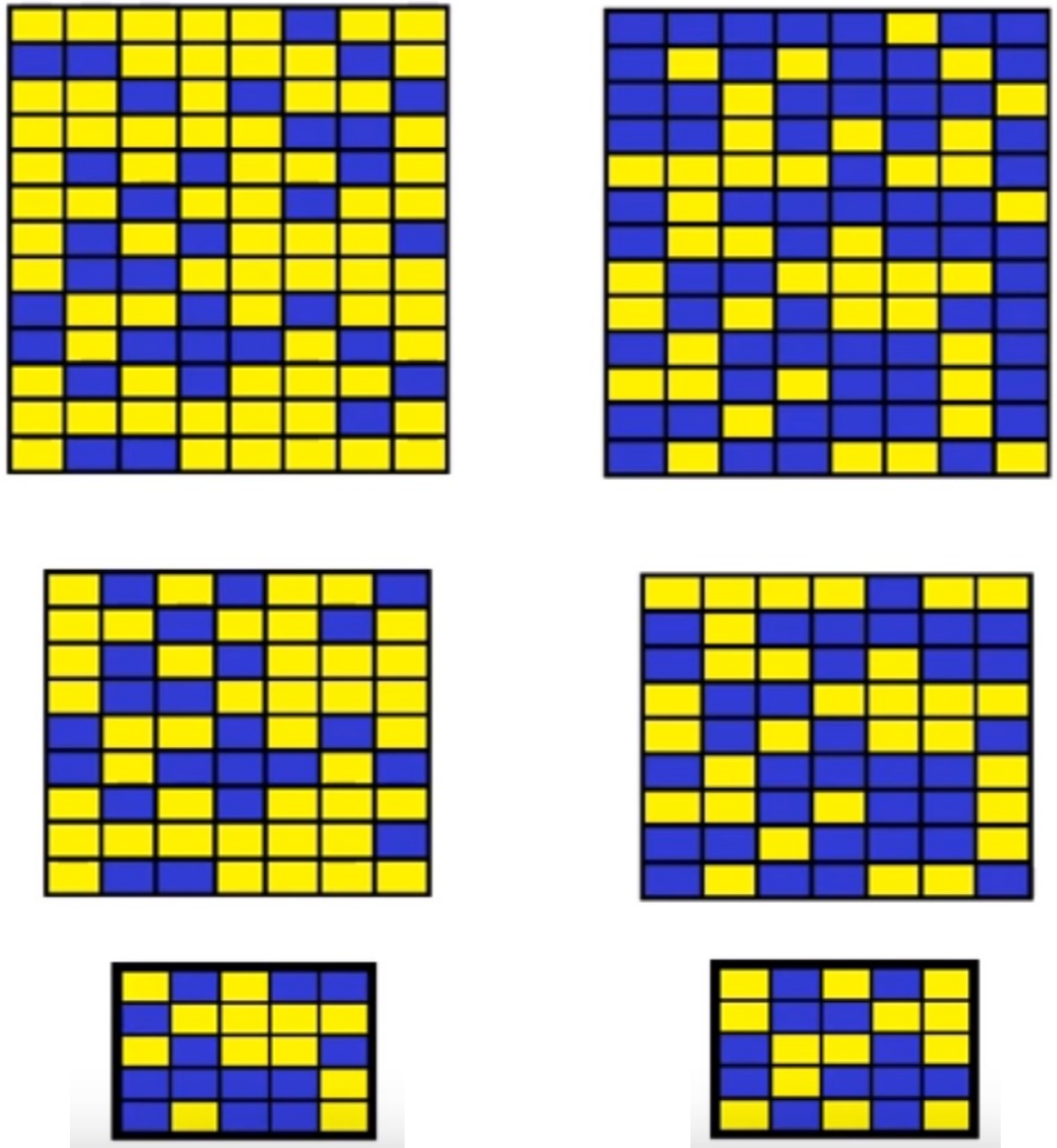
Relation fondamentale

Power =$1 - \beta$



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

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

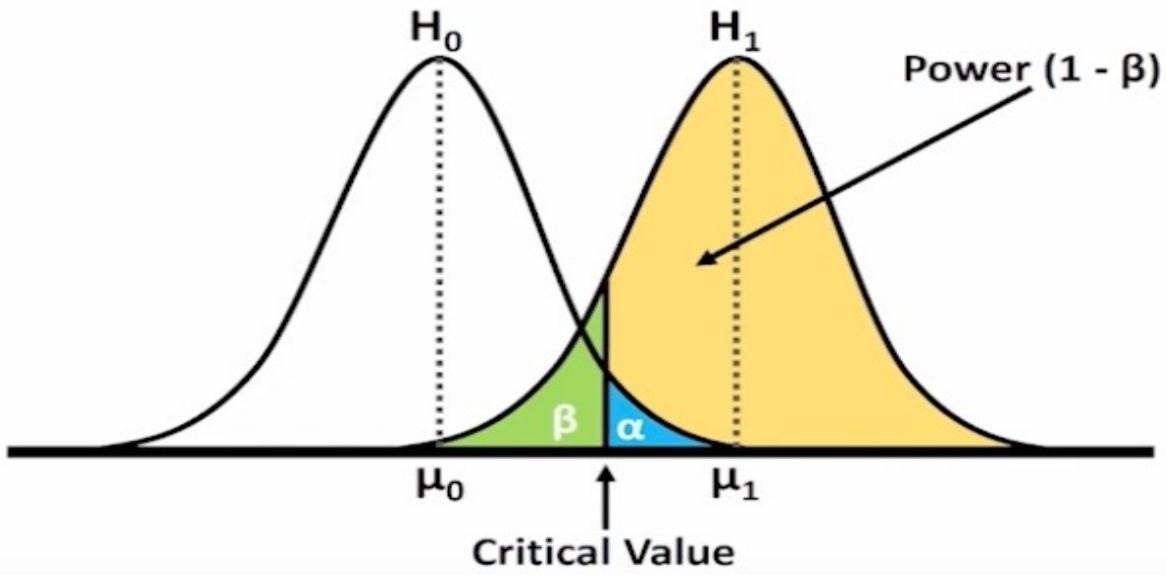


Bilan

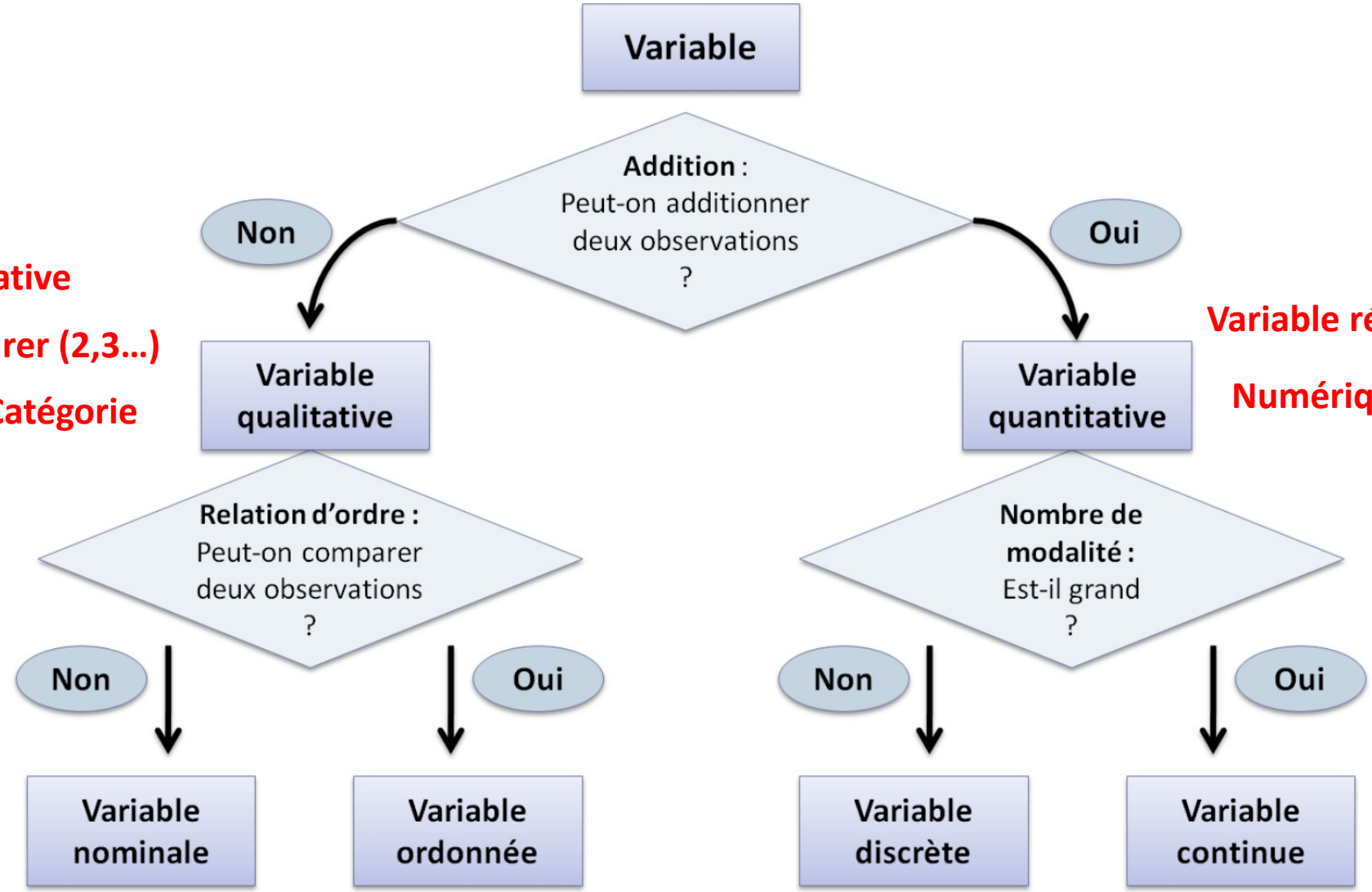
Population

TEST échantillons

	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



Rappel sur les variables… important pour tests statistiques



Variable explicative

Vos groupes à comparer (2,3…)

Catégorie

Variable réponse/à expliquer

Numérique

- Marié, Célibataire, Divorcé…
- Fumeur/non Fumeur
- >10; [10-20]; >20

Notes d'examen entre 0 et 20
Nombre limité

Taille, poids : nombre infini

Tests d'hypothèses bivariés

- 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 (réponse) et 1 qualitative (explicatif)!

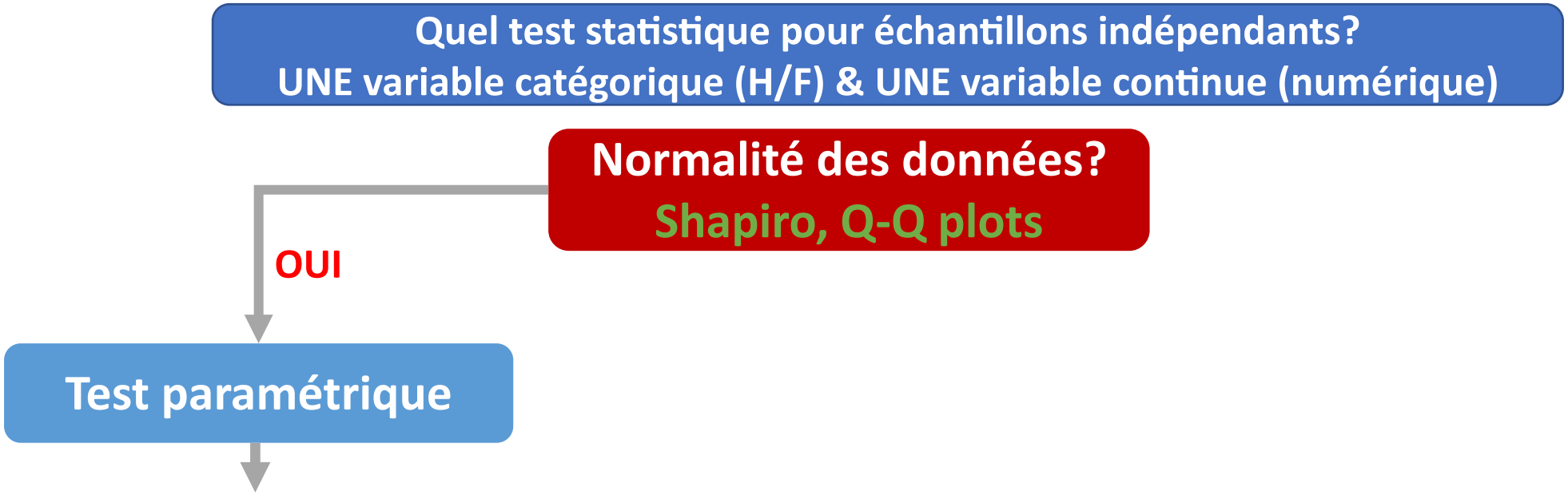
→ **Est-ce que les variations de la richesse en espèce (variable réponse) peuvent être expliquées par la variable explicative (facteur) Traitement?**

→ **Comparaison de moyennes entre groupes**

- **Utilisation de test paramétrique, non paramétrique??**
- **Choix du test?? Significativité ? (pvalue)**
- **Combien de groupes??**
- **Faut-il faire un test Post hoc ??**

Quel test statistique pour échantillons indépendants?
UNE variable catégorique (H/F) & UNE variable continue (numérique)

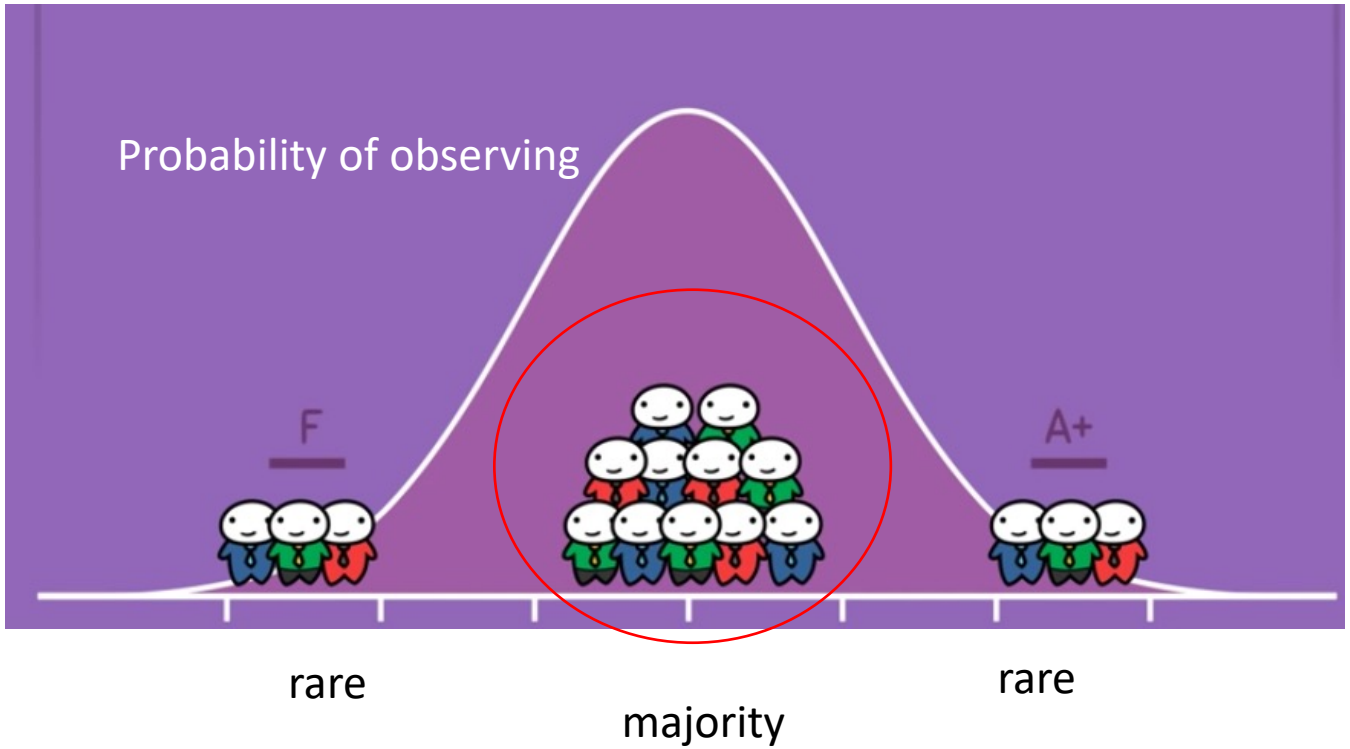
Normalité des données?
Shapiro, Q-Q plots


Quel test statistique pour échantillons indépendants?
UNE variable catégorique (H/F) & UNE variable continue (numérique)
Normalité des données?
Shapiro, Q-Q plots
OUI
Test paramétrique

Caractéristiques d'une distribution normale

Symétrique

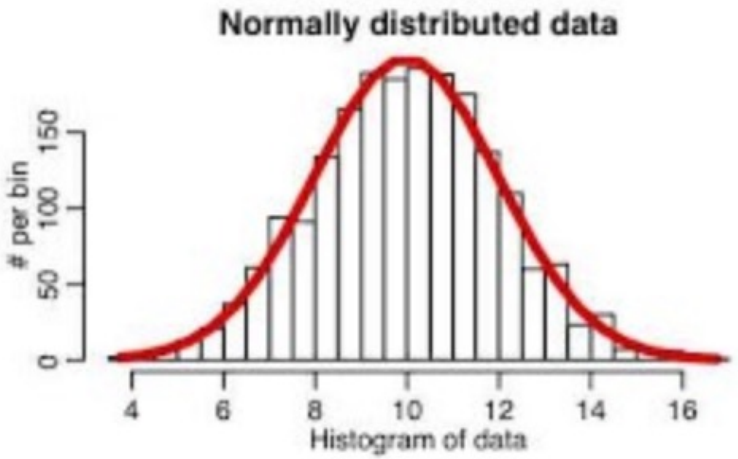
- **Centré sur la moyenne/mean**
- **Dispersion autour de la moyenne: Standard deviation (SD)**
- **95% data sont dans -/+ 2 SD**

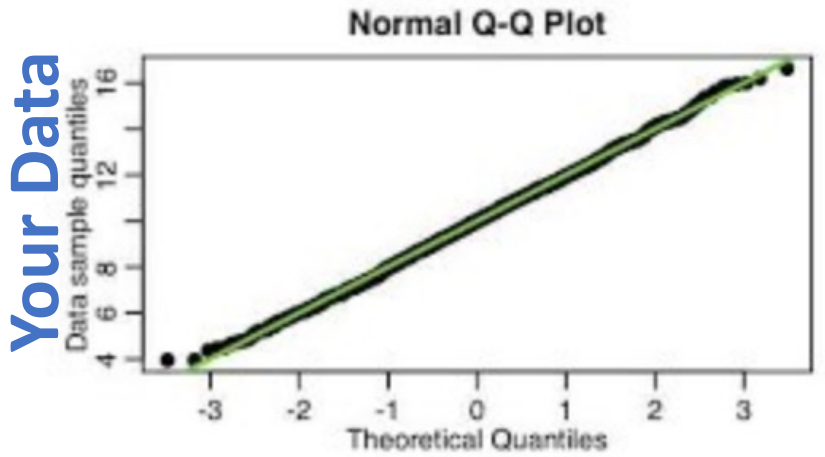


Vérifier la normalité des données: Shapiro Test & QQ-plots!!

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

Est-ce que mes données suivent une loi normale?



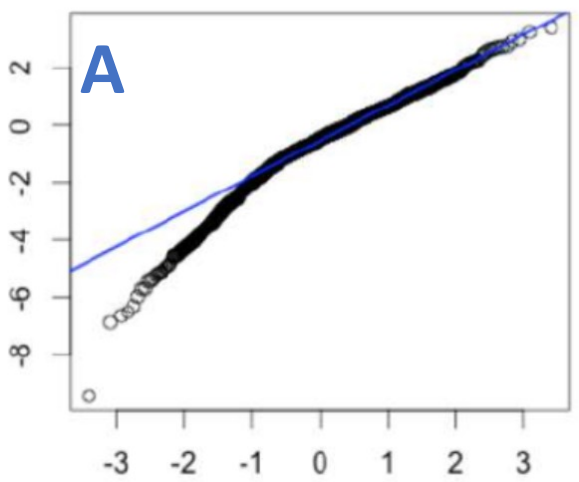


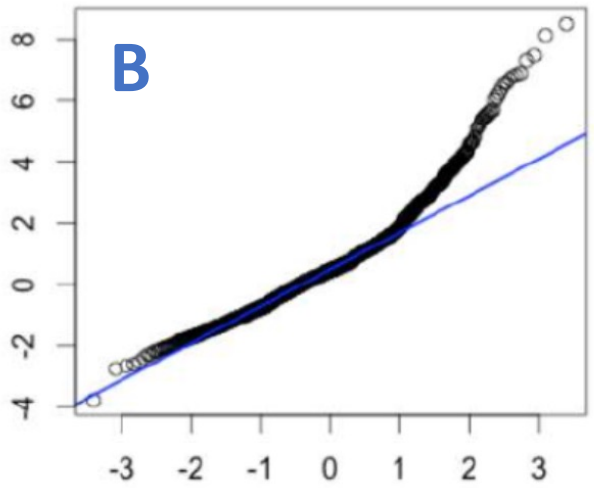
Normal Data (u= 0, SD=1)

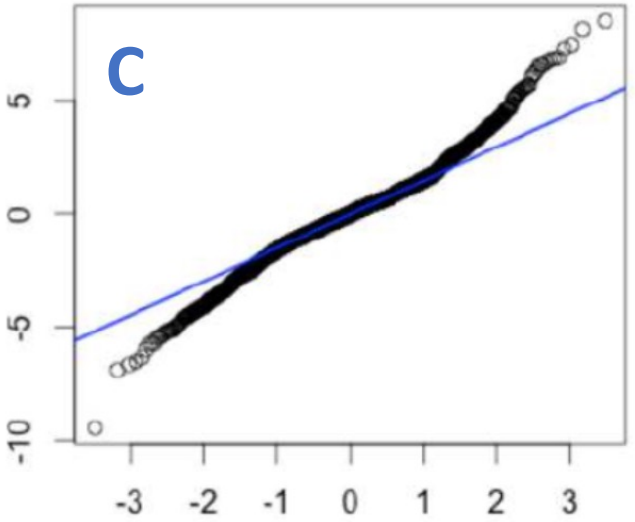
Conclusion?

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

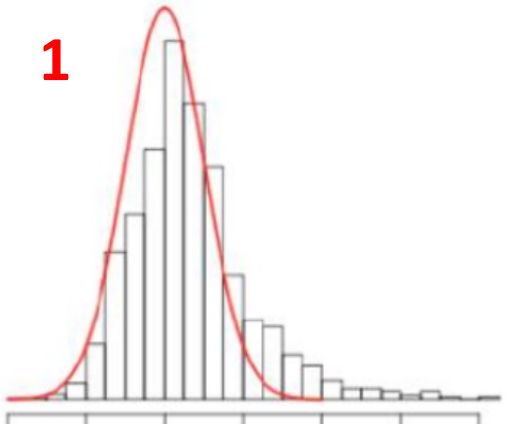
Quelles sont les distributions correspondantes à ces QQ-plots?

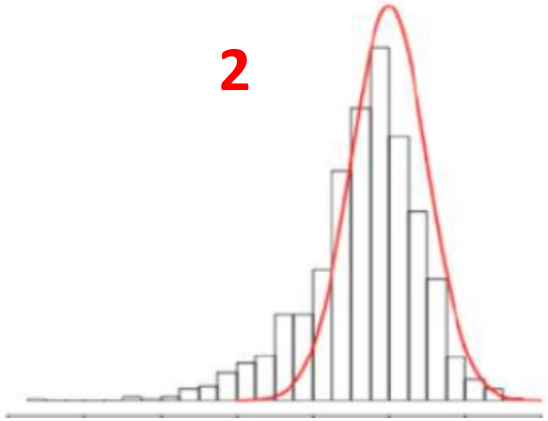


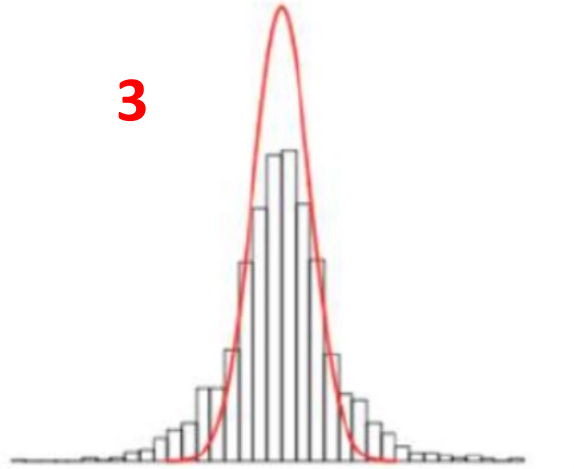


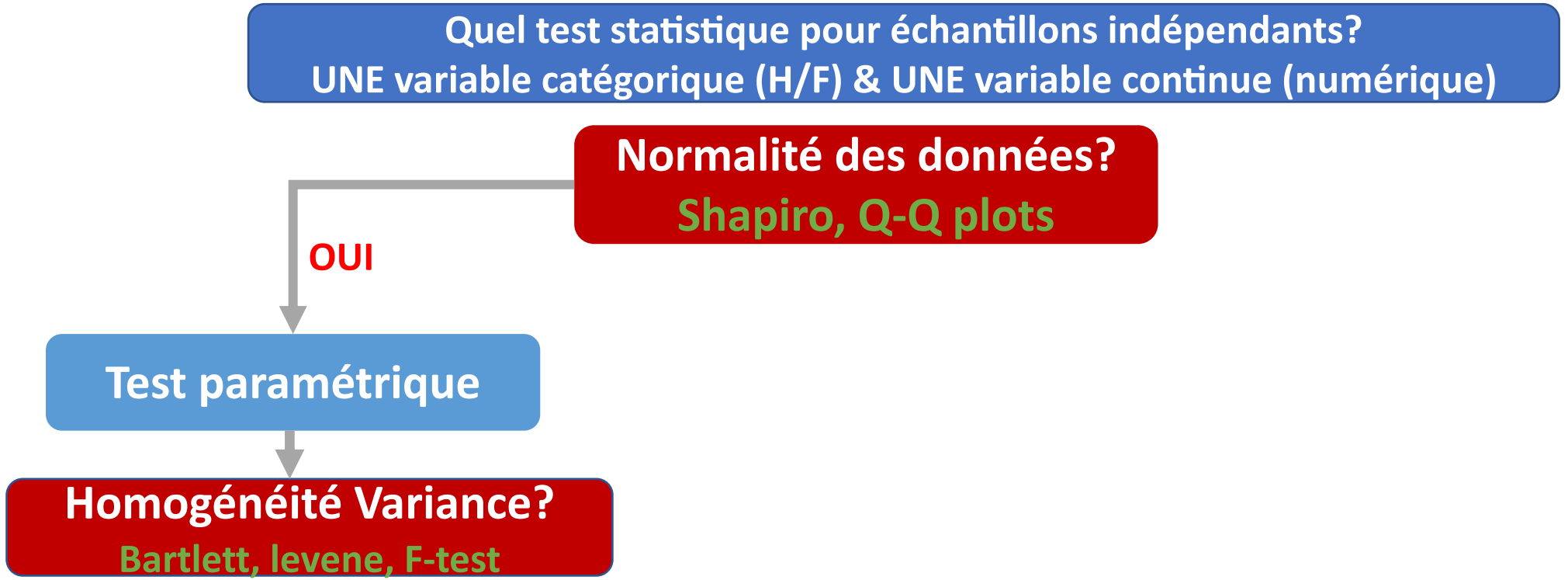


??????????








Quel test statistique pour échantillons indépendants?
UNE variable catégorique (H/F) & UNE variable continue (numérique)
Normalité des données?
Shapiro, Q-Q plots
OUI
Test paramétrique
Homogénéité Variance?
Bartlett, levene, F-test

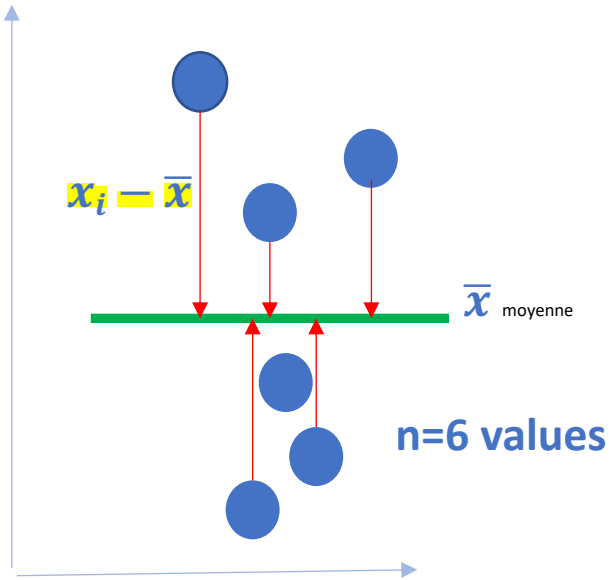
Variance= S^2/σ^2

- **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**

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



Ecart-type (Standard Deviation)=S/ σ

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

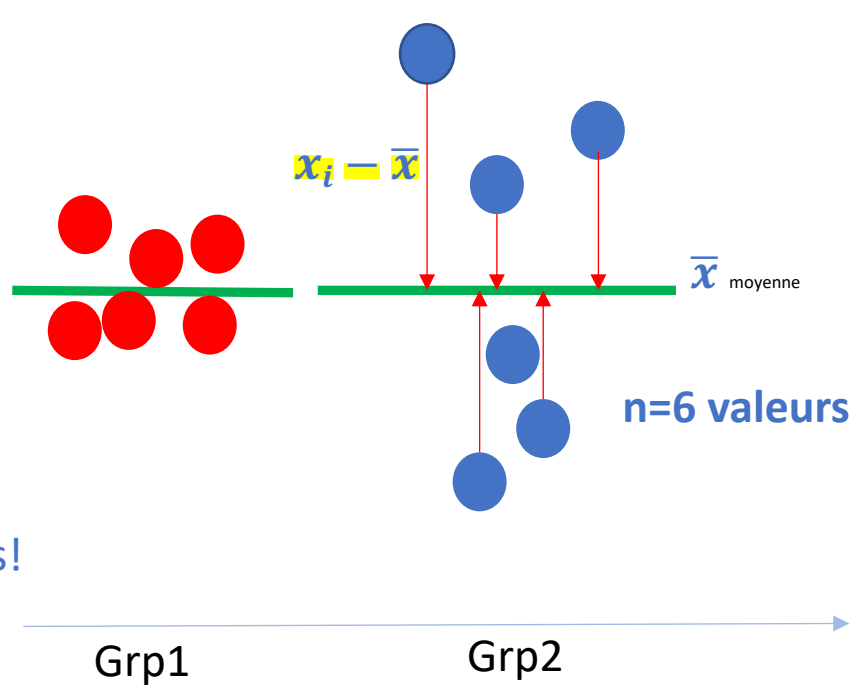
L'avantage de l'écart-type est de s'exprimer dans la même unité que la série de données

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

SS sera bien plus grande dans l'échantillon ….

Exple 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!

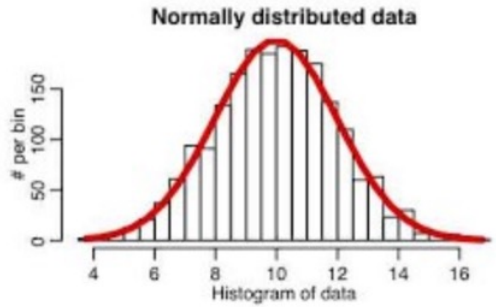
→ Numérateur de la Variance!!

- ***Mean Square (= Mean Sq= Toute la formule = VARIANCE!!!)***

Appliquer un test paramétrique… check-list!

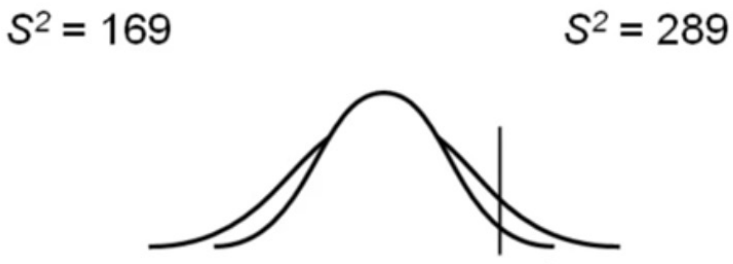
- Vérifier la normalité des données : Shapiro Test & QQ-plots

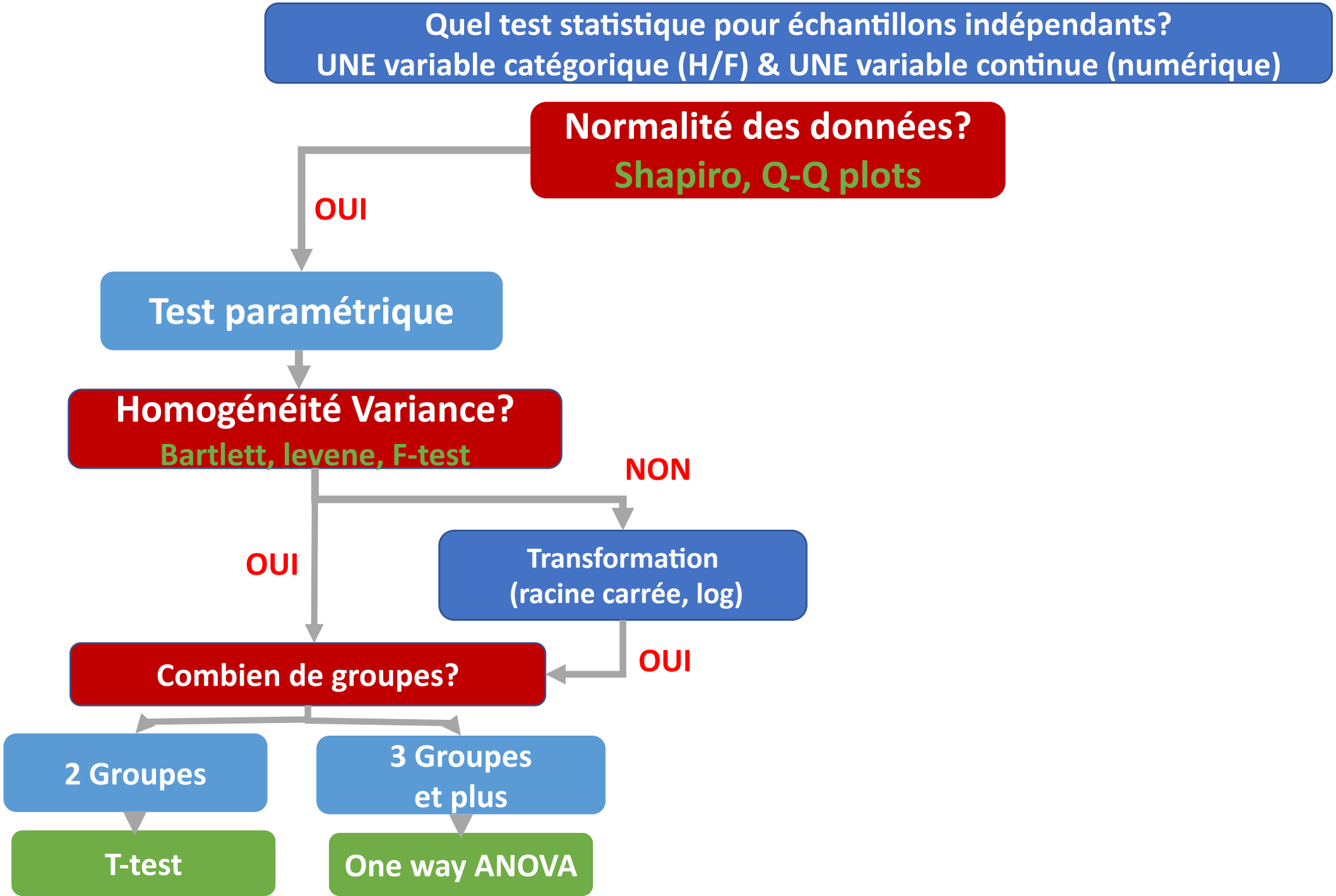
H0: « données suivent une distribution normale»



- Vérifier l'homogénéité de la variance : F-test (2 groups), Bartlett's & Levene's tests

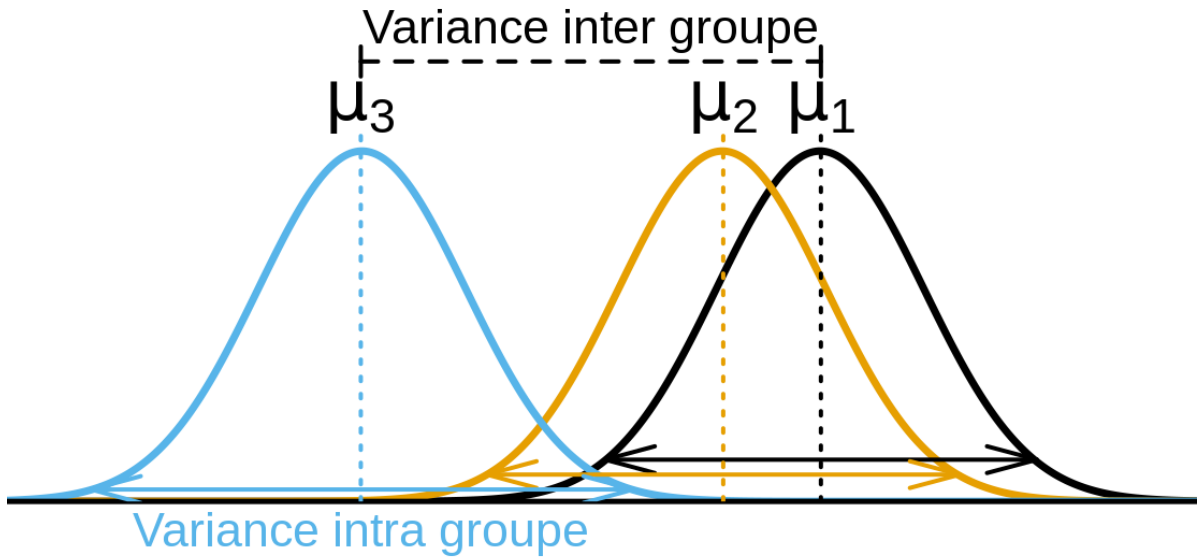
H0: « pas de différence »




Quel test statistique pour échantillons indépendants?
UNE variable catégorique (H/F) & UNE variable continue (numérique)
Normalité des données?
Shapiro, Q-Q plots
OUI
Test paramétrique
Homogénéité Variance?
Bartlett, levene, F-test
NON
OUI
Transformation
(racine carrée, log)
OUI
Combien de groupes?
2 Groupes
3 Groupes
et plus
T-test
One way ANOVA

ANOVA: ANalysis Of VAriance (One way Anova= Univariée) (3 groups minimum)

- Compare la variance des moyennes 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 (AU MOINS 3) sont si **faibles** qu'elles **s'expliquent** facilement **par le hasard!!!***

- **Evaluation : Compare la variance inter-groupes (DES MOYENNES) 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 :

Variance inter-groupes + **Variance intra-groupes**

attribuable au facteur

attribuable à l'expérimentale

(fluctuation de l'échantillonnage, hasard)

Effet du facteur!

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

Effet du hasard!

	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**!

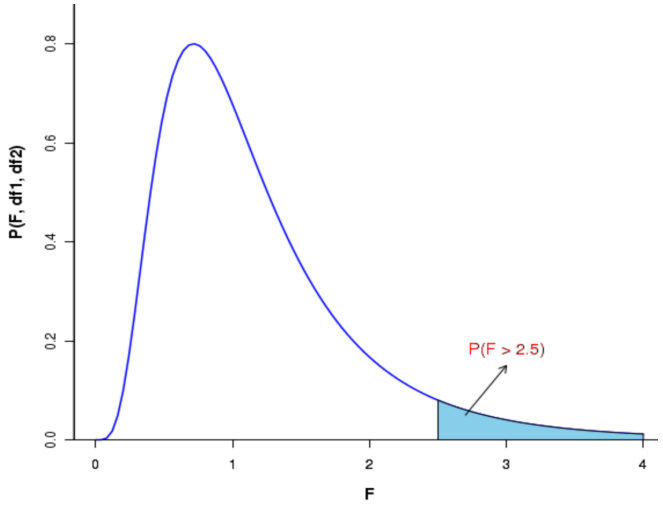
Statistique F suit une loi dites de Fisher-Snedecor : = Distribution F utilisée pour test des variances, distribution des variances n'étant pas normale.

- **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!**

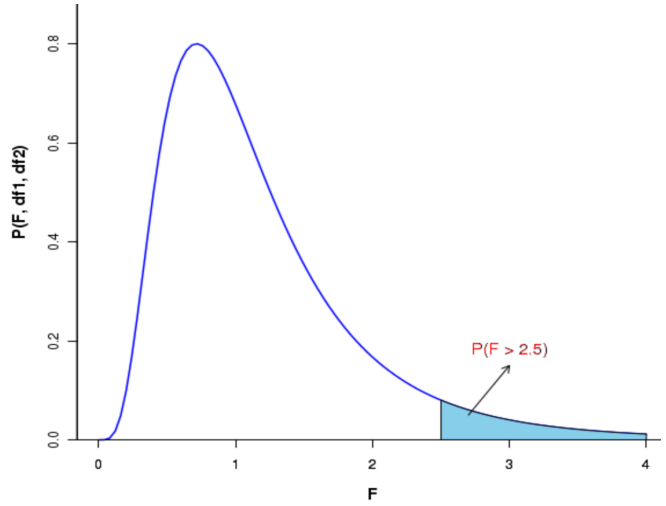
Dénominateur S^2 Numérateur S^2 S^2

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



- **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

	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		

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

Degré de liberté

- **Two-ways ANOVA : Influences de DEUX variables qualitatives sur UNE variable quantitative.**

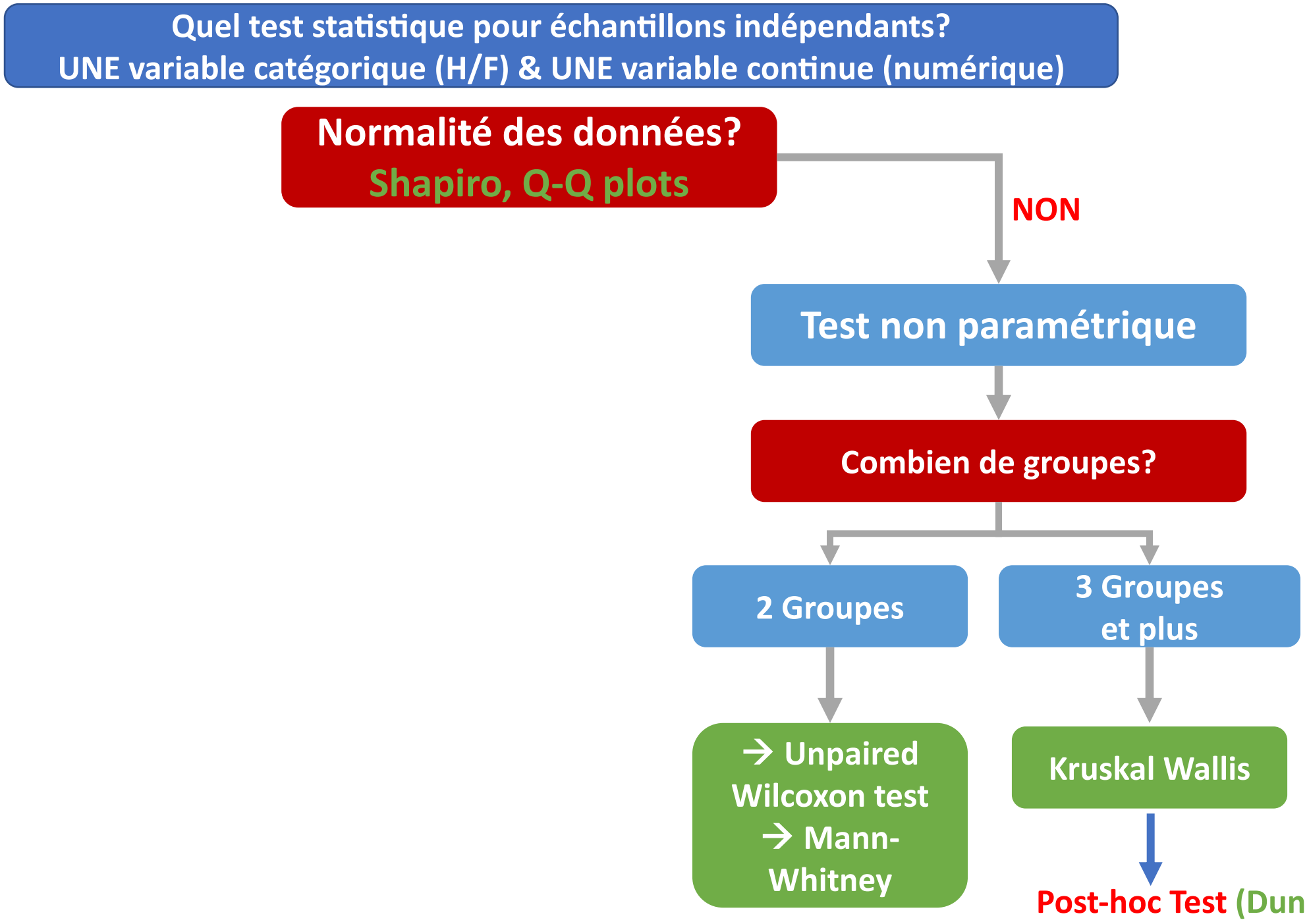
Exple: Influence du type de sol et de la température (variable ordinale) sur rendement de plantes

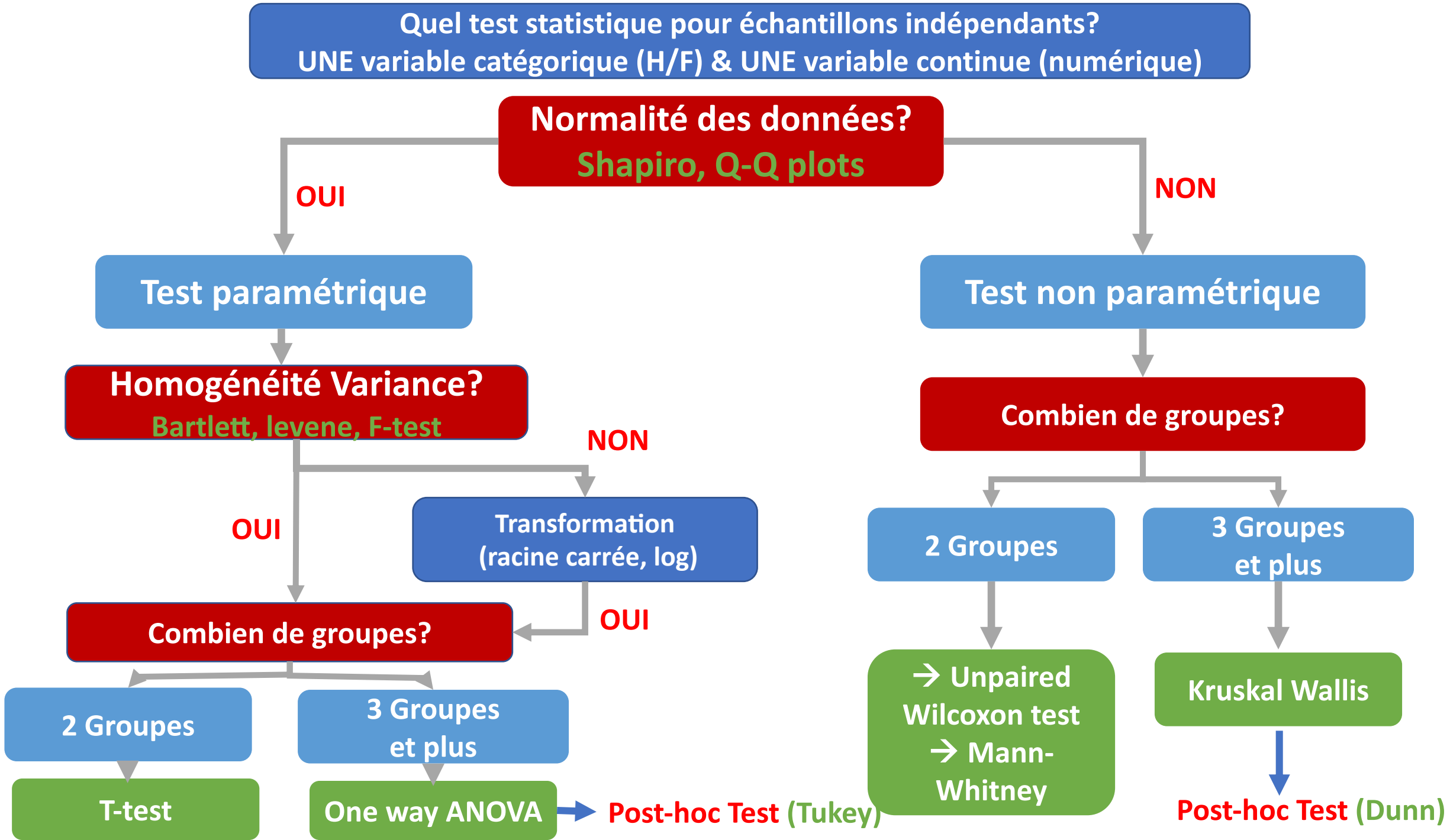
Tests non-paramétrique

Aucune hypothèse n'est faite sur la distribution des données :
Tests sans distribution, ils sont alternatifs aux tests paramétriques

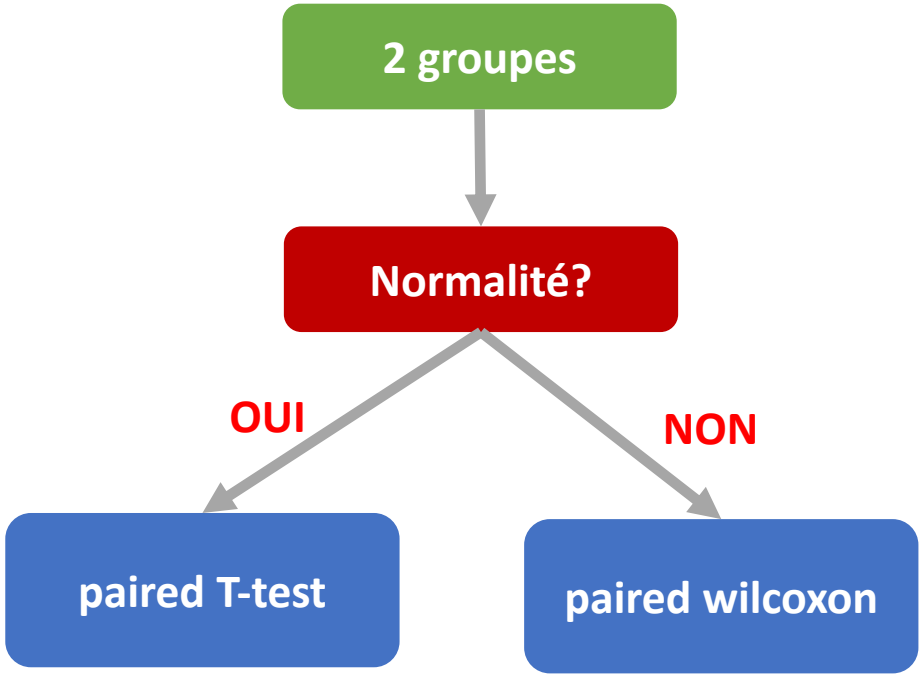
- **Wilcoxon Rank test : samples are paired/unpaired, 2 sample groups**
- **Mann-Withney test: Independent samples, 2 sample groups**
- **Kruskal wallis test : Independant samples, Three or more groups**

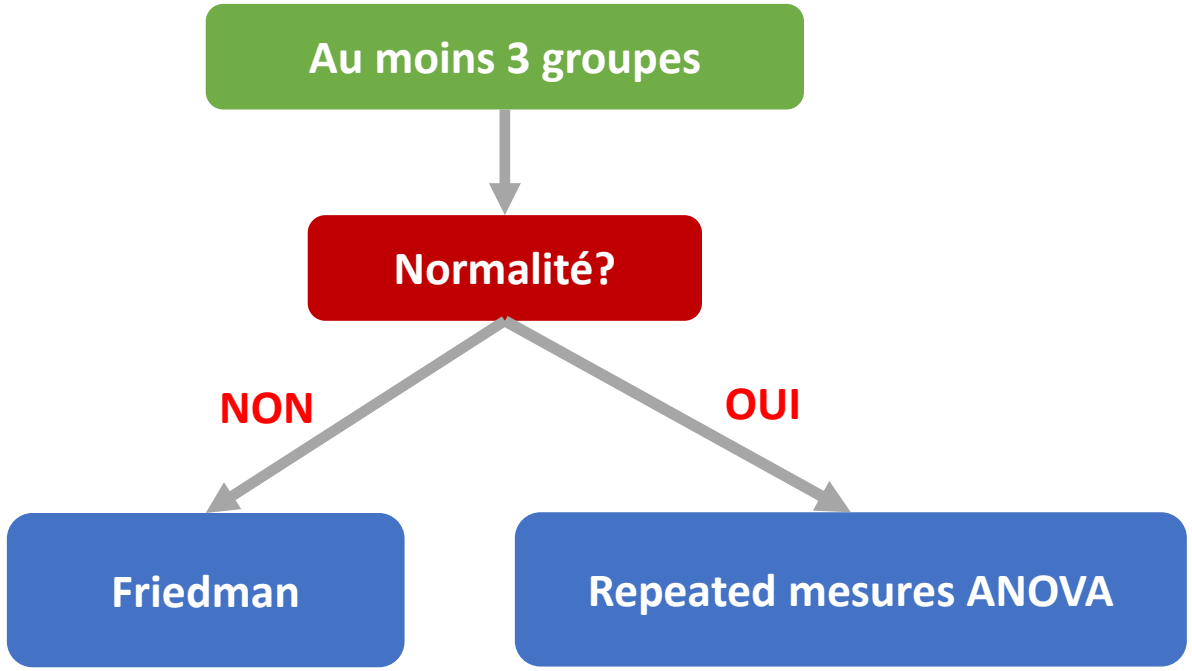
→ Basé sur la moyenne des rangs : on classe les valeurs, on remplace par une position (1,2 etc), Compare la moyenne des rangs entre les groupes


Quel test statistique pour échantillons indépendants?
UNE variable catégorique (H/F) & UNE variable continue (numérique)
Normalité des données?
Shapiro, Q-Q plots
NON
Test non paramétrique
Combien de groupes?
2 Groupes
3 Groupes et plus
→ Unpaired Wilcoxon test → Mann-Whitney
Kruskal Wallis
Post-hoc Test (Dunn)


Quel test statistique pour échantillons indépendants?
UNE variable catégorique (H/F) & UNE variable continue (numérique)
Normalité des données?
Shapiro, Q-Q plots
OUI
NON
Test paramétrique
Test non paramétrique
Homogénéité Variance?
Bartlett, levene, F-test
NON
OUI
Transformation
(racine carrée, log)
OUI
Combien de groupes?
Combien de groupes?
2 Groupes
3 Groupes
et plus
2 Groupes
3 Groupes
et plus
T-test
One way ANOVA
Post-hoc Test (Tukey)
→ Unpaired
Wilcoxon test
→ Mann-
Whitney
Kruskal Wallis
Post-hoc Test (Dunn)

CAS Mesures répétées – échantillons appariés
Exple= série temporelle, Avant-Après Traitement…





Post-hoc Test

Cas des tests statistiques avec **au moins 3 groupes!**

- ANOVA, Kruskal-wallis etc

→ Le résultat d'un test ANOVA est une **p-value Globale,**

Exple: Vous comparez l'effet de 3 types (de sol A,B,C) sur la croissance plante.
ANOVA vous renvoie une p-value de 0.03 (< au seuil de 0.05), donc significative (H0 est rejetée).
→ Cela ne vous indique pas quelles paires de groupes sont significativement différents!

Vous devez faire **un Test Post-hoc** pour le savoir :
→ comparaisons multiples (exple: Gp A vs. Grp. B; GrpB vs. Grp C; Grp C vs. Grp A!)

- **Paramétrique Post-hoc test (ANOVA)** → **Tukey Test**
- **Non-paramétrique Post-hoc test (Krukal wallis)** → **Dunn Test**

Connexion à l’évènement wooclap : YTEIXI

1 Allez sur wooclap.com

2 Entrez le code d'événement dans le bandeau supérieur

Code d'événement
YTEIXI

1 Envoyez @YTEIXI au 06 44 60 96 62

2 Vous pouvez participer

Le problème des tests multiples : augmentation du risque...

Les tests sont basés sur des probabilités, il y a donc toujours un risque de tirer des conclusions erronées !

→ **Aucun test d'hypothèse n'est fiable à 100 %**

Effectuer des tests d'hypothèse :

Vous avez deux hypothèses :

- H0 : Hypothèse nulle = l'hypothèse de référence : pas de différence
- H1 : Hypothèse alternative : Il y a une différence

Vous rencontrez : Erreur de type I : α = Risque alpha

- $\alpha = 0,05$ C'est la probabilité de rejeter H0 à tort ! (seuil de significativité)
- En d'autres termes, une chance acceptable de faux positif !!!

Abondance différentielle : Tests multiples !!

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

Prob complémentaire $P_{\text{no\_error}} = 1 - \underline{\alpha} = 0.95$

2 TESTS sans faire d'erreur : $P_{\text{no\_error in two tests}} = (1 - \underline{\alpha}) * (1 - \underline{\alpha}) = (1 - \underline{\alpha})^2$

Prob complémentaire $P_{\text{at\_least\_ONE\_error in two tests}} = 1 - (1 - \underline{\alpha})^2$

Generalisation à n TESTS $P_{\text{at\_least\_ONE\_error in\_n\_tests}} = 1 - (1 - \underline{\alpha})^n$

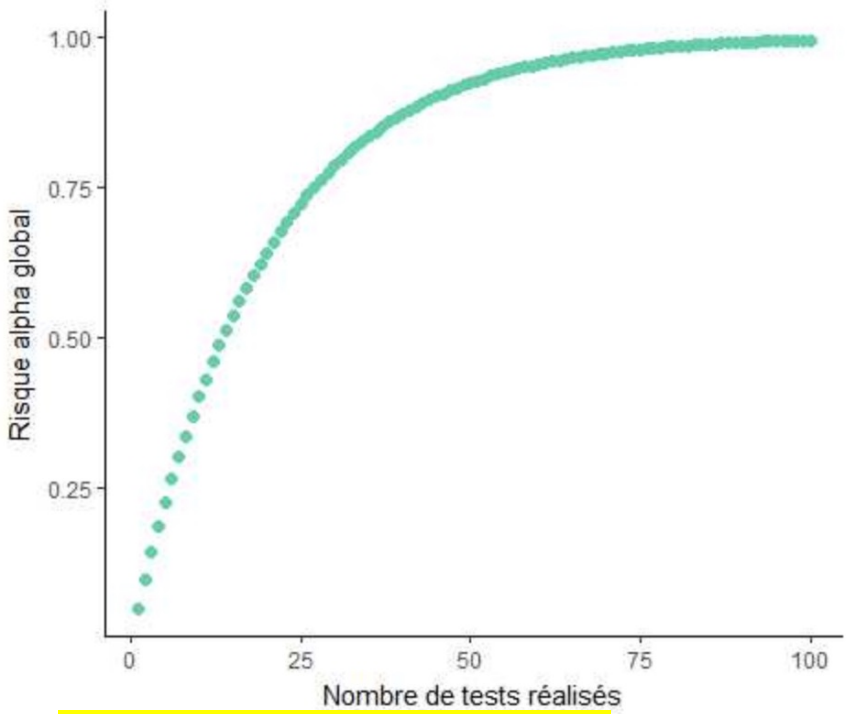
It's called the global $\underline{\alpha}$ risk

What does it means…

- You test ONE ASVs (n=1) for differential abundance: $1-(1-\alpha)^n = 1-(1-0.5)^1 = 0.05$
- You test 3 ASVs (n=3): $1-(1-0.05)^3 = 0.14$
- You test 100 ASVs (n=100): $1-(1-0.05)^{100} = 0.9941$

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

→ 99% de rejet à tort H0 au moins une fois



Il faut ajuster ce phénomène en utilisant la p.value ajustée !

FDR : False Discovery Rate : Benjamini-Hochker

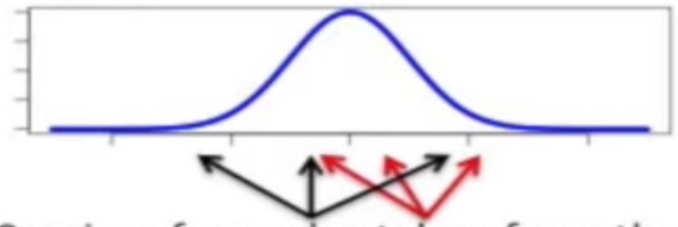
L'idée : Rejeter les mauvaises données qui ont l'air bonnes!!!

Benjamini-hocherk **ajuste la p-values**
Pour limiter le nombre de **faux positifs**
qui sont signalés comme **significatifs** (pvalue < 0.05)

p-values ajustées
= augmente la valeur!

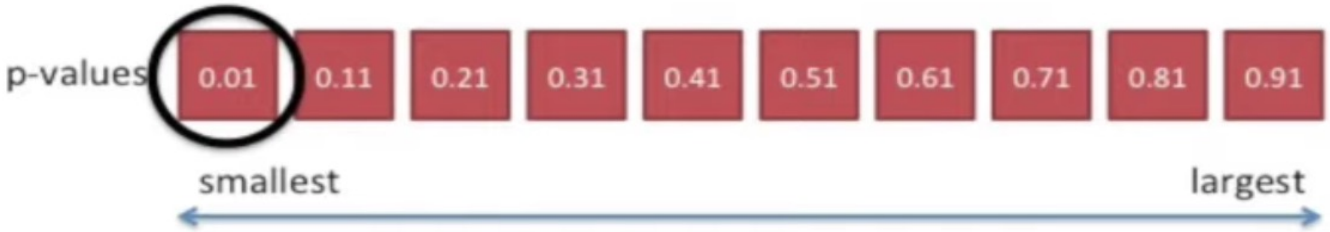
Using FDR cutoff < 0.05
signifie que moins de 5% des résultats significatifs seront des faux positifs

Mathematical approach FDR-Benjamini-Hochker



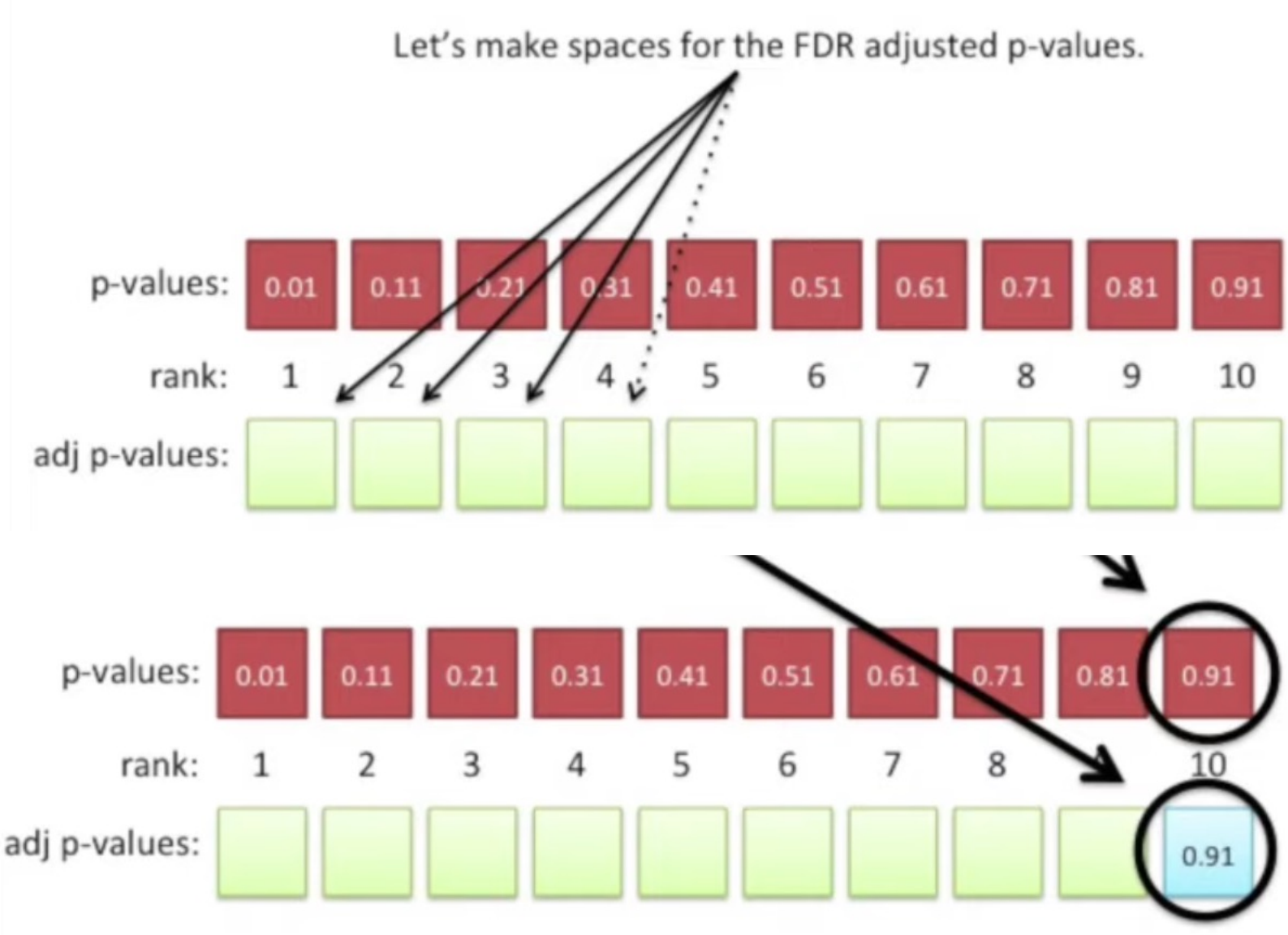
10 pairs of samples taken from the same distribution. (i.e. 10 genes that were not effected by the drug).

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



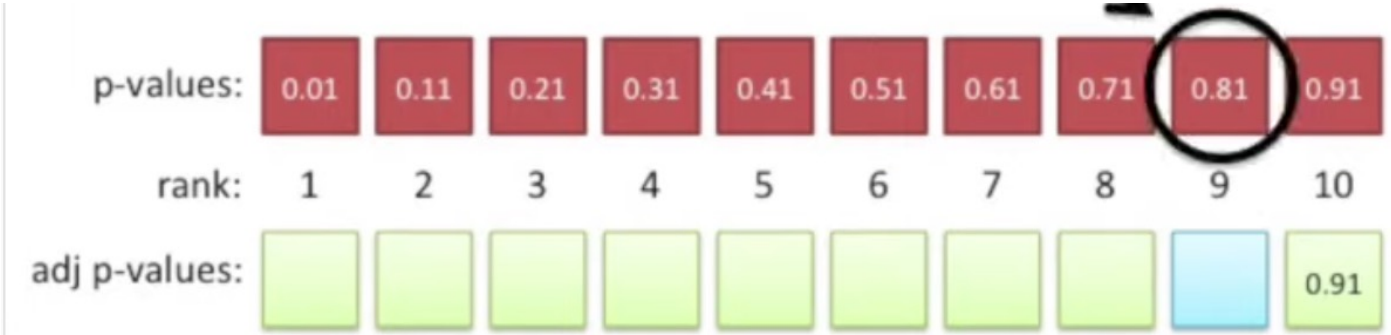
1- Ranking pvalue

Prepare space for adjusted p-value



2- Largest adjusted pvalue and larger pvalue are same

Next adjusted pvalue ….

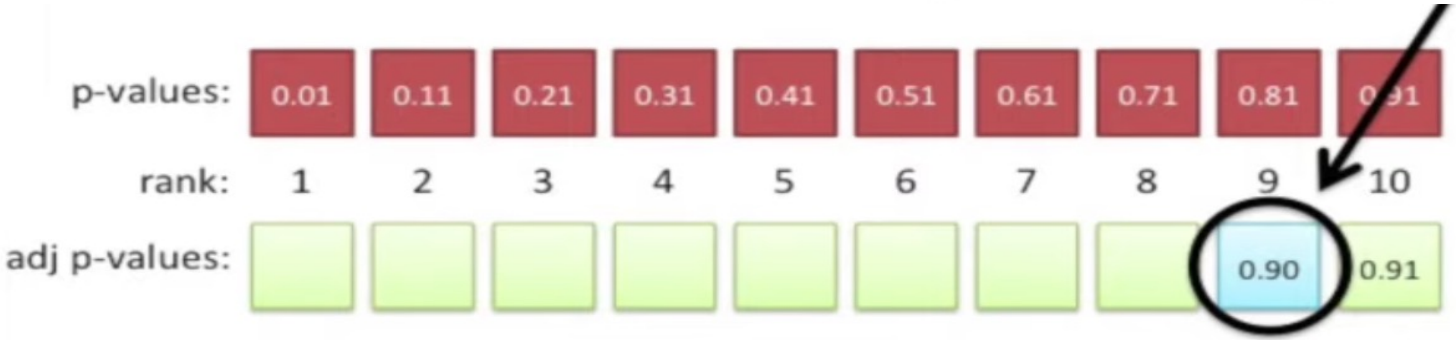


The smallest of the two options

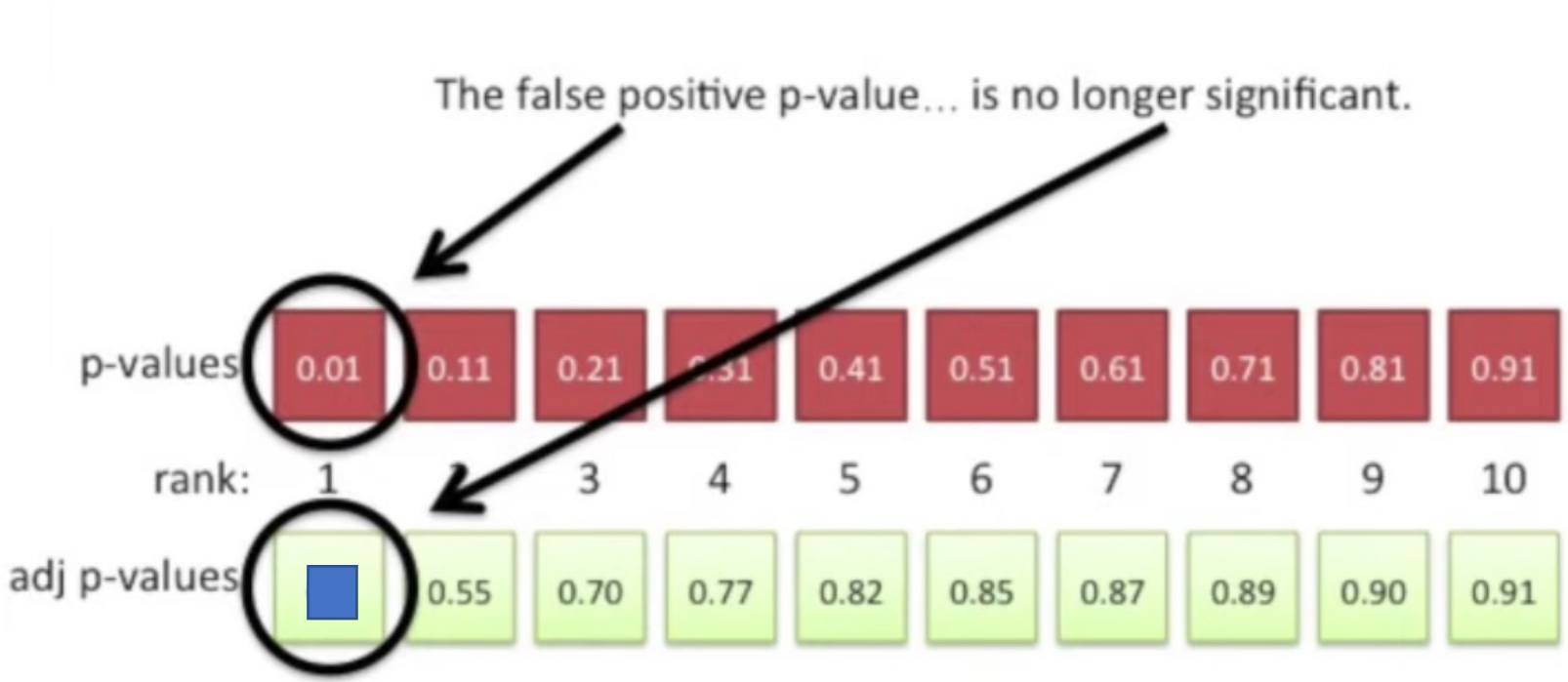
$$\text{b: the current p-value} * \left(\frac{\text{total \# of p-values}}{\text{p-value rank}} \right)$$

a: The previous adjusted p-value = 0.91

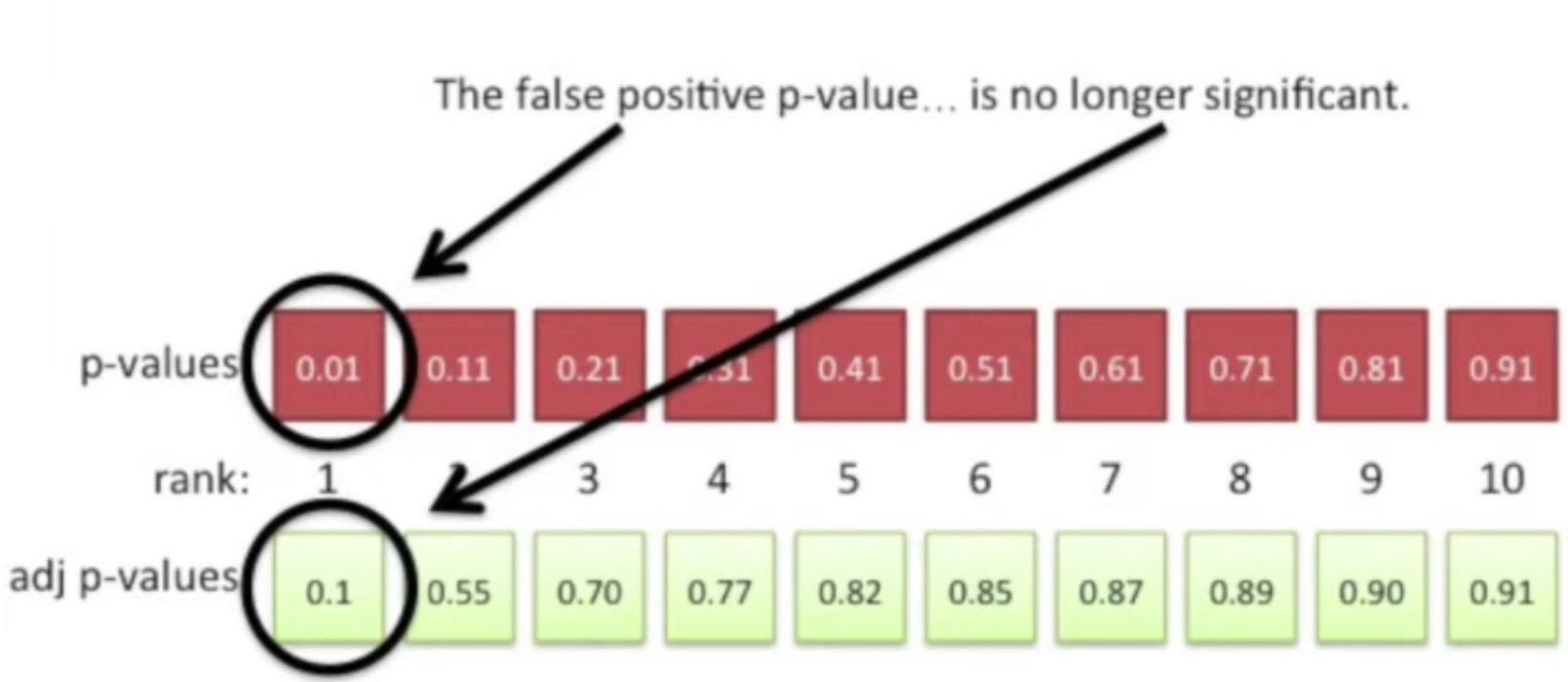
$$\text{b: 0.81} * \left(\frac{10}{9} \right) = 0.90$$



Finalement…


The false positive p-value… is no longer significant.
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
0.55
0.70
0.77
0.82
0.85
0.87
0.89
0.90
0.91

Finalement…



Le faux positif ne l'est plus grâce à cette méthode d'ajustement !!

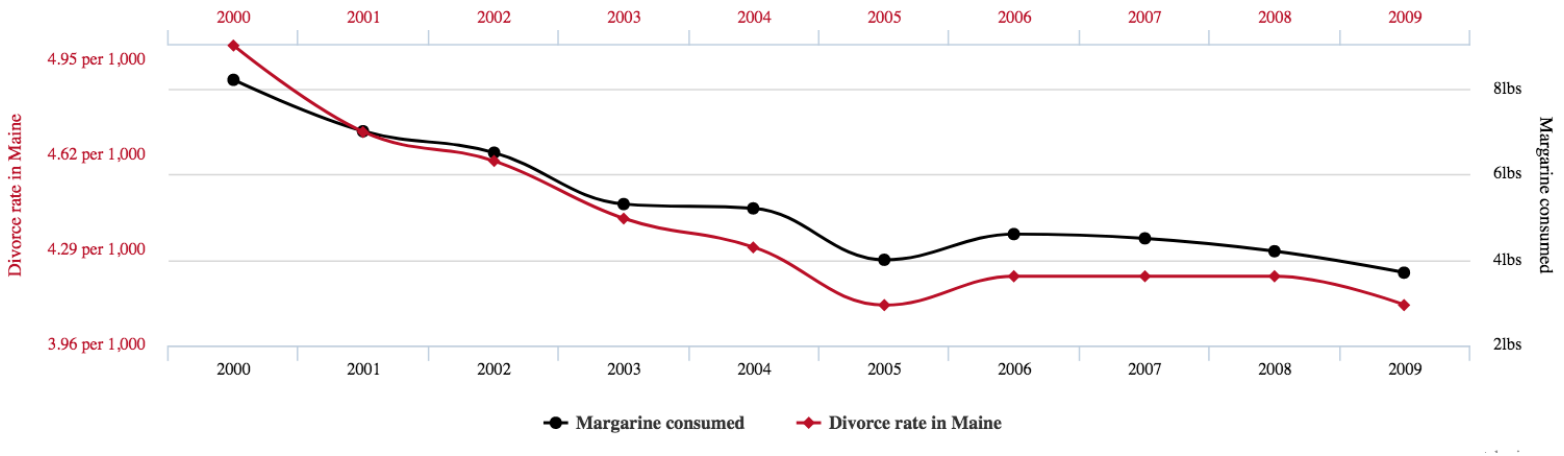
Régression linéaire & Corrélation (analyses bivariées)

Objectif : Analyser le lien qui peut exister entre deux variables (ici : quantitatives)
(Deux variables qualitatives -> test Khi2)

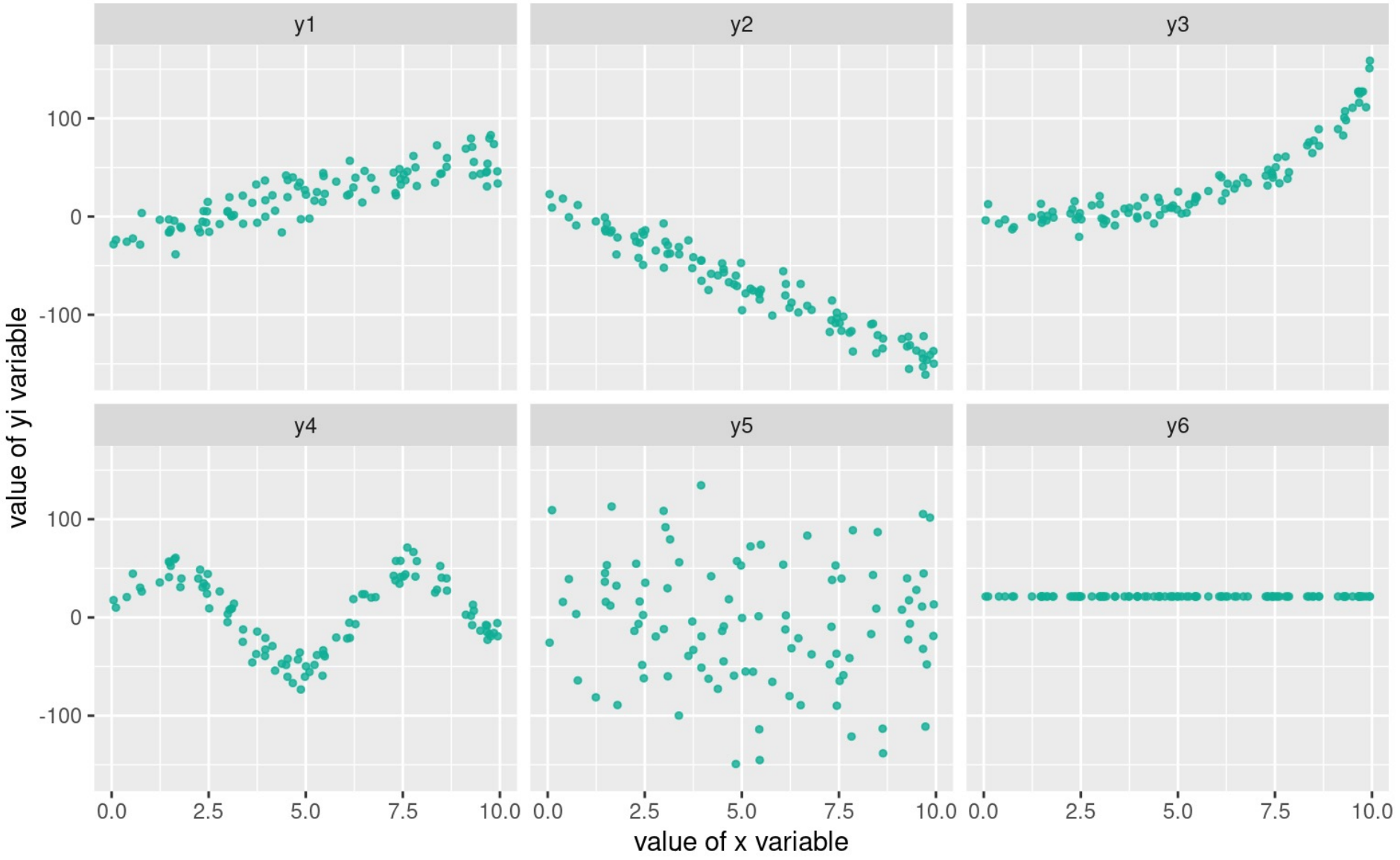
Lien/relation/dépendance entre les variables

→ Les valeurs de deux variables **n'évoluent pas indépend**amment mais présentent au contraire une certaine forme, une certaine régularité.

⚠ L'intensité de l'association n'indique pas un lien de causalité …

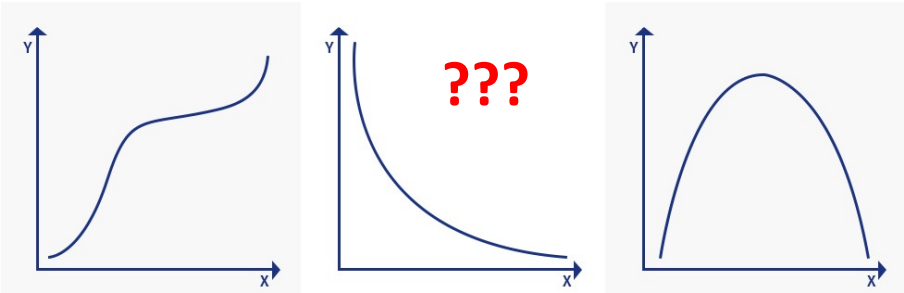


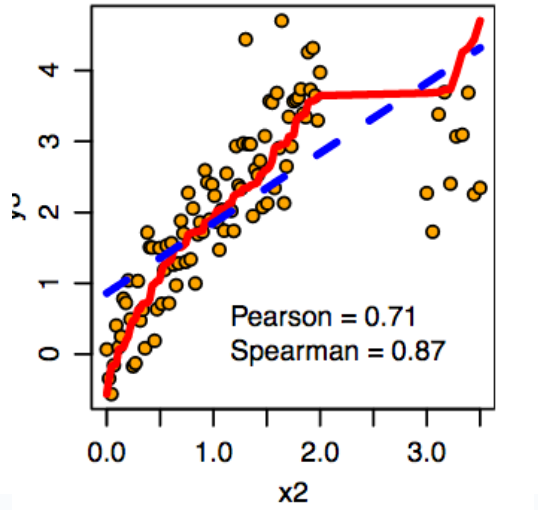
Quelle est la relation entre les variables dans chaque graphique ?



Association: Coefficient de corrélation r
Intensité & Direction de l'association entre deux variables

- **Relation linéaire stricte : Pearson** (**r**, paramétrique)
- **Relation monotone: Spearman** (**Rho**, non-paramétrique)
  Kendall (**Tau**, non-paramétrique), Alternative to Spearman (petit échantillonnage)

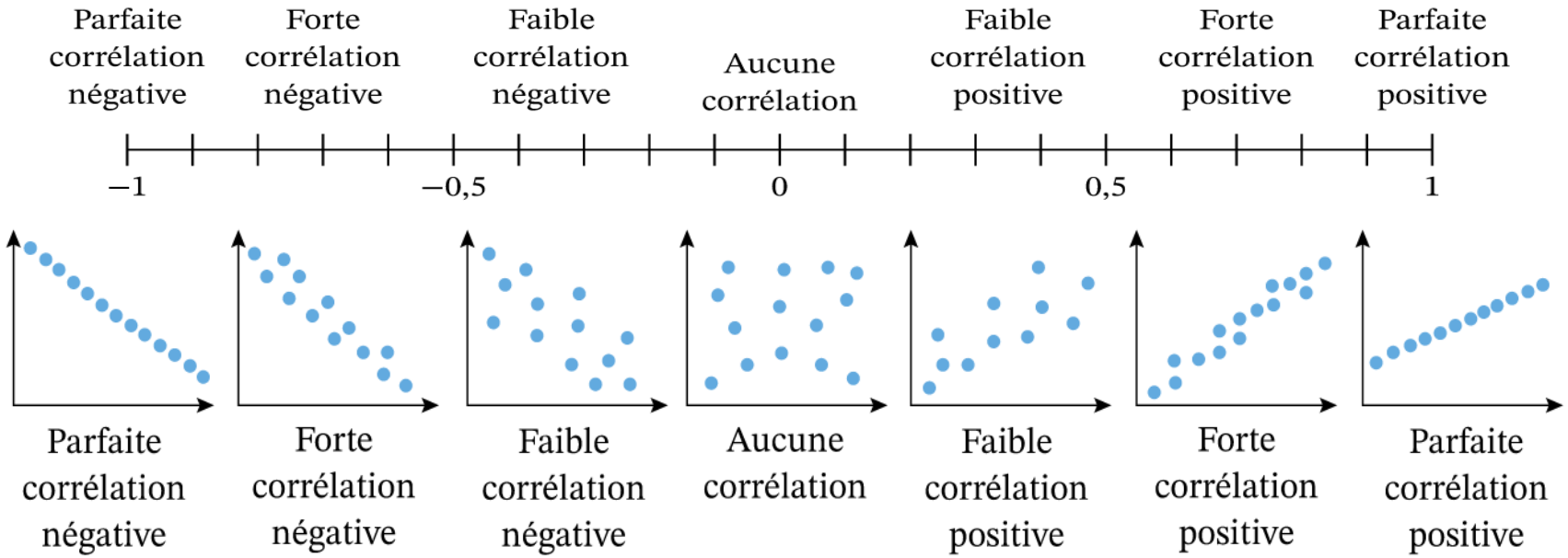




Coefficient r range between -1 et 1

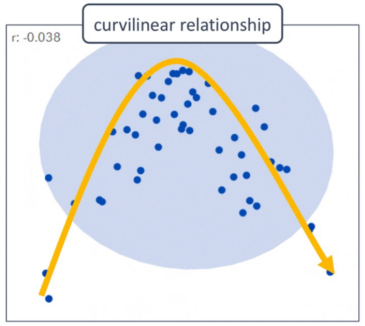
- **Corrélation positive**: Les valeurs des deux variables ont tendance à augmenter ensemble
- **Corrélation négative**: Les valeurs d'une variable ont tendance à augmenter et les valeurs de l'autre variable à diminuer
- **Zero** : association non **LINEAIRE** (**Pearson**)

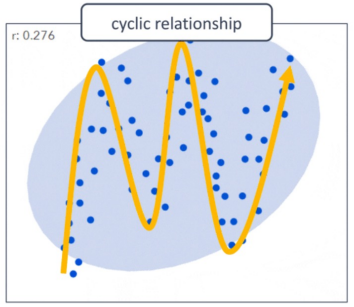
pour information!!!



Parce qu'il n'est jamais inutile de contrôler ses résultats...

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

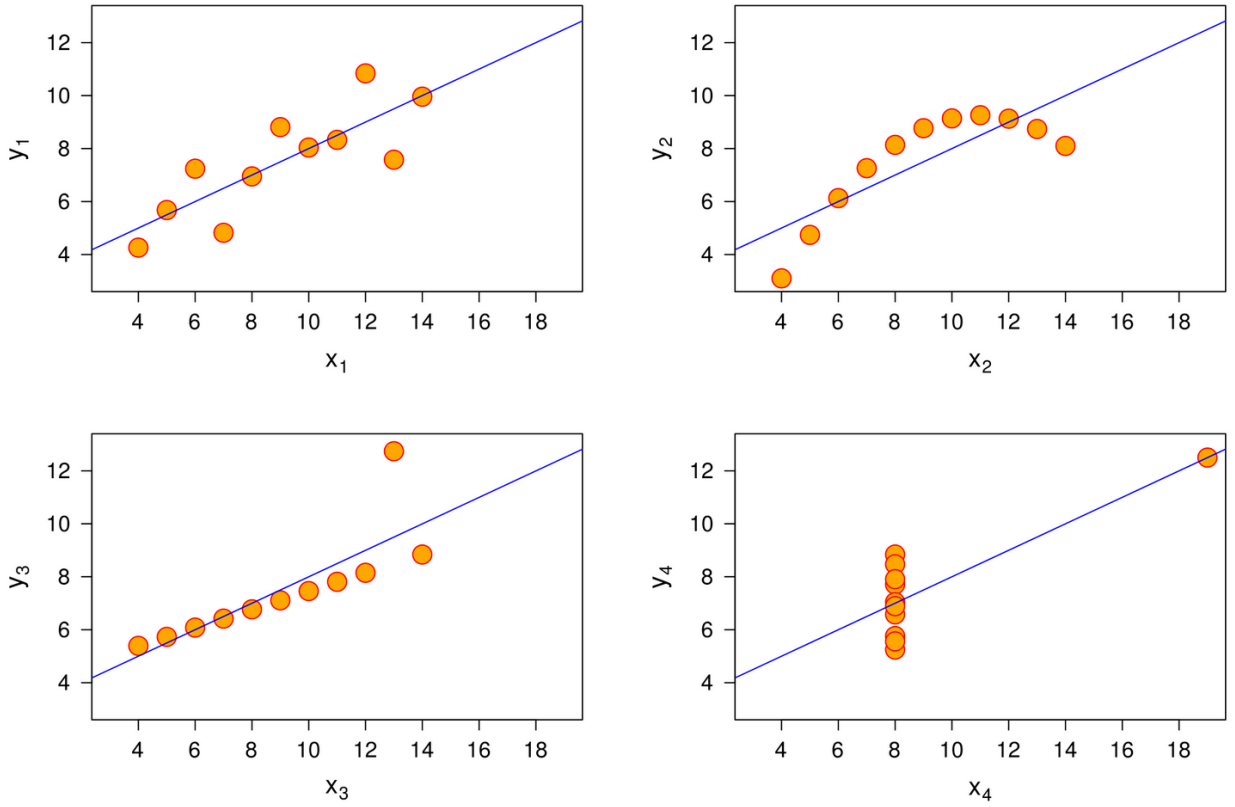




Vraiment pas inutile : Anscombe…

- **4 jeux de données avec les mêmes statistiques desciptives**

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



r =?

☺

0.3?
0.5?
0.8?

- **Loi de distribution de r sous l'hypothèse Ho : Pas de lien statistique entre X et Y**

→ Accès aux p-values

TP !

Régression linéaire simple

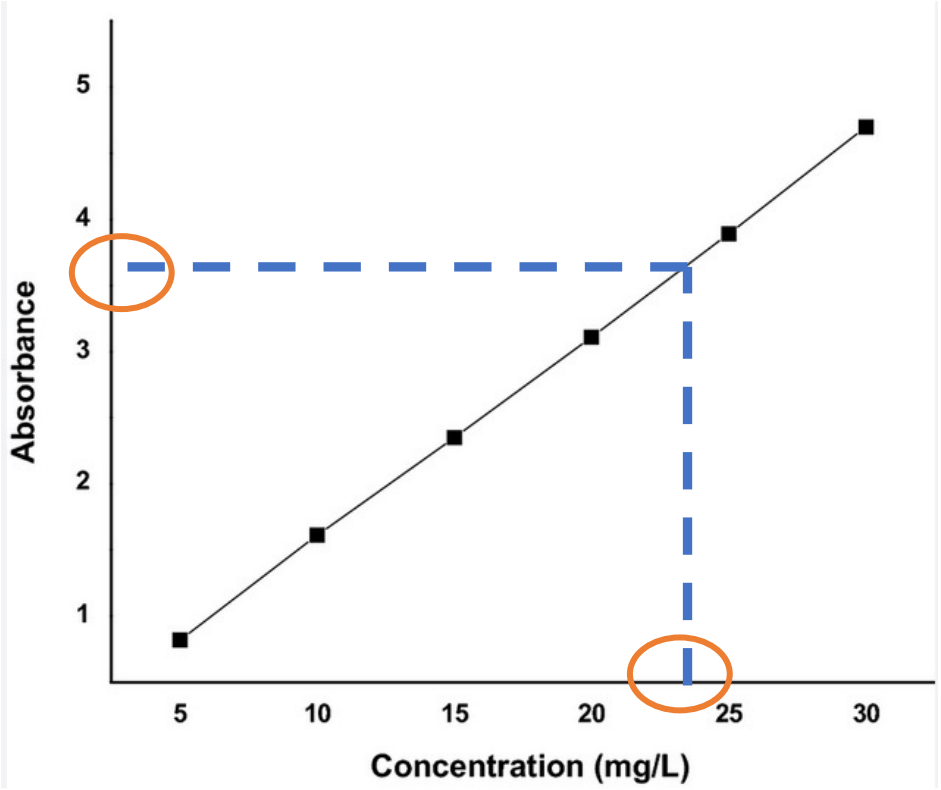
- Uniquement pour les variables quantitatives

- Tracer le diagramme de dispersion : Existe-t-il une **relation**?
- Est-elle **linéaire** ?
- Quelle est son **orientation** (positive, négative)?
- Si l'association est **linéaire** → Faire une **régression**

Nécessite

- Distribution normale
- Homogénéité des variances

Votre régression linéaire favorite… courbe de calibration !!!

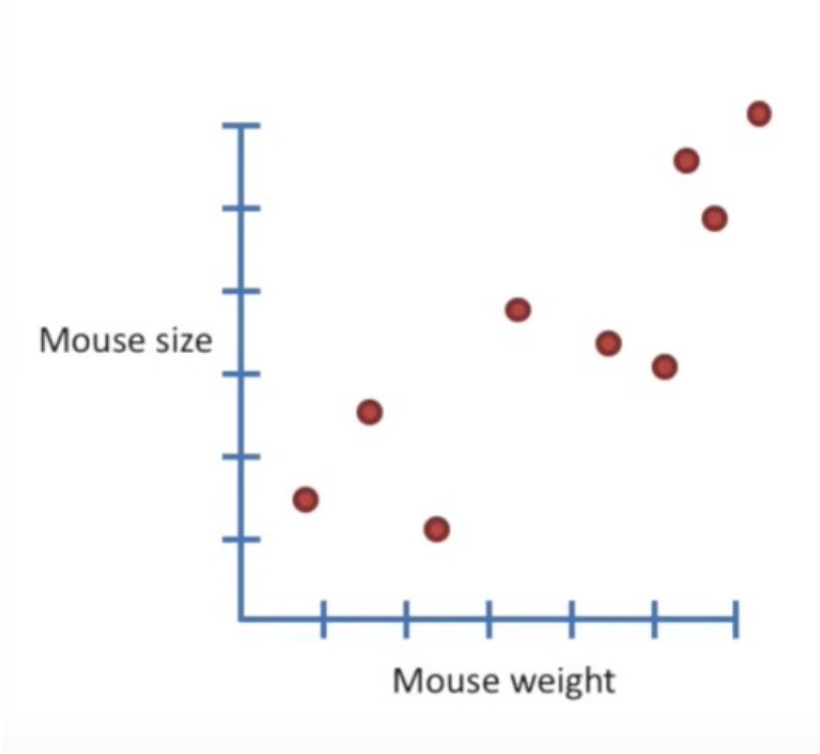


Expliquer & prédire !

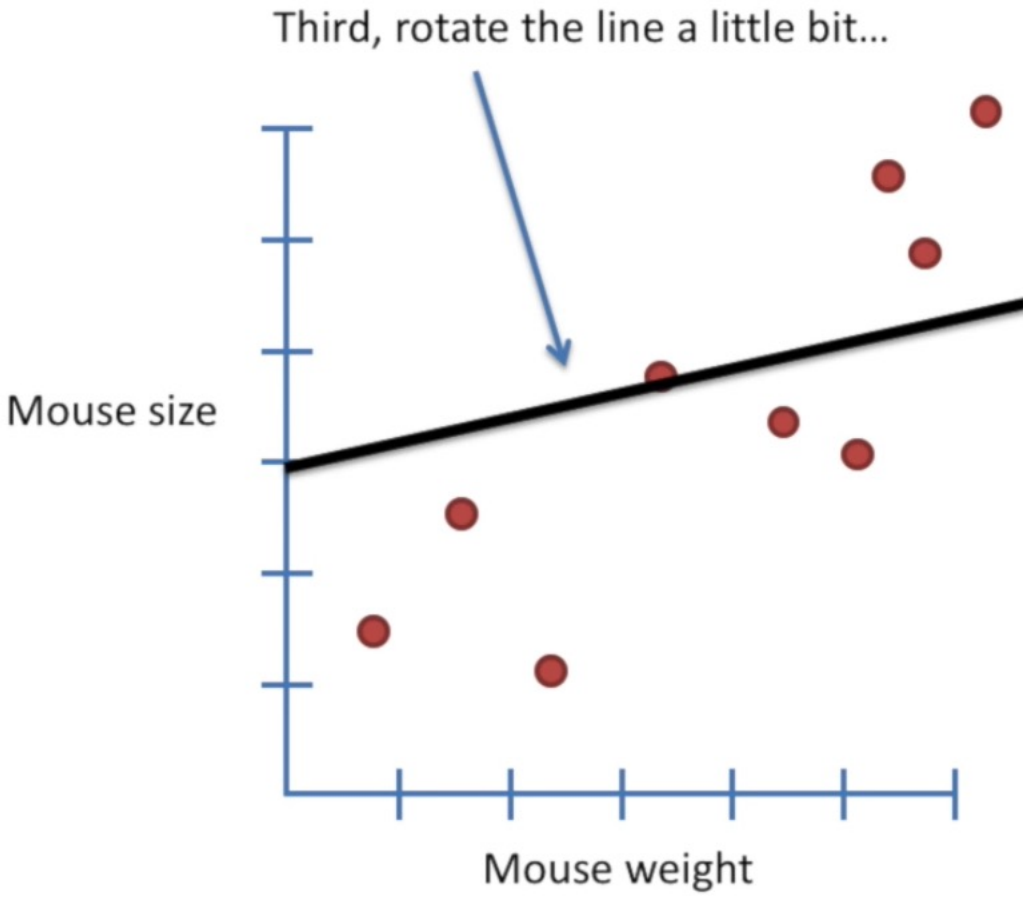
Modélisation d'une relation de type linéaire(Y=aX+b)

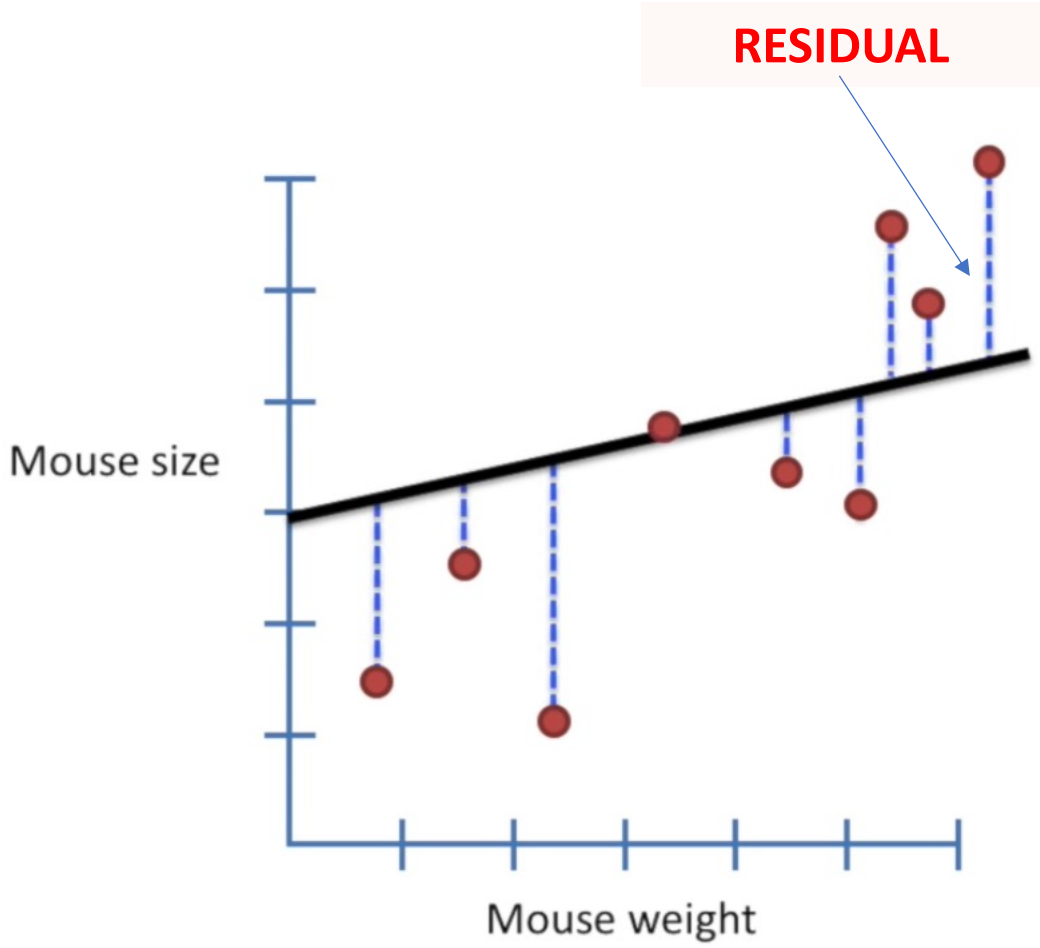
Modèle cherchant à établir une **relation linéaire** entre une variable, dite **expliquée/dépendante** (Y), et une autre dite **explicative/indépendante** (X).

Le poids de la souris permet-il de prédire correctement la taille (R^2)?
La relation est-elle due au hasard ? (p-value)

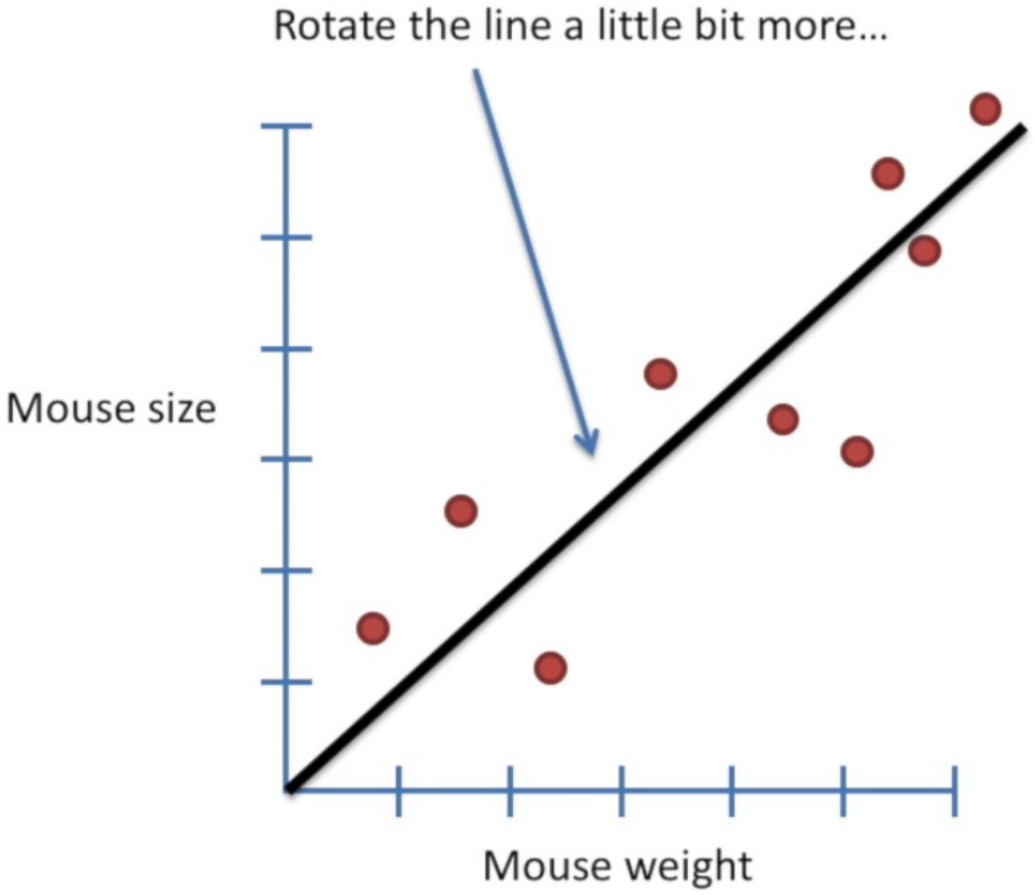


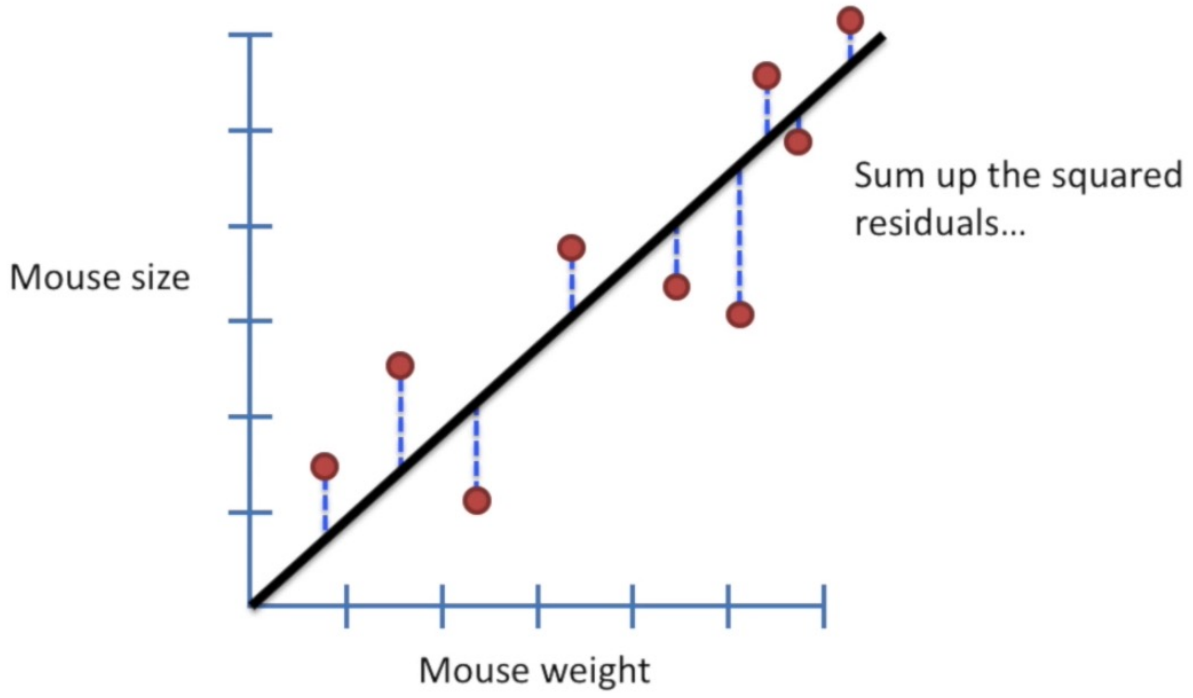
Méthode des moindres carrés





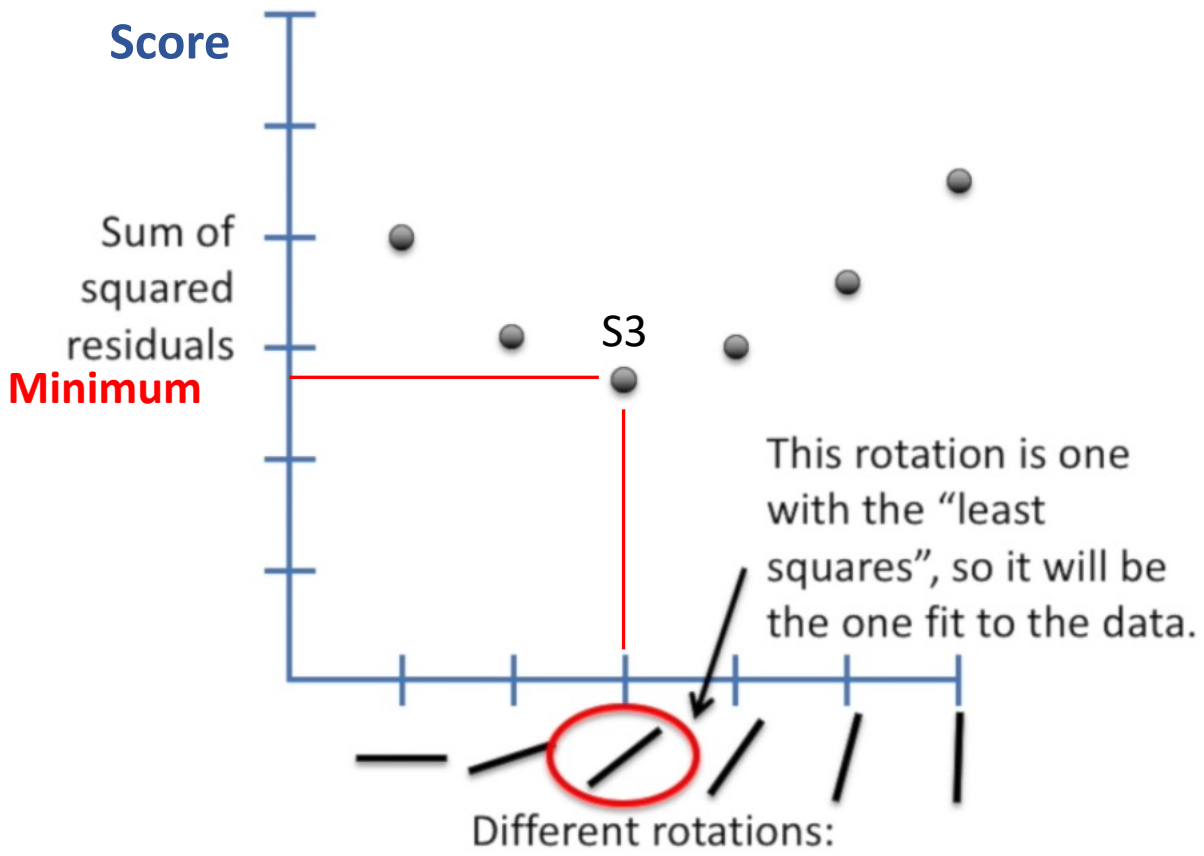
Méthode qui consiste à déterminer la droite dite « de régression de y en x » qui minimise la somme des distances des valeurs/points par rapport à la droite





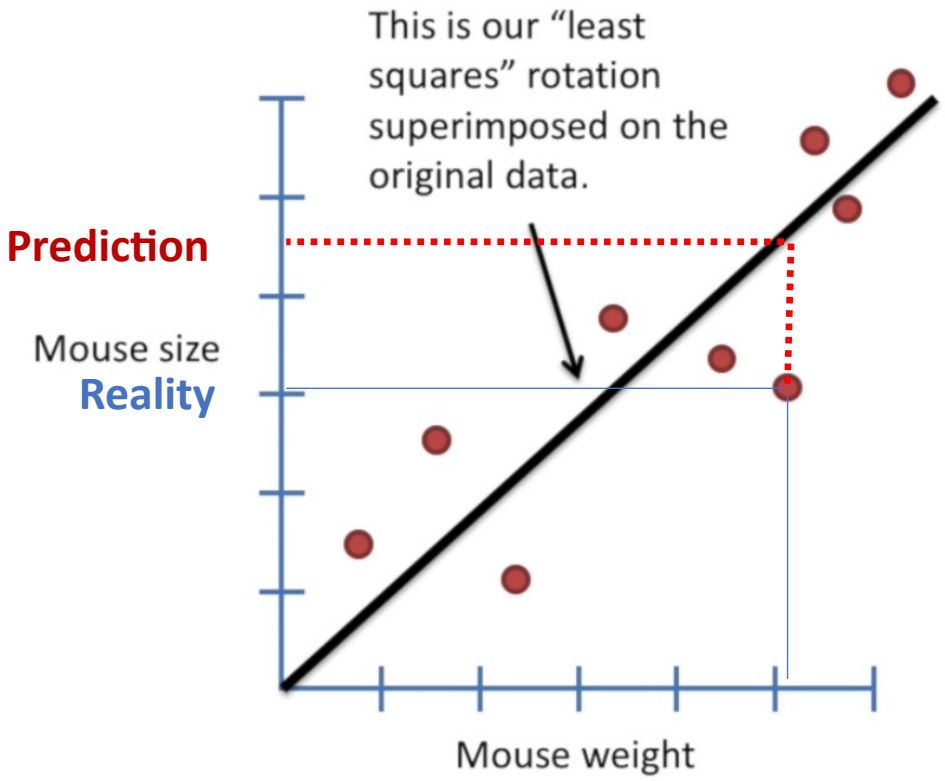
Again & again, recalculate

Resume : Sums of squared residuals for each rotation



Meilleure rotation (= position de la droite), celle qui minimise le score de la somme des carrés des distances résiduelles !!!!

Le poids de la souris permet-il de prédire correctement la taille (R^2)?



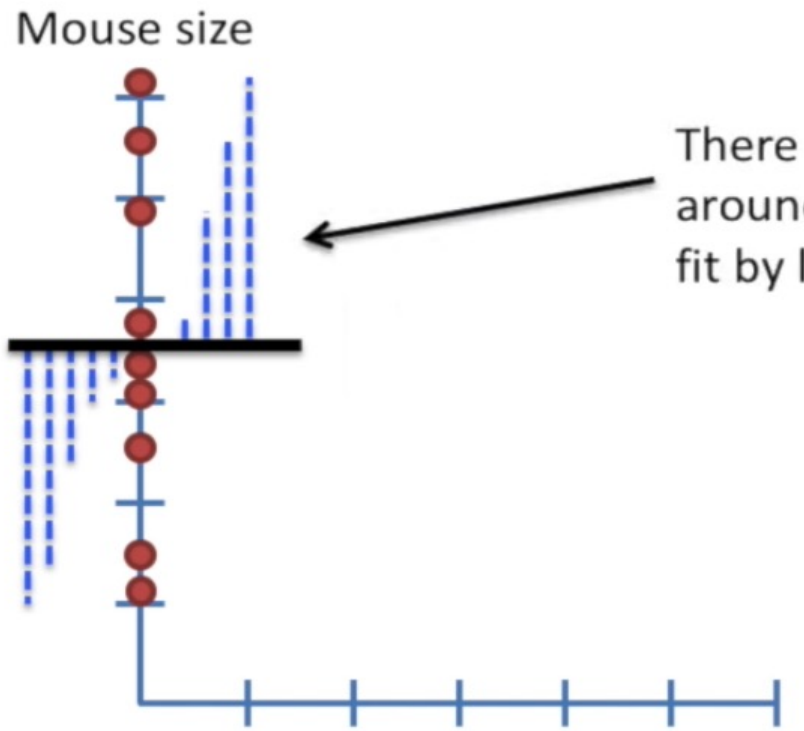
$$y = 0.1 + 0.78x$$

Dependence to « Mouse weight »

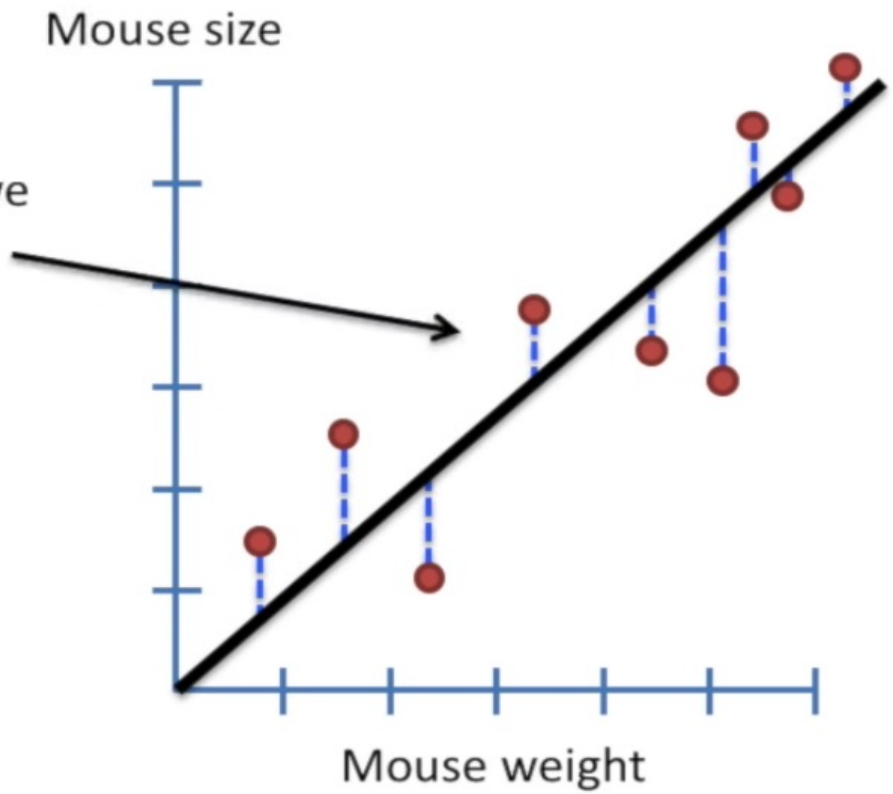
Coefficient R^2 = qualité de la prédiction

A quel point le modèle permet de correctement prédire la taille de la souris en tenant compte de son poids ?

R^2 : Détermination du Coefficient



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



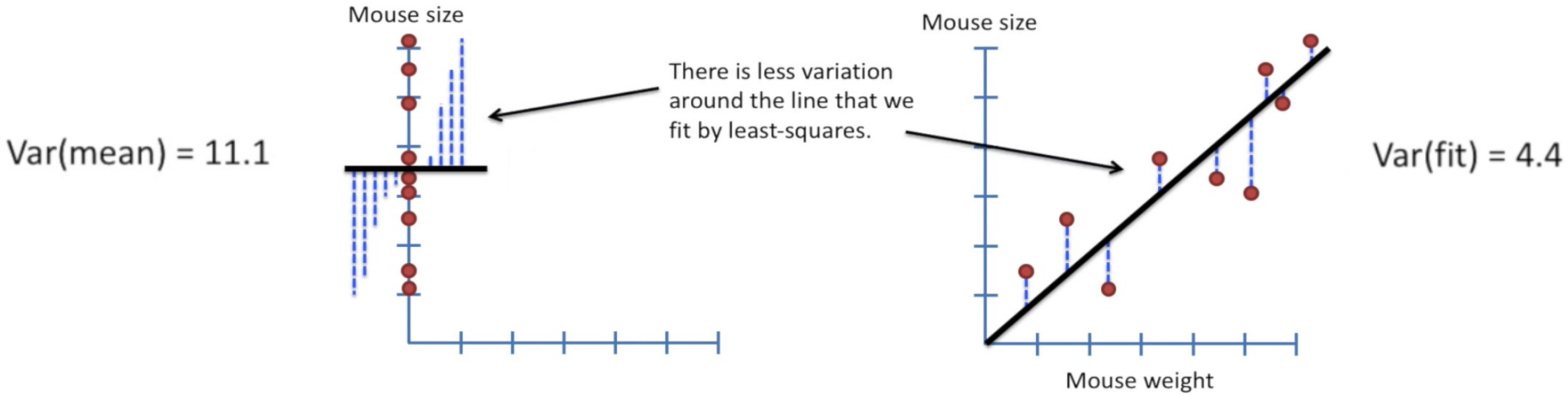
$$\text{Var(mean)} = \frac{\text{SS(mean)}}{n}$$

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

- **Taking into account « weight », less variations?? (SSfit < SSMean)!**

R^2 = % de variation de la variable réponse expliquée par un modèle linéaire (variable poids)

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



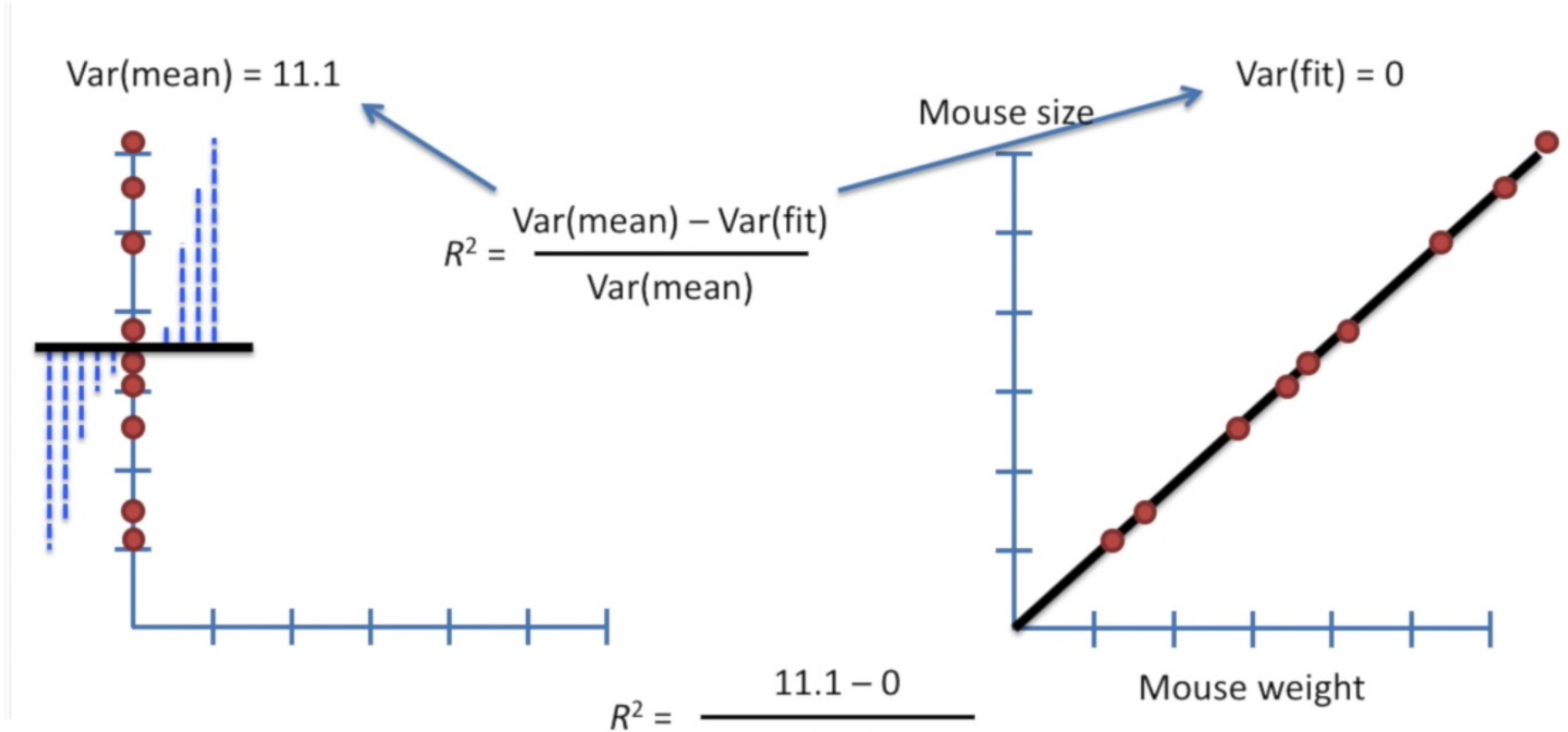
$$R^2 = \frac{11.1 - 4.4}{11.1} = 0.6 = 60\%$$

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

→ Le modèle établi explique 60% de la variabilité/variance de la "taille de la souris".

→ R^2 entre 0 et 1

TO be sure …


Var(mean) = 11.1
Var(fit) = 0
Mouse size
$R^2 = \frac{\text{Var(mean)} - \text{Var(fit)}}{\text{Var(mean)}}$
Mouse weight
$R^2 = \frac{11.1 - 0}{11.1}$
$R^2 = 1 = 100\%$

R^2 & significativité ?

- **Nécessite une p-value…**
- **Variance … la p-value est données par le ratio F & distribution F**

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

Relation entre r & R^2

Le coefficient de corrélation de Pearson r peut être lié à la régression linéaire R^2
Son carré est la variance expliquée par la régression(R^2)

r =0.5 -> R^2 = 0.25 -> 25% of the Y variance explained by X variable… ☹