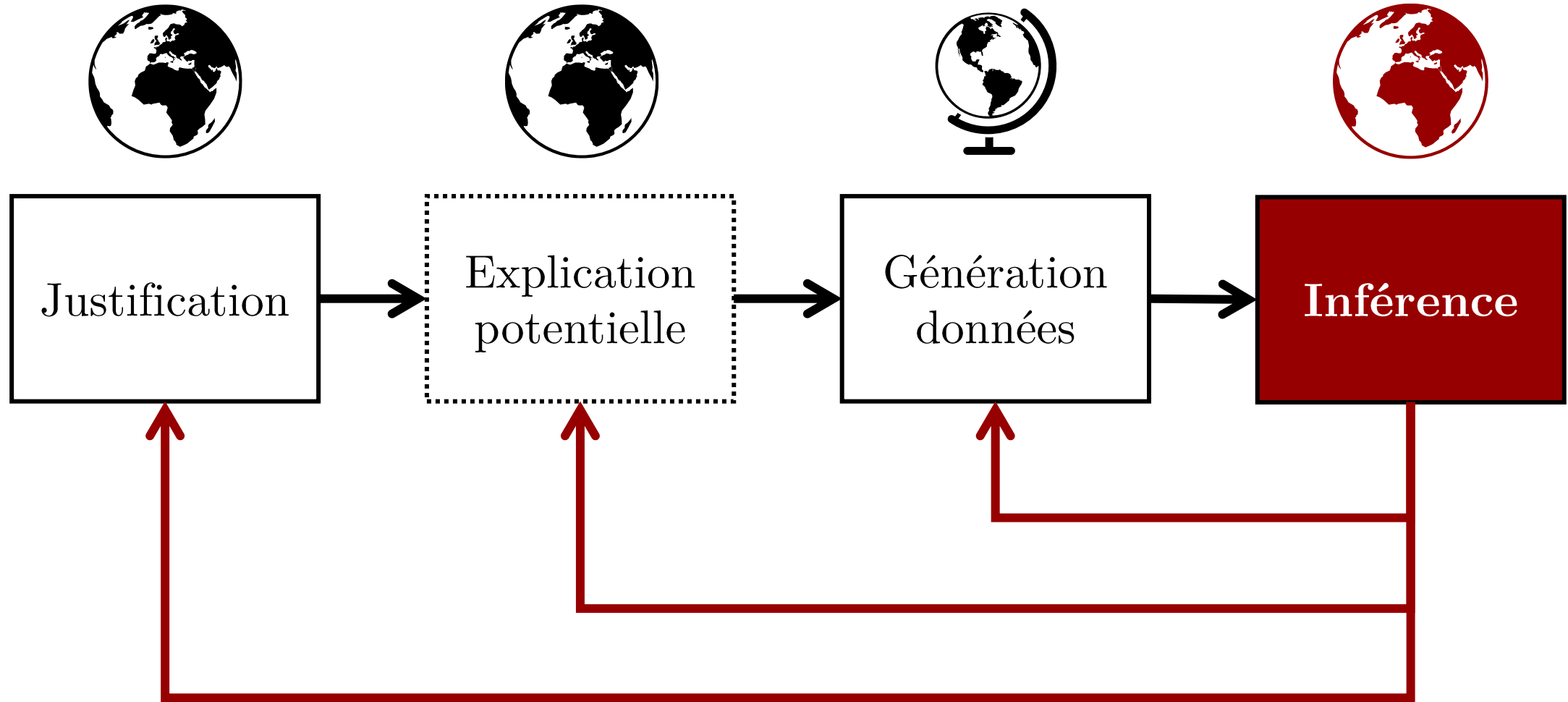


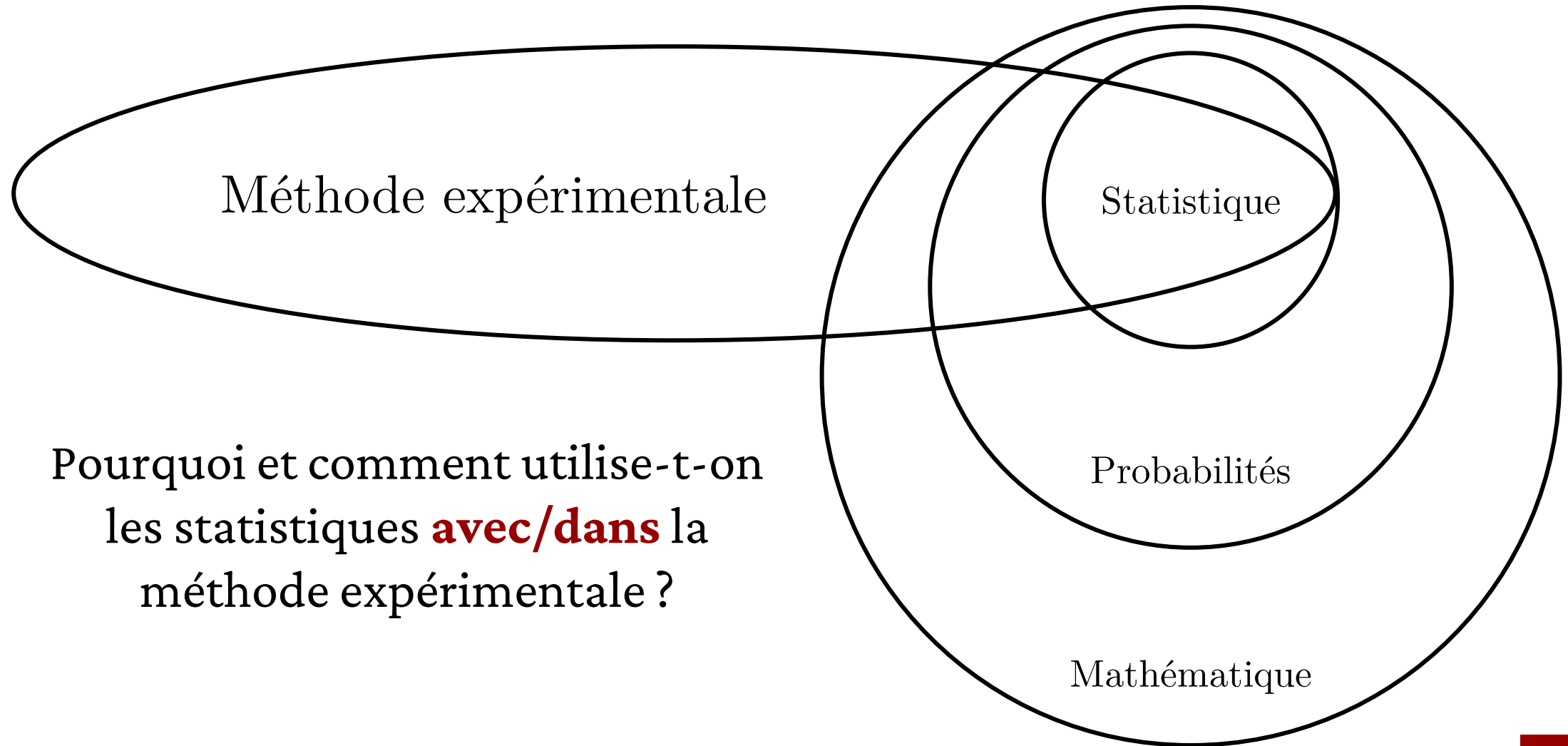
Fondements statistiques de la méthode expérimentale

Mattia A. Fritz
TECFA, Université de Genève

Étapes principales d'une expérience



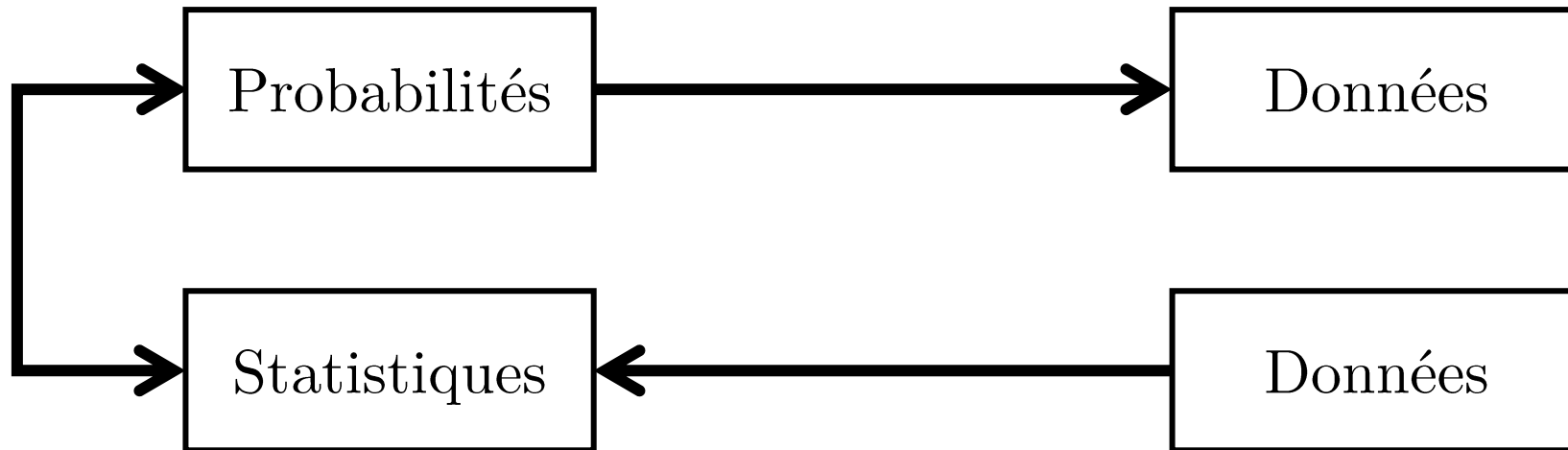
Articulation expérience-statistiques



Pourquoi et comment utilise-t-on
les statistiques **avec/dans** la
méthode expérimentale ?

Définition naïve de probabilité

Les événements qui peuvent se produire de plus de façons que d'autres sont plus probables/plausibles (et vice-versa).



Probabilités/statistiques et causalité

On peut utiliser le langage des probabilités/statistiques dans une perspective causale (e.g., Pearl, 2000 ; Pearl et al., 2016) :

- **Si $P(Y | X) = P(Y)$ alors X n'a pas d'effet sur Y**

Principe de la probabilité conditionnelle/indépendance des événements. Si la probabilité de réussir un examen avec/tenant compte de l'intervention X est égale à la probabilité sans intervention X, alors l'intervention n'a pas d'effet.

- **Si $P(Y | do(X)) \neq P(Y | X)$ alors il y a une 3ème variable**

S'il y a plus de probabilité que les élèves dans une expérience randomisée aient plus de facilité avec le logiciel A qu'avec le logiciel B, mais en observant deux classes non randomisées on voit le contraire, il y a probablement un effet d'une troisième variable (e.g. les compétences techniques de l'enseignant-e).

Raisons pour utiliser statistiques

- › **Pseudo-déterminisme dans un système complexe**

Étudier augmente les chances de réussir un examen, *mais* on peut échouer en étudiant ou réussir sans étudier

- › **Abstraction des particularités**

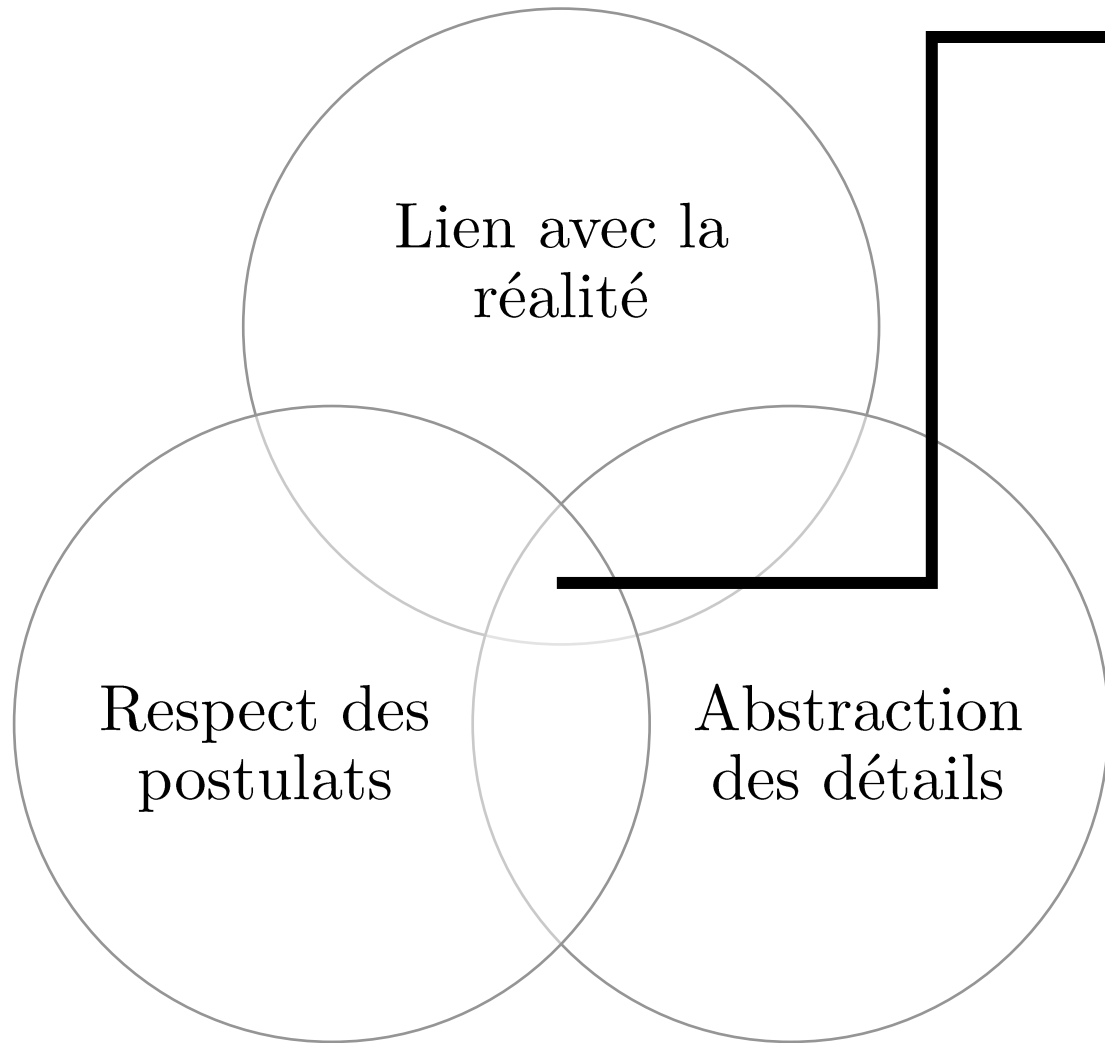
Étudier augmente les chances de réussir un examen, *si* l'examen porte sur les contenus du cours, *si* l'examen peut avoir lieu, *si* l'étudiant-e n'est pas malade, ...

- › **Précision de la **modélisation** mathématique**

« Les modèles mathématiques sont plus facilement **falsifiables**, ils forcent la **précision théorique**, leurs hypothèses peuvent être plus facilement étudiées, ils favorisent l'analyse des données et ils ont plus d'applications pratiques. »

– Rodgers, 2010, p.2, traduction libre

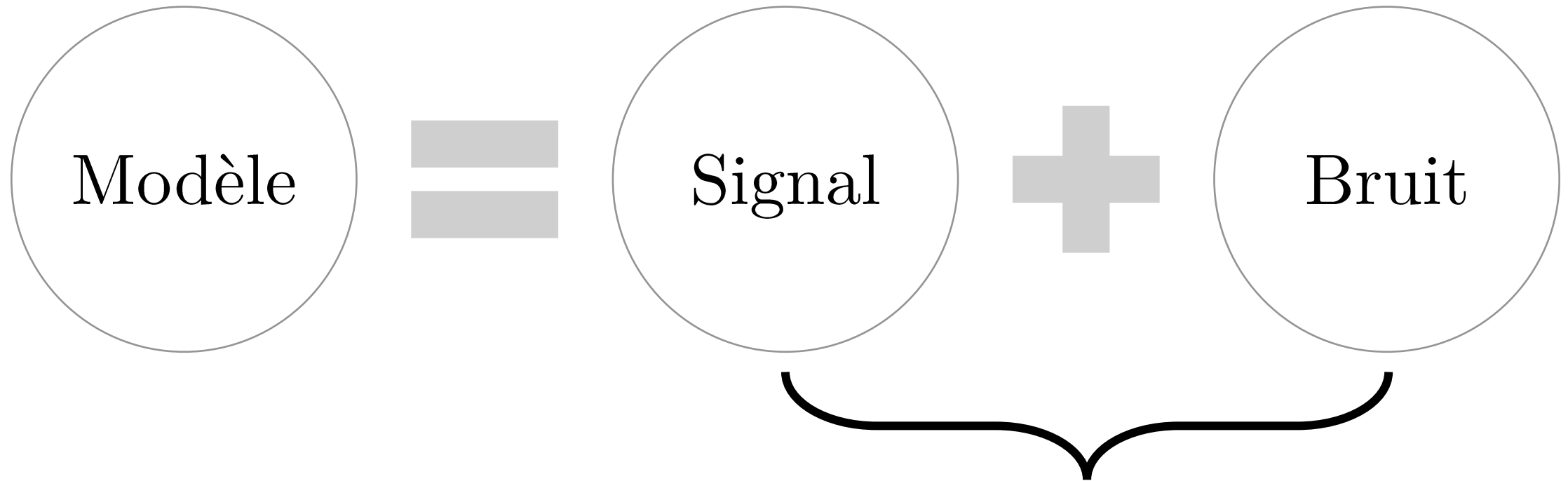
Qu'est-ce qu'un modèle ?



« Une représentation
idéalisée de la réalité
qui met en évidence
certains aspects et en
ignore d'autres »

– Pearl, 2000, p. 202
Traduction libre

Qu'est-ce que l'inférence statistique ?



Déterminer si le rapport entre le signal et le bruit est suffisamment élevé pour faire confiance au modèle.

3 modélisations en sciences sociales

- **Divergence par rapport à un modèle *nul***

Les données sont comparées à une modélisation basée sur le présupposé qu'il n'y a que du bruit. Si les données ne sont pas en adéquation avec le modèle *nul*, alors il y a probablement un signal.

- **Comparaison entre des modèles concurrents**

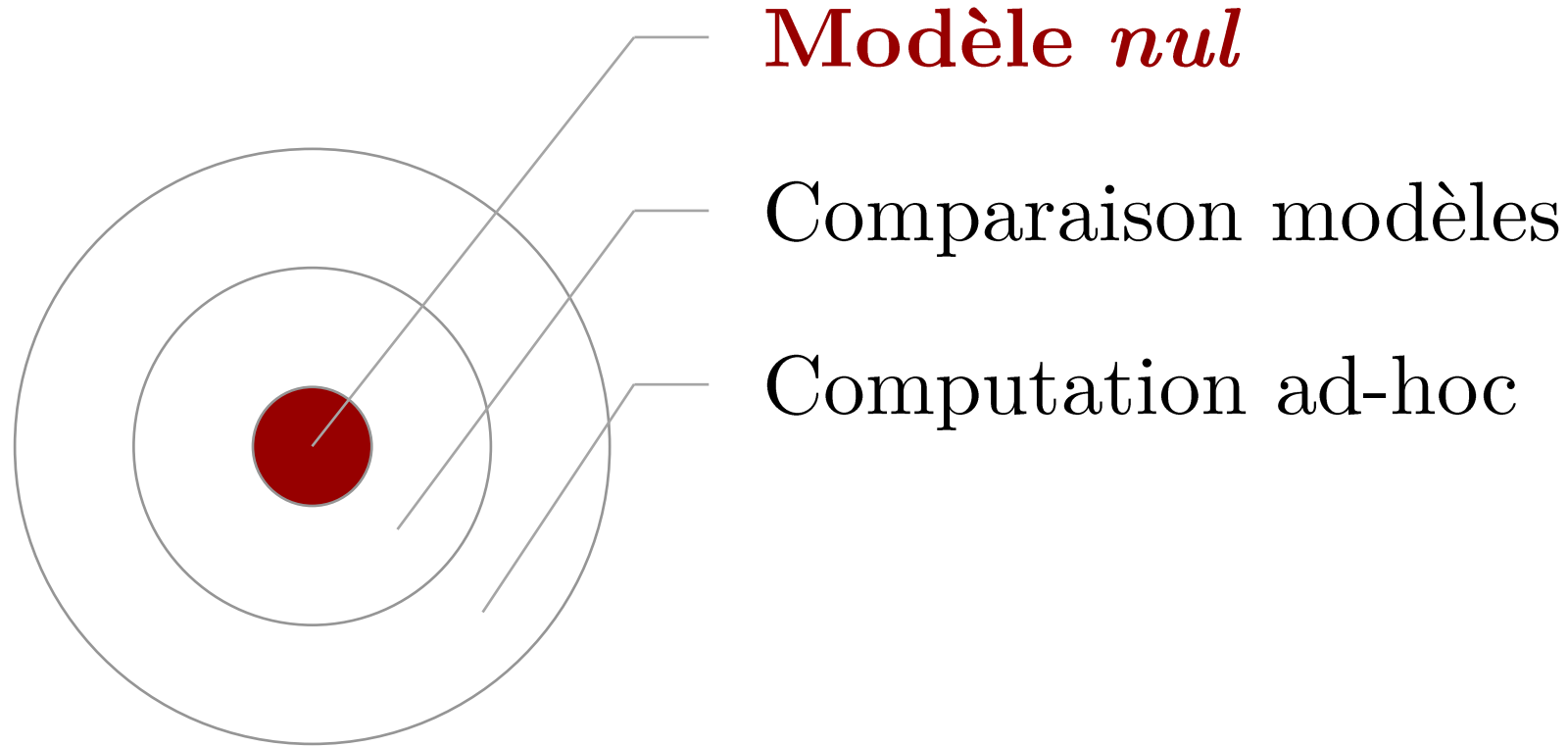
Deux modélisations ou plus sont comparées en termes de leur capacité à maximiser le signal et minimiser le bruit par rapport aux données observées.

- **Construction d'un modèle computationnel *ad-hoc***

La modélisation est construite à partir de paramètres dérivés théoriquement (e.g. modèle de la mémoire) et/ou empiriquement (e.g. *machine learning*).

Adapté de Rodgers (2010)

3 modélisations en sciences sociales



La divergence par rapport à un modèle *nul* est actuellement la **modélisation dominante** en sciences sociales, mais aussi **la moins informative**. On peut la considérer une forme basique de comparaison entre modèles (Rodgers, 2010)

2 philosophies inférentielles

- **Statistiques fréquentielles ou *classiques***

L'objectif est de prendre une décision, en essayant de limiter le risque de baser cette décision sur du bruit plutôt que sur le signal.

- **Statistiques Bayésiennes**

L'objectif est de mettre à jour des connaissances antérieurs en fonction de la plausibilité des nouvelles informations récoltées.



Le choix de traiter les statistiques fréquentielles est dû au fait qu'elles sont actuellement encore les plus répandue dans la littérature. D'un point de vue pédagogique, les statistiques Bayésiennes peuvent être plus intuitives (McElreath, 2020). Dans la boîte à outil des chercheurs il y a la place pour les deux, à utiliser selon des décisions épistémiques.

Statistiques fréquentielles

- › Approche de Ronald Fisher
- › **Approche de Jerzy Neyman et Egon Pearson**
- › *Null Hypothesis Significance Testing (NHST)*



Gravure de Cerbère et Héraclès par Antonio Tempesta.
Musée d'Art du comté de Los Angeles

Approche Neyman-Pearson

«Dans l'approche Neyman-Pearson, le but des tests statistiques est de **guider le comportement des chercheurs par rapport à une hypothèse**. Sur la base des résultats d'un test statistique, et sans jamais savoir si l'hypothèse est vraie ou non, les chercheurs choisissent d'agir provisoirement comme si **l'hypothèse nulle** ou **l'hypothèse alternative** était vraie. »

– Lakens, 2021, p. 1-2
Traduction libre

Hypothèse nulle et alternative

L'hypothèse **nulle** H_0 et **alternative** H_1/H_A dépendent de ce qu'on veut tester/savoir :

- **Présence d'un effet (très souvent)**

H_0 : Il n'y a pas d'effet, par exemple $M(VD \mid VI_0^1) - M(VD \mid VI_1^1) = 0$

H_1 : Il y a un effet

- avec hypothèse *non directionnelle* $M(VD \mid VI_0^1) \neq M(VD \mid VI_1^1)$,
- avec hypothèse *directionnelle* $M(VD \mid VI_0^1) > M(VD \mid VI_1^1)$ – ou vice-versa

- **Absence d'un effet (plus rarement)**

H_0 : Il existe un effet inférieur ou supérieur à un certain seuil

H_1 : L'effet est entre deux limites qui le caractérisent comme *ininfluent*

Types d'utilisations des statistiques

Dans une expérience, on utilise les statistiques pour :

- **Analyse de puissance statistique (*si possible*)**

On détermine la **taille de l'échantillon** minimale nécessaire pour détecter la présence ou décreter l'absence d'un effet de X sur Y (i.e., tester l'hypothèse)

- **Modélisation du micro-monde**

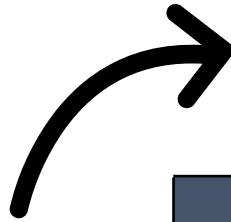
Statistiques/graphiques qui illustrent les **caractéristiques de l'échantillon** (nombre d'observations retenues, moyenne, écart type, ...)





- **Modélisation inférentielle du macro-monde**

Déterminer la **présence** (ou **absence**), la **direction**, la **magnitude** et l'**incertitude** de l'effet de X sur Y sur la base de la relation entre le(s) VI et la VD

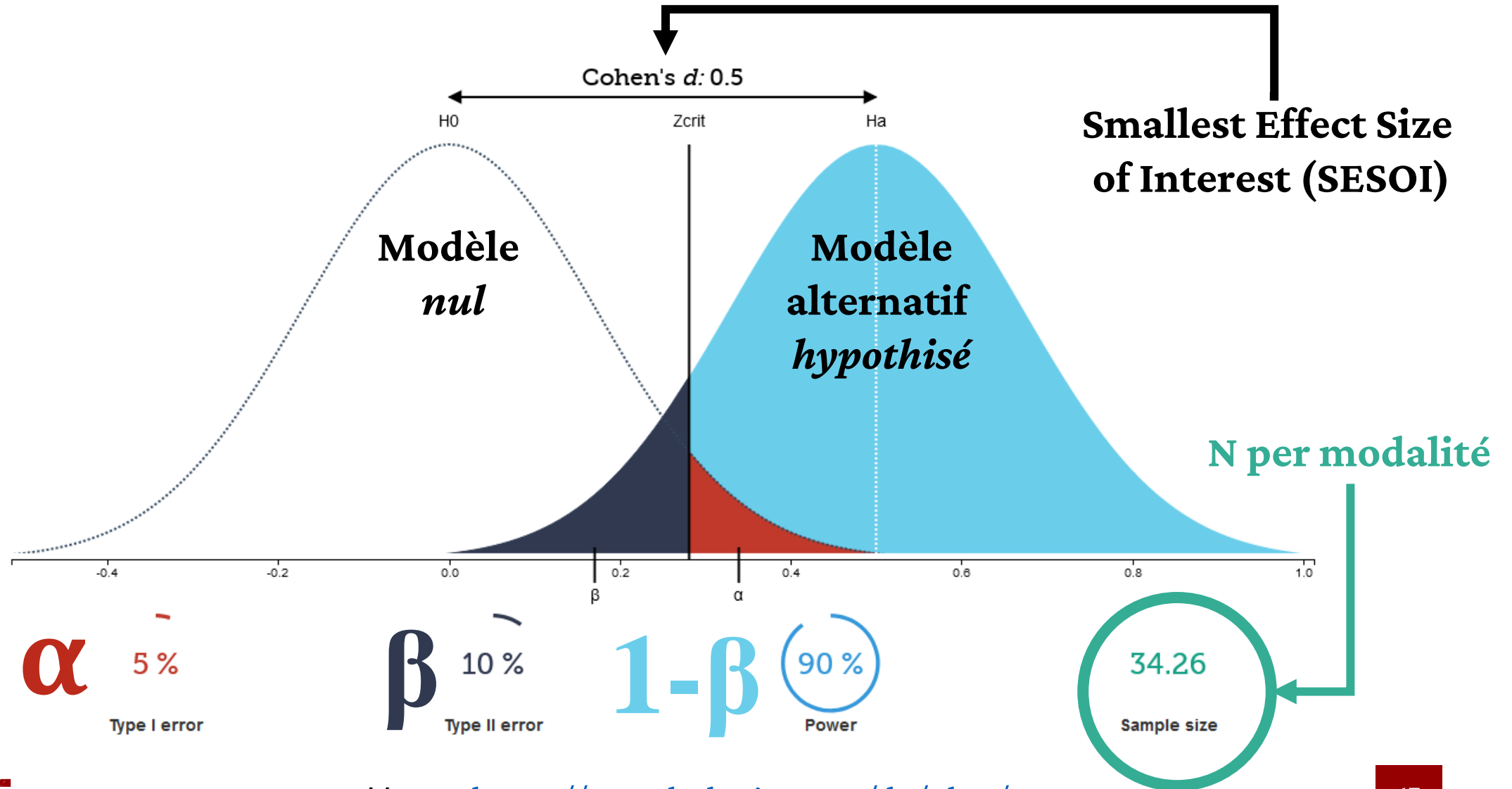
Possibilités dans un test d'hypothèse

Inférence



	Effet de X sur Y	Pas d'effet de X sur Y
Effet de VI sur VD	 Inférence correcte	 Erreur de Type I
Pas d'effet de VI sur VD	 Erreur de Type II	 Inférence correcte

Analyse de puissance statistique



Comment déterminer le SESOI ?

- › **Par rapport à la littérature**

Tailles des effets (ou M et SD) disponibles dans d'autres contributions. Attention aux études pilotes : quand N est petit, l'incertitude autour de la taille est grande

- › **Par rapport aux connaissances du domaine**

Quel effet minimal est considéré intéressant théoriquement/pratiquement ? E.g., si on applique l'intervention X_1 , quel gain d'apprentissage le justifie-t-il vs. X_0 ?

- › **Seuils conventionnels/sugérés**

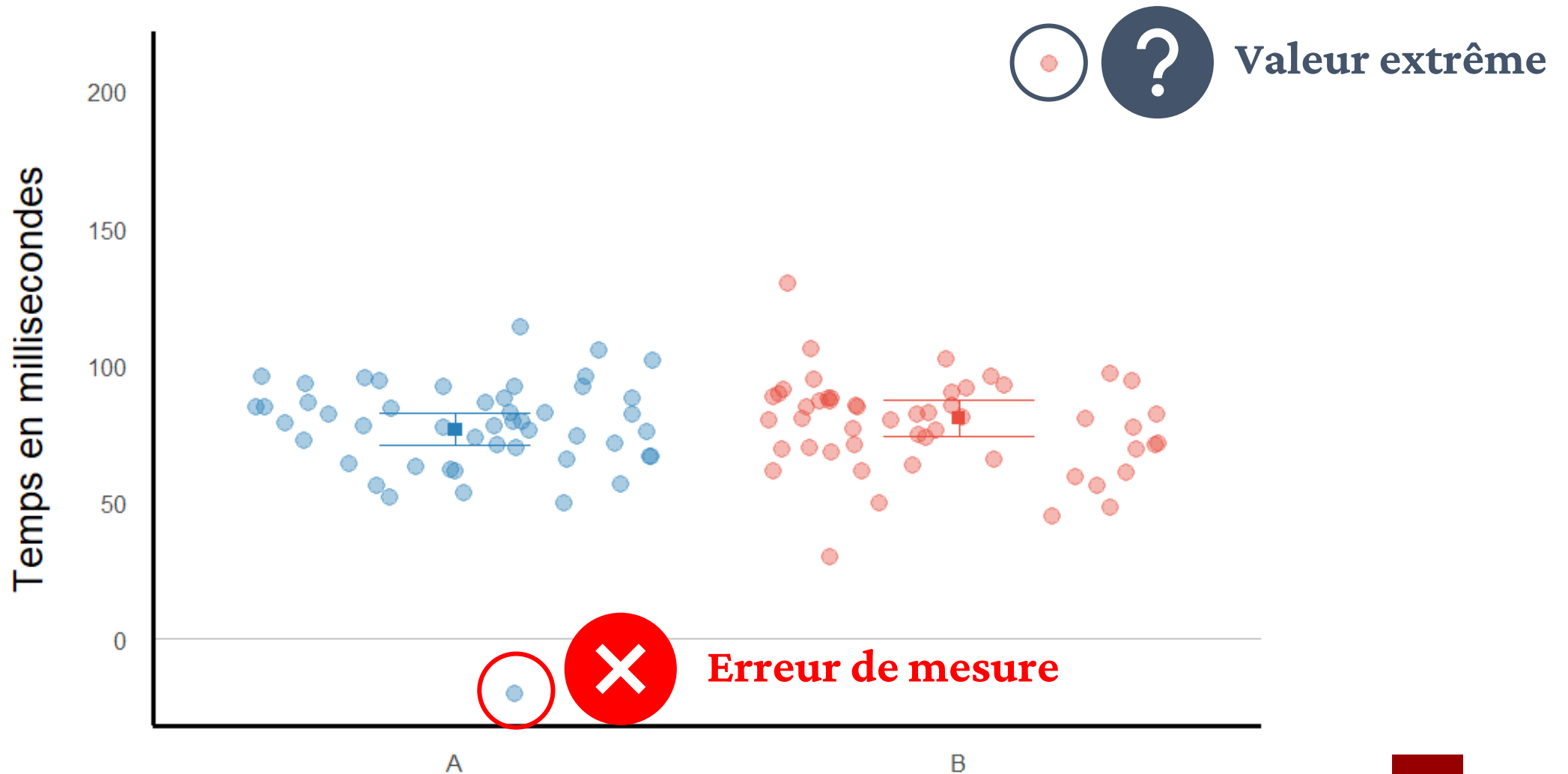
Il existe des valeurs suggérées dans la littérature qui dépendent du type de test mené et de la famille de la taille d'effet adoptée, à utiliser avec précaution !

Modélisation du micro-monde

« Les quantités numériques se concentrent sur les valeurs attendues, les résumés graphiques sur les valeurs inattendues. »

– John Tukey
Traduction libre

Exploratory Data Analysis

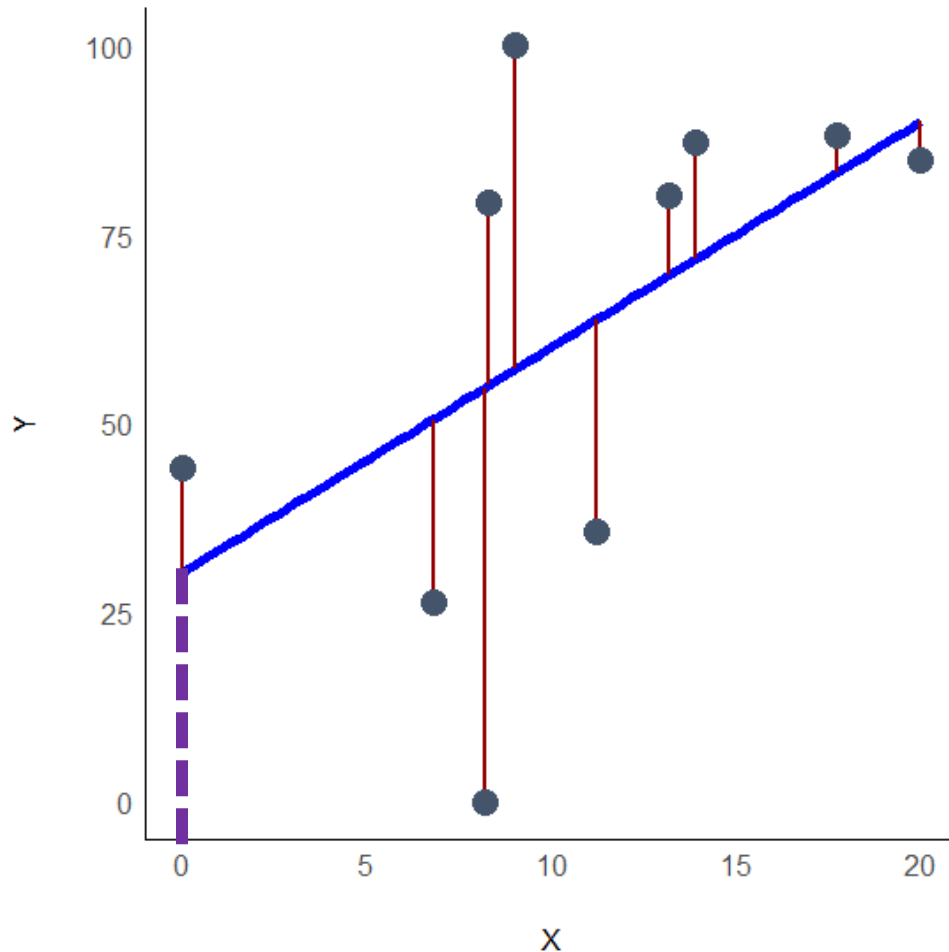


Caractéristiques de l'échantillon

	N	M	SD
Groupe A	48	78.6	14.5
Groupe B	53	80.7	24.6
Total	101	79.7	20.4

Elles sont appelées souvent « **statistiques descriptives** »,
mais elles sont déjà **une forme de modélisation**.

Modélisation du macro-monde



Les tests statistiques adoptés en sciences sociales sont basés très souvent sur le **modèle linéaire**.

$$Y_i = \underbrace{\beta_0 + \beta_1(X_i)}_{\text{Parameters}} + \varepsilon_i$$

Intercept Slope Residual

Paramètres : estimation dans le macro-monde

Cas spéciaux, mais c'est le même test

Nom du test	Type de VI	Type de VD
Regréssion linéaire simple	1 continue	1 continue
Régression linéaire multiple	2+ continues (et 1+ discrètes)	1 continue
Welch (ou Student) <i>t</i> -test	1 discrète avec 2 modalités	1 continue
ANOVA simple (one-way)	1 discrète avec 2+ modalités	1 continue
ANOVA factorielle	2+ discrètes	1 continue
ANCOVA	1+ discrètes et 1+ continues	1 continue

Modèles plus articulés

Les cas particuliers du modèle linéaire **ne s'adaptent souvent pas** à des design avec **mesure répétée** et/ou avec des **entités hiérarchisées** (e.g. binôme dans une tâche, étudiant-es dans des classes, ...), car certains postulats ne sont pas satisfaits . À ce moment on utilise plutôt des modèles linéaires multi-niveaux (Brown, 2021) :

$$Y = X\beta + Ku + \varepsilon$$

Fixed effects

(Intervention et facteurs d'influence)

Random effects

(Mesures répétées et/ou structure hiérarchique)

Indicateurs dans un test statistique

› Degrés de liberté et résultat du test statistique

Les degrés de liberté dépendent du nombre de VI/modalités et d'observations utilisées et déterminent la distribution *nulle* de référence pour le résultat du test.

› *p*-valeur associée au résultat du test statistique

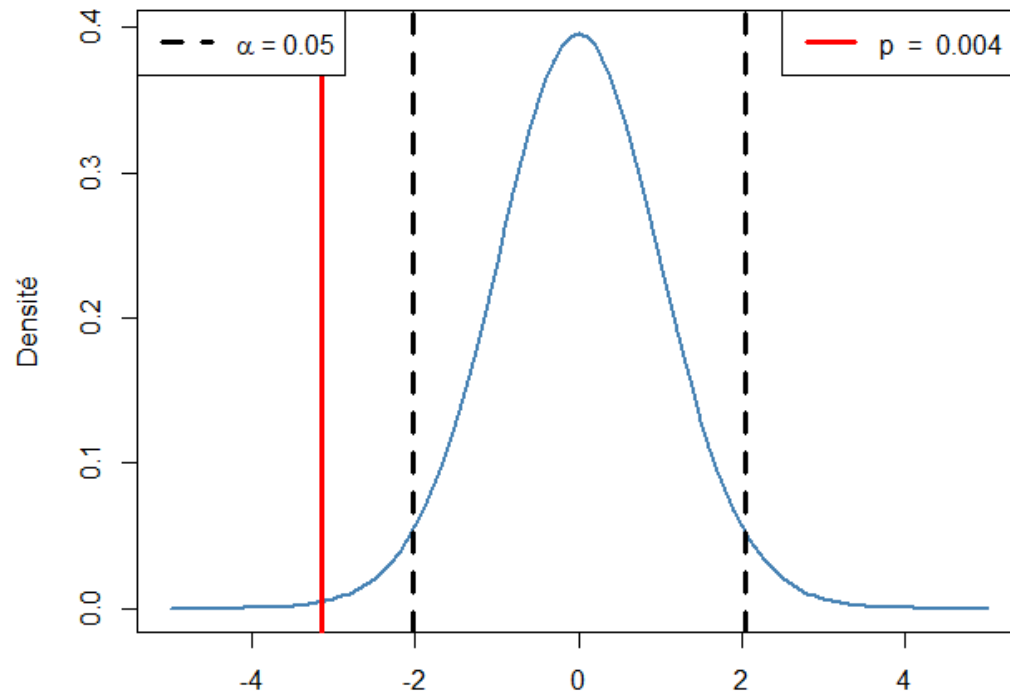
La *p*-valeur correspond à la probabilité d'obtenir des données aussi divergentes ou encore plus divergentes du modèle *nul* de celles observées dans l'échantillon **si l'hypothèse nulle était vraie.**

› Taille de l'effet brute et standardisée

La taille de l'effet brute est indiquée en utilisant l'échelle de la VD. La taille de l'effet standardisée utilise un indicateur de type ***d*** (différence standardisée entre moyennes) ou ***r*** (variance expliquée par la VI ou force de la relation VI-VD).

Exemples de distributions *nulles*

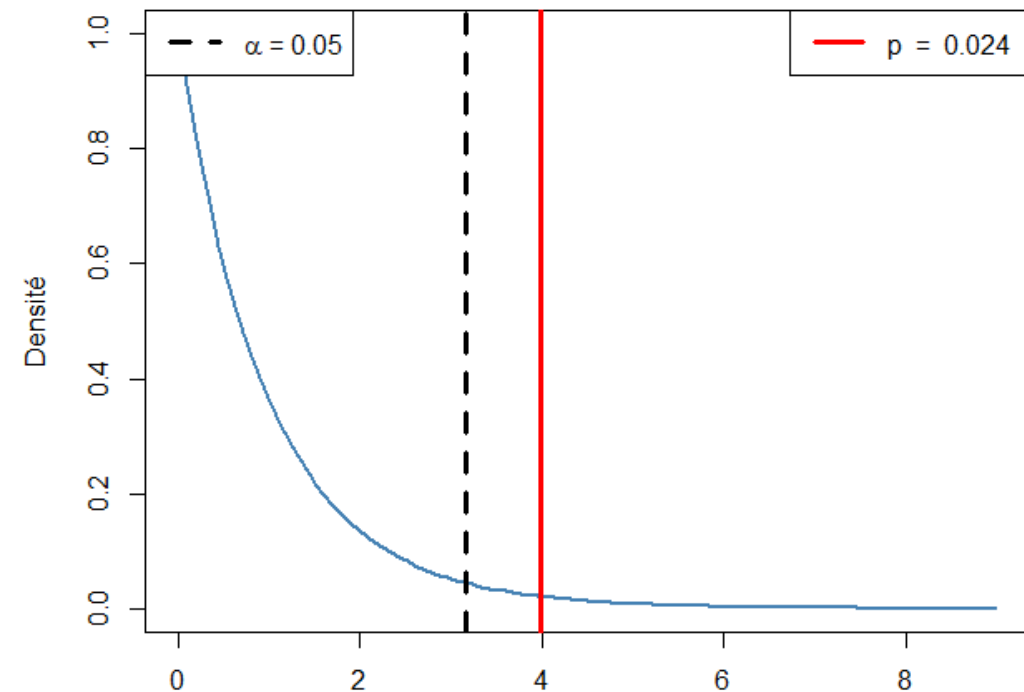
Distribution t (df = 32.99)



Student's t-distribution

Un résultat à gauche/droite des lignes en tirets est considéré surprenant : **rejet H_0**

Distribution f (df1 = 2 df2 = 57)

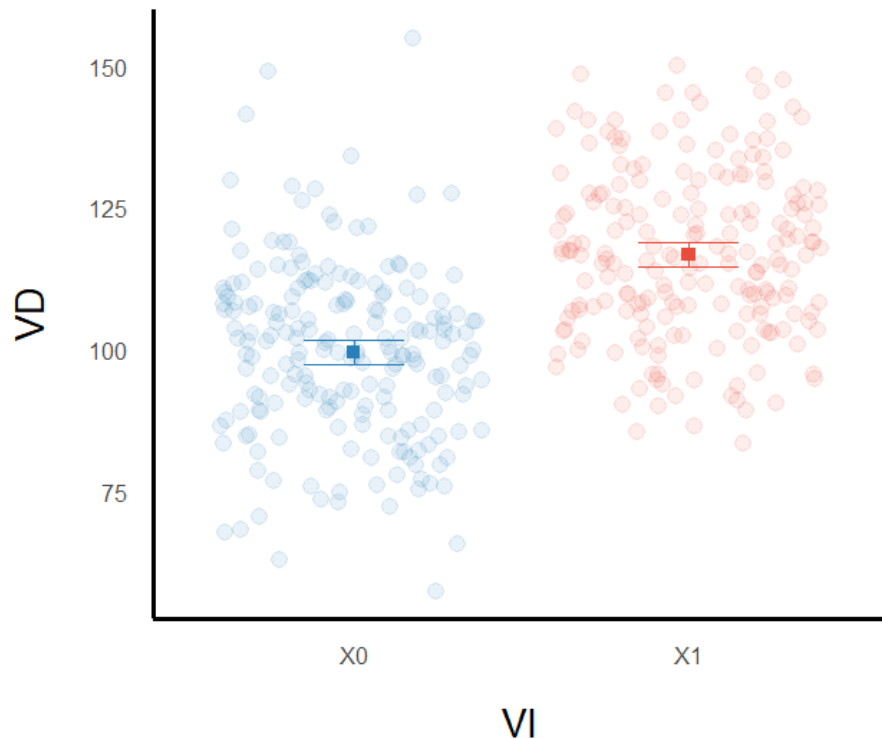


F-distribution

Un résultat à droite de la ligne en tirets est considéré surprenant : **rejet H_0**

Hypothèse : effet existe (i.e. > SESOI)

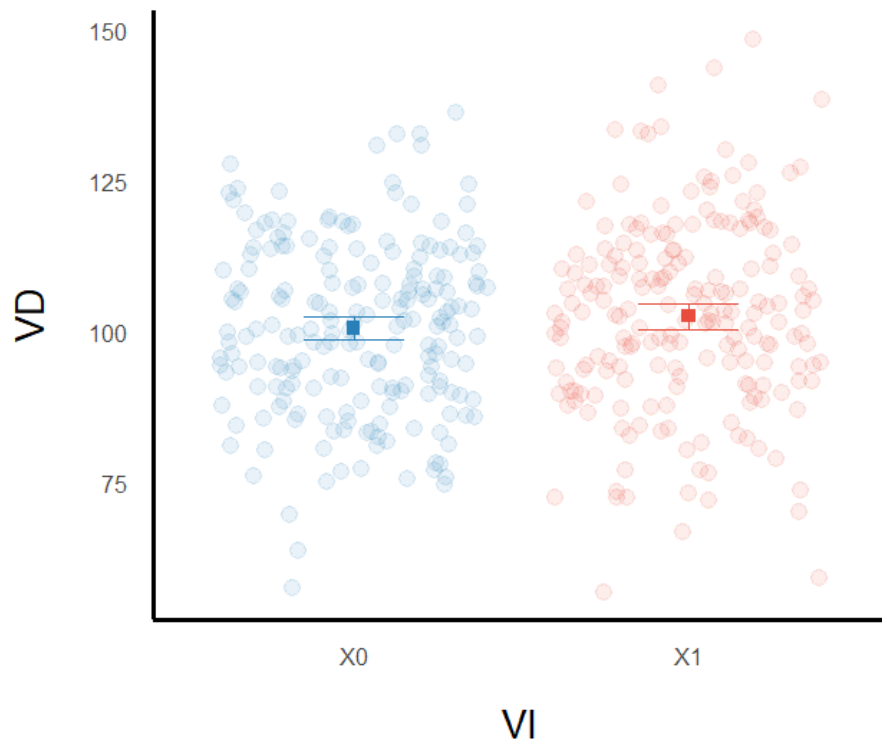
Quand **p-valeur** < α (e.g. 0.05), **H_1 acceptée/corroborée ***



- › Degrés de liberté et résultat test statistique
 $t(397.26) = -11.42$
- › P-valeur associée au test statistique
 $p < .001$
- › Taille de l'effet brute
 $\Delta M = -17.26$, 95% CI $[-20.23, -14.29]$
- › Taille de l'effet standardisée
Cohen's $\delta = -1.15$, 95% CI $[-1.36, -0.93]$

Hypothèse : effet existe (i.e. > SESOI)

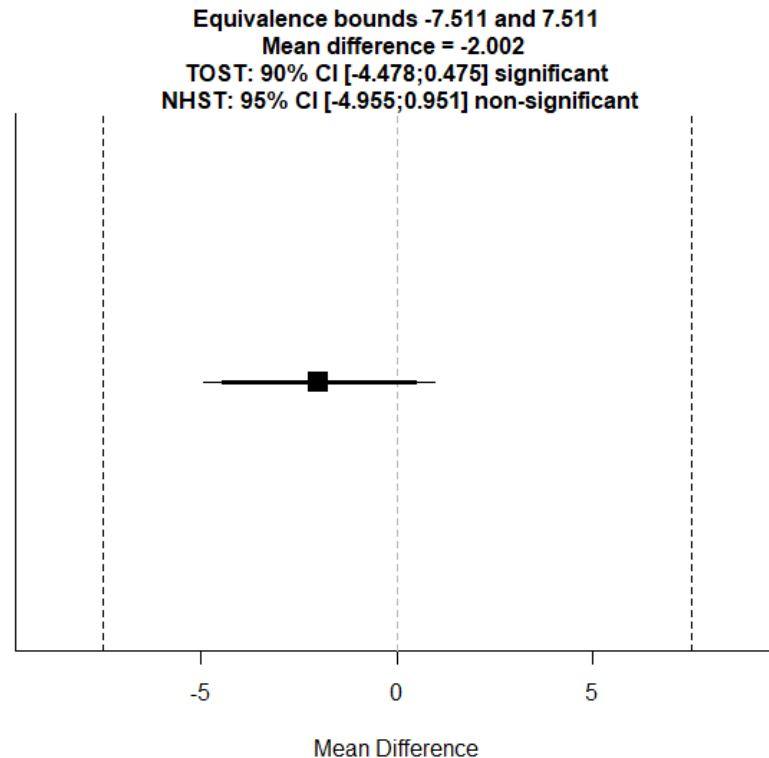
Quand **p-valeur** > α (e.g. 0.05), **H_1 rejetée/infirmée ***



- › Degrés de liberté et résultat test statistique
 $t(393.39) = -1.33$
- › P-valeur associée au test statistique
 $p = .183$
- › Taille de l'effet brute
 $\Delta M = -2.00$, 95% CI $[-4.96, 0.95]$
- › Taille de l'effet standardisée
Cohen's $\delta = -0.13$, 95% CI $[-0.33, 0.06]$

Hypothèse : effet n'existe pas |limites|

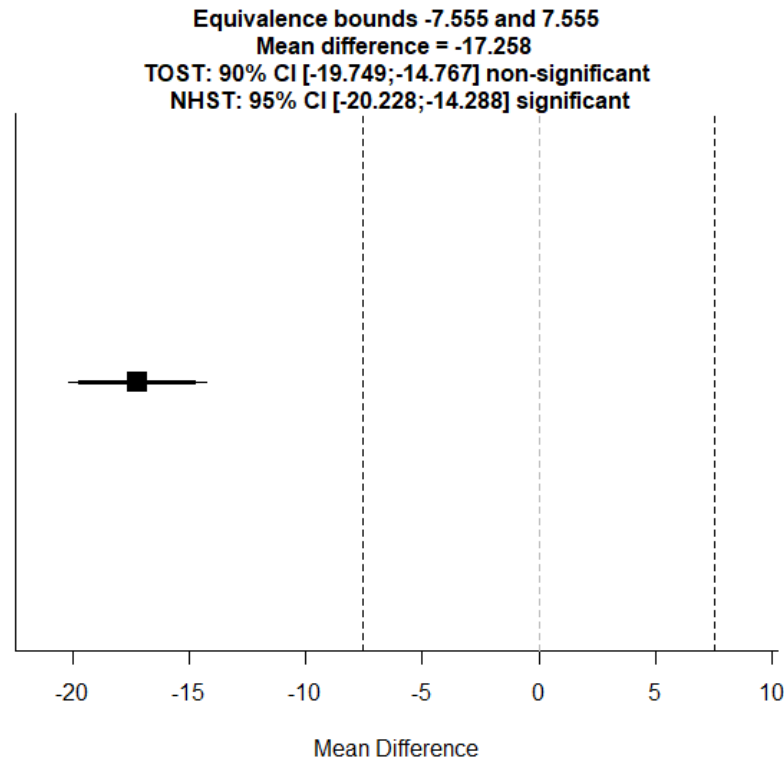
Quand **p-valeur** $< \alpha$ (e.g. 0.05), **H_1 acceptée/corroborée ***



- › **SESOI limites inférieur et supérieur**
Cohen's δ entre -0.5 et 0.5
- › **Degrés de liberté et résultat test statistique**
 $t(397.26) = -6.42$
- › **P-valeur associée au test statistique**
 $p < 0.001$
- › **Taille de l'effet brute**
 $\Delta M = -2.00$, 90% CI $[-4.48, 0.48]$

Hypothèse : effet n'existe pas |limites|

Quand **p-valeur** $> \alpha$ (e.g. 0.05), **H_1 rejetée/infirmée ***



- › **SESOI limites inférieur et supérieur**
Cohen's δ entre -0.5 et 0.5
- › **Degrés de liberté et résultat test statistique**
 $t(397.26) = -6.42$
- › **P-valeur associée au test statistique**
 $p = 1.00$
- › **Taille de l'effet brute**
 $\Delta M = -17.26$, 90% CI [-19.75, -14.77]

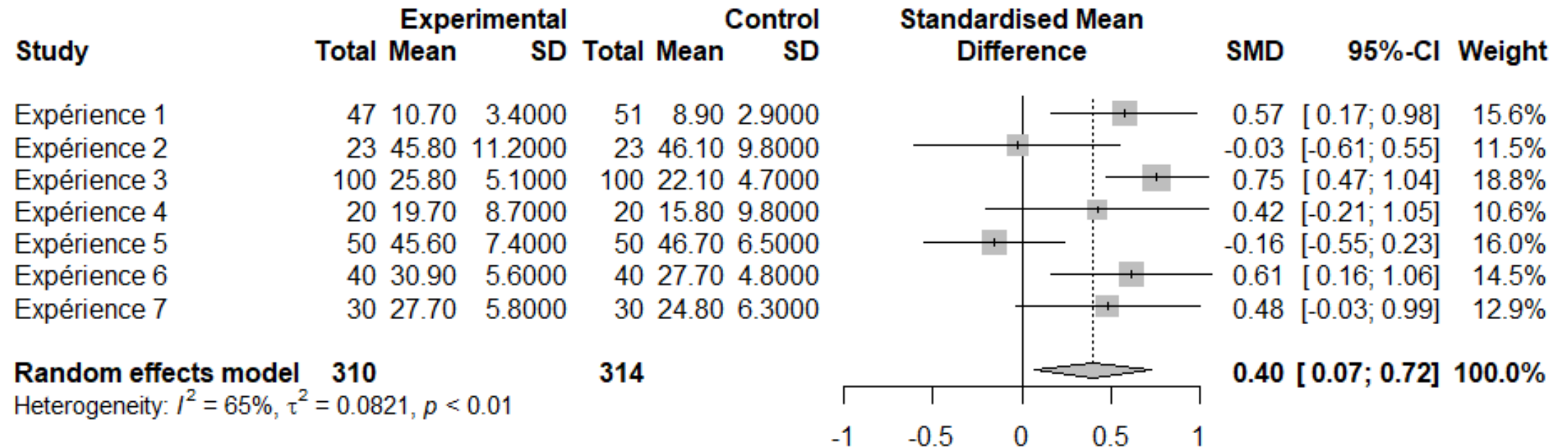
Combiner effet et équivalence

On peut combiner les deux tests dans la même expérience (Lakens, 2018), avec trois résultats possibles :

- › **Le test sur la présence de l'effet est inférieur à α**
- › **Le test sur l'absence de l'effet est inférieur à α**
- › **Les deux tests sont supérieurs à α**

Expérience inconcluante (trop de bruit, SESOI sur-estimé, ...)

Méta-analyse : plusieurs expériences



Combiner de **manière systématique** plusieurs études/expériences sur un même sujet (e.g. intervention similaire) et calculer **un effet « cumulé/pondéré »** qui tient compte du poids de chaque échantillon.

Inférence/implication *pratique*

La discussion dépend de l'ensemble de l'expérience :

- **Problèmes dans la génération/récolte/analyse**

Est-ce que des éléments dans le processus de génération des données peuvent biaiser les résultats du test statistique ?

- **Discussion sur la base de la taille de l'effet**

Baser l'inférence sur la taille de l'effet brute permet de raisonner en termes très pratiques (effet de l'intervention sur des unités de la mesure du phénomène).
L'effet standardisé permet de se faire une idée générale de la magnitude.

- **Incertitude autour des effets même si $p < \alpha$**

Des larges intervalles autour d'un effet suggèrent précaution (hétérogénéité).

Conclusion

- › Maîtriser les statistiques est **compliqué** et nécessite de temps/pratique
- › Comprendre la **logique** est plus important de mémoriser les procédures/détails
- › Sans **connaissances dans le domaine**, les chiffres ne peuvent nous dire rien d'intéressant !

Failing Grade: 89% of Introduction-to-Psychology Textbooks That Define or Explain Statistical Significance Do So Incorrectly



Scott A. Cassidy, Ralitza Dimova, Benjamin Giguère, Jeffrey R. Spence^{id}, and David J. Stanley
Department of Psychology, University of Guelph

Advances in Methods and
Practices in Psychological Science
2019, Vol. 2(3) 233–239
© The Author(s) 2019
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/2515245919858072
www.psychologicalscience.org/AMPP
SAGE

The prevalence of statistical reporting errors in psychology (1985–2013)

Michèle B. Nuijten¹ • Chris H. J. Hartgerink¹ • Marcel A. L. M. van Assen¹ •
Sacha Epskamp² • Jelte M. Wicherts¹

RESEARCH ARTICLE SUMMARY

PSYCHOLOGY

Estimating the reproducibility of psychological science

Open Science Collaboration*

substantial decline. Ninety-seven percent of original studies had significant results ($P < .05$). Thirty-six percent of replications had significant results; 47% of original effect sizes were in the 95% confidence interval of the replication effect size; 39% of effects were subjectively rated to have replicated the original result; and if no bias in original results is assumed, combining original and replication

ON OUR WEB SITE
Read the full article at <http://dx.doi.org/10.1126/science.aac4716>

Merci pour votre attention !

Mattia A. Fritz

TECFA, Université de Genève

mattia.fritz@unige.ch



This work is licensed under Attribution-NonCommercial-ShareAlike 4.0 International.
To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc-sa/4.0/>