Choisir une taille d'échantillon à l'aide de l'analyse de puissance

(presque sans maths!)

Atelier pratique - CIRRIS 9 février 2023
Michel-Pierre Coll

Présentation:
https://github.com/mpcoll/2023_puissance_cirris

Objectifs de l’atelier

- Comprendre la nécessité d’une approche réfléchie dans le choix d’une taille d’ échantillon.
- Comprendre les conséquences d’un manque de puissance sur l’inférence statistique.
- Comprendre la logique de l’analyse de puissance a priori.
- Comprendre la démarche pour le choix des paramètres d’une analyse de puissance a priori (taille d’effet, seuil alpha, taille d’échantillon).

Choix du nombre d'unités expérimentales

- Le choix du nombre d'unités expérimentales est une étape cruciale de la planification d'une étude quantitative.
- Choisir un nombre **trop petit** d'unités expérimentales peut **limiter la capacité d'une étude à répondre correctement à la question de recherche** et donc mener à un gaspillage des ressources.
- Choisir un nombre **trop élevé** d'unités expérimentales peut mener à l'**utilisation inutile de ressources limitées** (temps, argent, animaux...), ralentir le progrès scientifique et dans certains cas causer un fardeau inutile aux participants de recherche.
- Il importe donc de faire un choix prenant en compte à la fois ces **considérations scientifiques, éthiques et pratiques** et de **JUSTIFIER** ce choix.

Justification du choix du nombre d'unités expérimentales

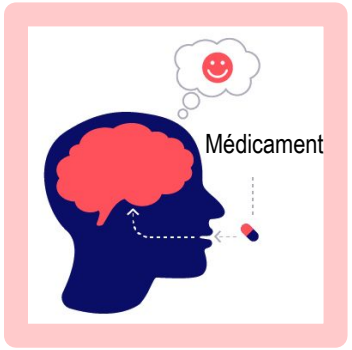
Selon Lakens (2022), cette justification peut prendre différentes formes:

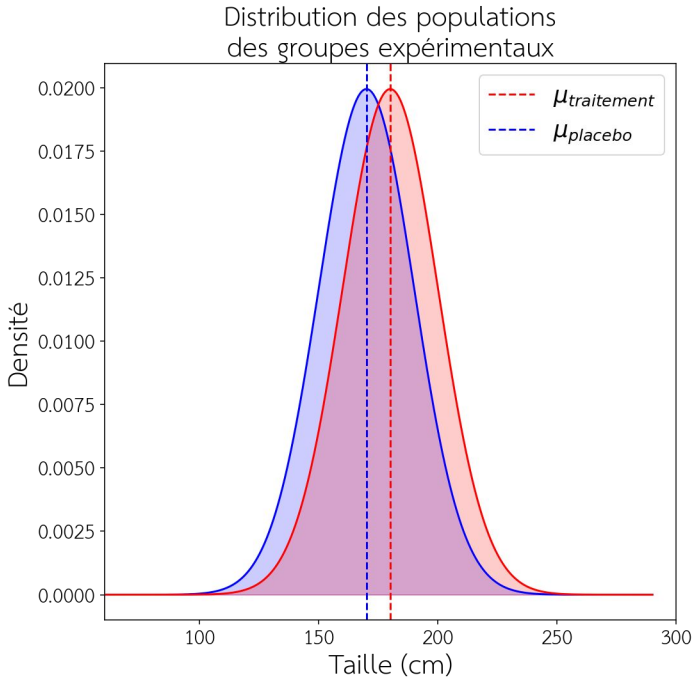
1. Collecter les données de (presque) toute la population.
    - Par exemple, si on étudie une petite population (e.g. employés d'une compagnie). Statistiquement idéal mais souvent impossible et peut entraîner l'investissement inutile de ressources.
2. Justifier la taille de l'échantillon sur la base d'une **analyse de puissance *a priori.***
3. Justifier la taille de l'échantillon sur la base de la précision désirée dans l'estimation de l'effet.
4. Choisir la taille de l'échantillon sur la base des ressources disponibles.
    - Les ressources sont toujours limitées. Toutefois, l'interprétation des résultats devrait tenir compte de la puissance statistique atteinte. Peut être justifié dans certains contextes.
5. Justifier sur la base de normes, de traditions ou de la littérature antérieure.
    - Souvent problématique, peut mener à perpétuer des pratiques inadéquates.
    - Peut parfois être utilisé si on vise à dépasser les standards en l'absence d'une meilleure justification.
6. Clairement indiquer que la taille d'échantillon a été choisie de façon arbitraire.
    - Problématique, mais il vaut mieux être transparent que de prétendre *a posteriori* que la taille d'échantillon avait été choisie *a priori.*

(Brève) révision de la statistique inférentielle dans l’approche fréquentiste

Mise en situation

- Vous désirez tester l'effet d'un nouveau médicament qui pourrait permettre d'augmenter la taille des personnes. Vous comparez la taille moyenne d'un groupe de personnes ayant reçu le médicament à un groupe de personnes ayant reçu un placebo.
- En réalité, ce médicament augmente la taille d'en moyenne 10 cm.
- Combien de personnes devriez-vous recruter dans votre essai clinique pour rejeter l'hypothèse nulle d'absence d'effet?

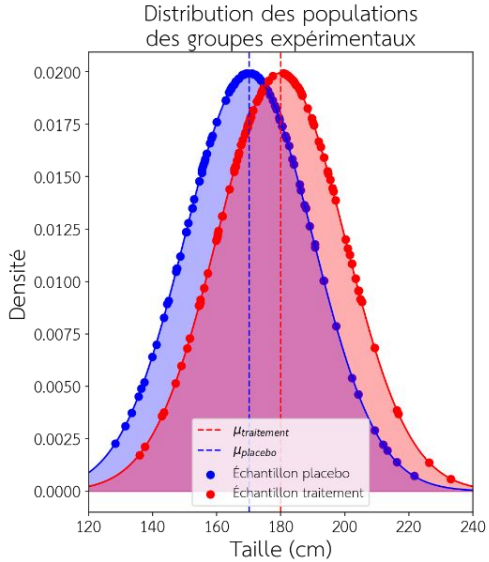


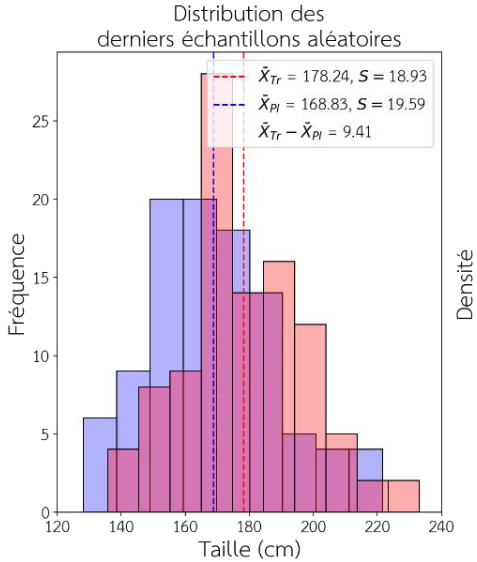


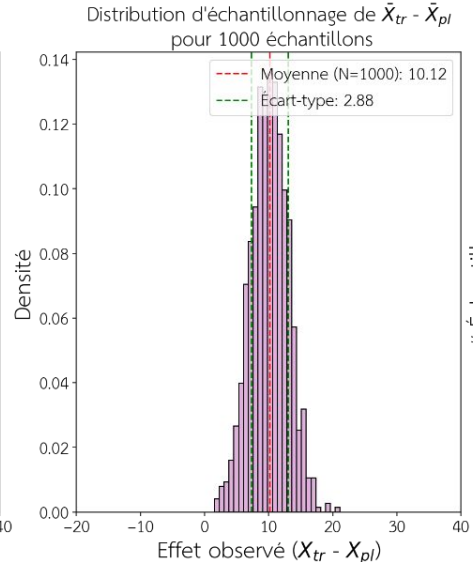
Approche fréquentiste et test d'hypothèse nulle

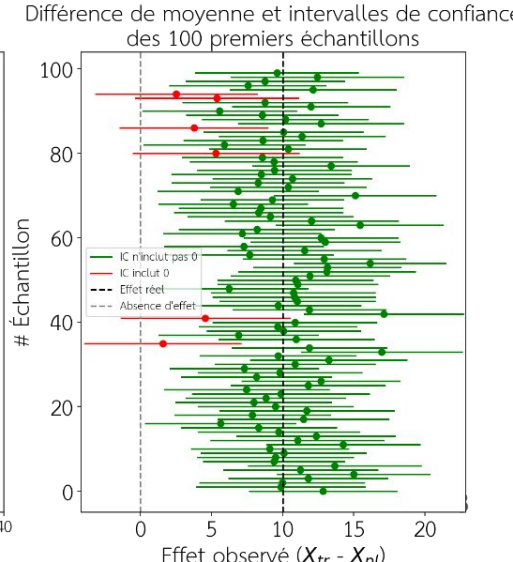
Échantillonnage aléatoire

- L'inférence statistique est le procédé par lequel l'analyse de données est utilisée pour inférer les propriétés d'une distribution de probabilités à partir de son échantillonnage.
- L'incertitude associée à l'échantillonnage peut être quantifiée à partir de l'intervalle de confiance.
- Si l'intervalle de confiance autour d'un effet inclut 0 (absence d'effet), nous ne pouvons rejeter l'hypothèse nulle d'absence d'effet.
- Si l'intervalle de confiance autour d'un effet exclut 0, nous pouvons rejeter l'hypothèse nulle d'absence d'effet.







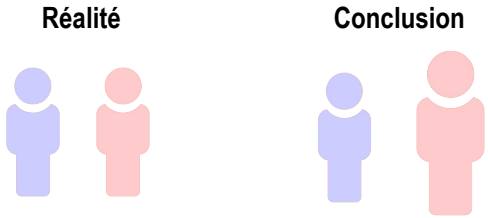


Approche fréquentiste et test d'hypothèse nulle

- Étant donné l'incertitude associée à l'échantillonnage, il est possible de tirer une conclusion erronée.

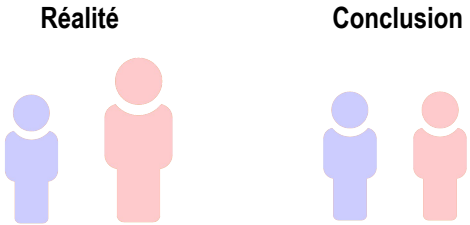
Erreur de type 1

- Rejet de l'hypothèse nulle alors qu'elle est vraie.
- Probabilité = seuil alpha (e.g. 0.05)



Erreur de type 2

- Non rejet de l'hypothèse nulle alors qu'elle est fausse
- Probabilité (beta) = 1 - puissance statistique

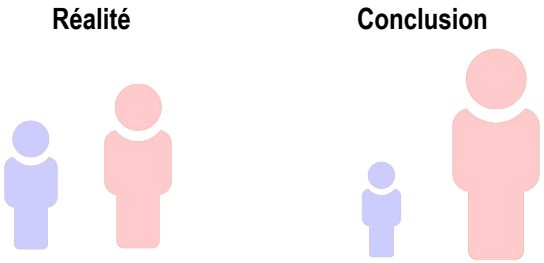


Approche fréquentiste et test d'hypothèse nulle

- Étant donné l'incertitude associée à l'échantillonnage, il est possible de tirer une conclusion erronée.

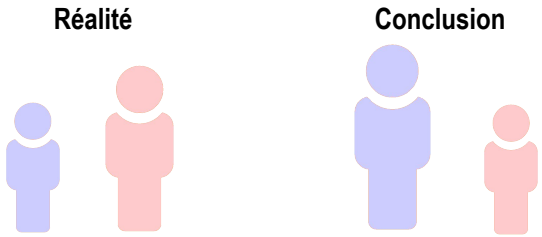
Erreur de type M

- Rejet correct de l'hypothèse nulle, mais exagération de la taille d'effet
- Probabilité liée à la puissance statistique



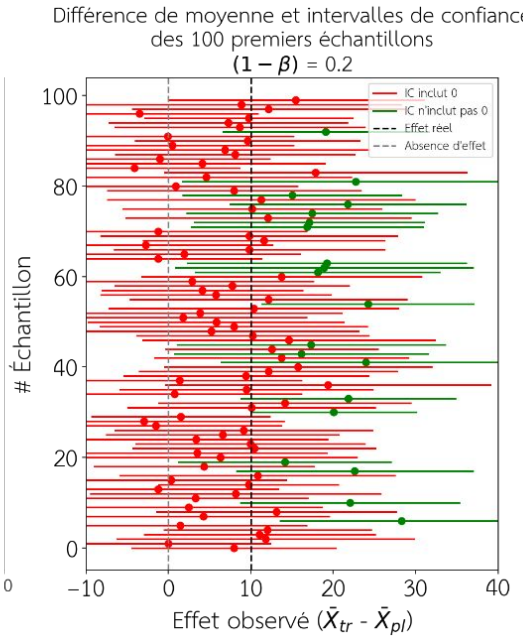
Erreur de type S

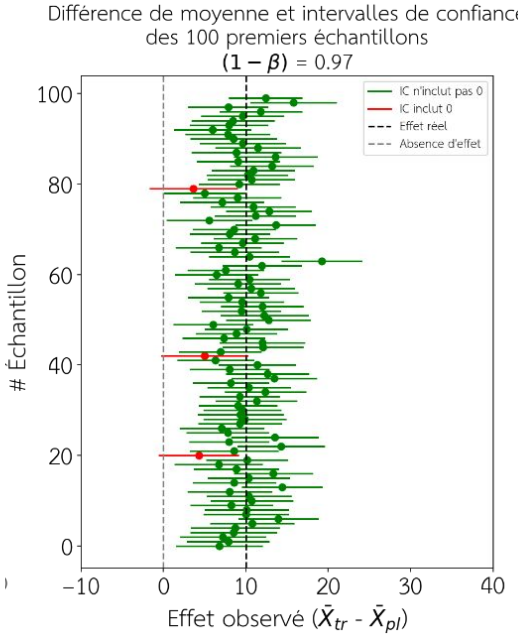
- Rejet correct de l'hypothèse nulle, mais mauvaise direction de l'effet
- Probabilité liée à la puissance statistique



Puissance statistique

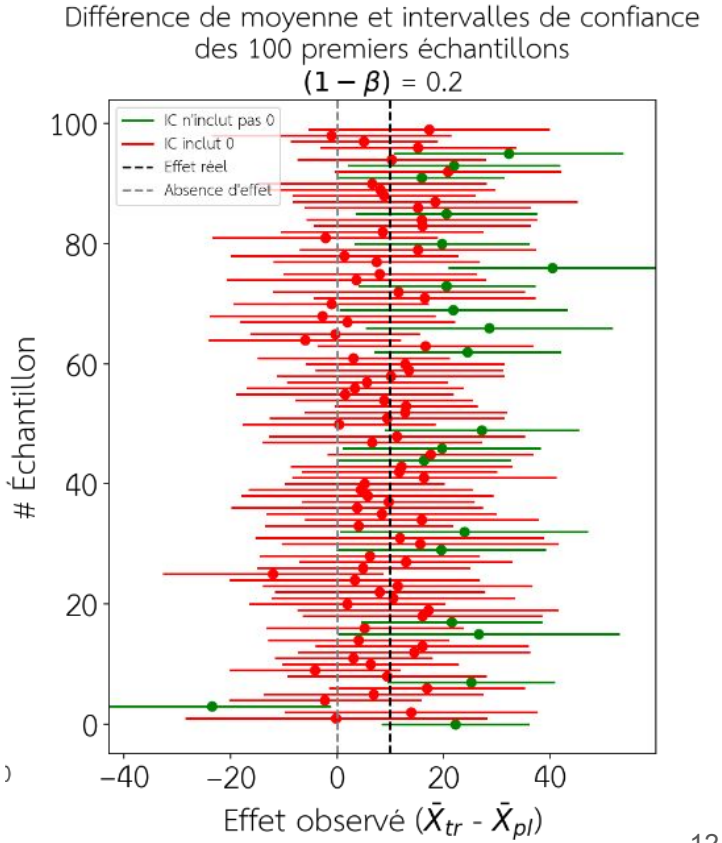
- La puissance statistique réfère à la probabilité pour notre étude de rejeter l'hypothèse nulle s'il existe un véritable effet de taille *d*.
- La puissance statistique est donc l'inverse de la probabilité de faire une erreur de type 2 (1-β).





Puissance statistique

- La puissance statistique **n'influence pas** la probabilité de faire une erreur de type 1.
- Toutefois, un devis avec une faible puissance a une plus **forte probabilité de mener à une erreur de type M ou de type S**, soit de faire une erreur dans la magnitude ou le signe de l'effet (Button, 2013).
- **<u>Une expérience qui manque de puissance est donc peu informative:</u>**
  - Si le résultat est non significatif, on ne peut exclure que l'effet réel est plus petit que ce qui pouvait être détecté (erreur de type 2).
  - Si le résultat est significatif, l'estimation de la taille de l'effet est imprécise et probablement exagérée.



Le manque de puissance statistique est un problème persistant dans la littérature scientifique et une des menaces principales à la crédibilité et robustesse des résultats scientifiques.

Psychological Methods
2004, Vol. 9, No. 2, 147–163

1082-989X/04/$12.00 DOI: 10.1037/1082-989X.9.2.147

The Persistence of Underpowered Studies in Psychological Research: Causes, Consequences, and Remedies

Scott E. Maxwell

Journal of Abnormal and Social Psychology
1962, Vol. 65, No. 3, 145–153

THE STATISTICAL POWER OF ABNORMAL–SOCIAL PSYCHOLOGICAL RESEARCH: A REVIEW [1]

JACOB COHEN
New York University

Power failure: why small sample size undermines the reliability of neuroscience

Katherine S. Button, John P. A. Ioannidis, Claire Mokrysz, Brian A. Nosek, Jonathan Flint, Emma S. J. Robinson & Marcus R. Munafò

Nature Reviews Neuroscience **14**, 365–376 (2013) | Cite this article

PLOS MEDICINE

OPEN ACCESS

ESSAY

Why Most Published Research Findings Are False

John P. A. Ioannidis

Published: August 30, 2005 • https://doi.org/10.1371/journal.pmed.0020124

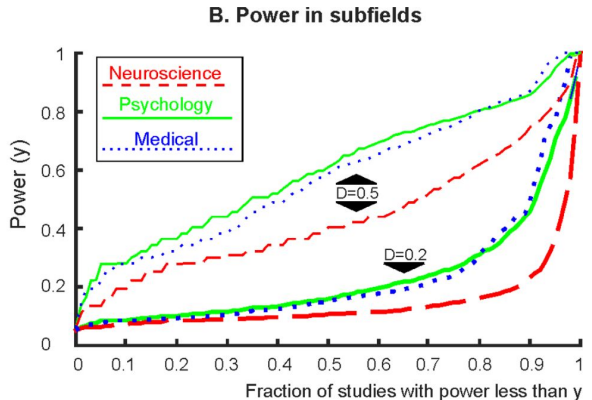
PLOS BIOLOGY

OPEN ACCESS PEER-REVIEWED

META-RESEARCH ARTICLE

Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature

Denes Szucs, John P. A. Ioannidis



L’analyse de puissance *a priori*

Analyse de puissance *a priori* - Paramètres

- La puissance statistique pour un test statistique spécifique est fonction de trois paramètres du devis de recherche:
  - La taille de l'effet à détecter
  - Le seuil alpha.
  - La taille de l'échantillon.

$$\text{Puissance statistique} \sim (\text{Seuil alpha, Taille d'effet, Taille d'échantillon})$$

- L'analyse de puissance *a priori* vise à choisir un niveau de puissance statistique, un seuil alpha et une taille d'effet pour ensuite calculer la taille d'échantillon qui permettra d'atteindre le niveau de puissance voulue.
- Il existe d'autres types d'analyse de puissance:
  - *Analyse de puissance de compromis:* Tailles d'effet et d'échantillon fixes, calcul d'un seuil alpha/beta pour atteindre la puissance voulue. Peut être utile si contrainte pratique pour taille d'échantillon.
  - *Analyse de puissance post-hoc:* Calculer la puissance atteinte en utilisant la taille d'effet d'une étude. Rarement pertinent (Zhang et al., 2019)
  - *…*

Puissance statistique ~ (Seuil alpha, Taille d'effet, Taille d'échantillon)

?

Analyse de puissance *a priori* - Choisir un niveau de puissance

- Traditionnellement, en psychologie et en recherche biomédicale on vise une puissance minimale de 80%.
- Par contre, de plus en plus de pression pour adopter des devis avec une puissance statistique > 90% (e.g. Chambers, 2013)
- Pour les études préenregistrées, la plupart des journaux nécessitent une puissance > 90%:
  - *Cortex*: > 90%
  - *Nature Human Behaviour / Nature communications*: > 95%
  - …
- Le choix du niveau de puissance précis dépend de plusieurs facteurs, dont les ressources disponibles et les conséquences d'une erreur de type 2.

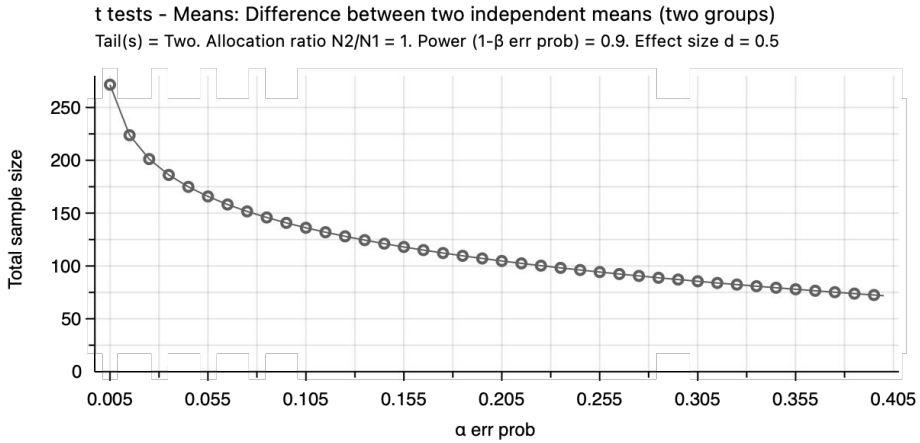
Puissance statistique ~ (Seuil alpha, Taille d'effet, Taille d'échantillon)

0.90

?

Analyse de puissance *a priori* - Choisir un seuil alpha

- Le seuil alpha détermine la taille de l'intervalle de confiance. Plus il est élevé, plus il sera facile de rejeter l'hypothèse nulle et plus le devis sera puissant (mais plus la probabilité de faire une erreur de type 1 augmente).
- Traditionnellement, le seuil alpha en recherche biomédicale est fixé à 0.05.
- Toutefois, rien n'empêche de choisir un seuil plus libéral ou conservateur en fonction des objectifs de recherche, des ressources disponibles et des conséquences de faire une erreur de type 1 ou 2 (voir Lakens et al. 2018).
- En fonction des hypothèses et de la question de recherche, considérer un test unilatéral peut aussi permettre d'augmenter la puissance statistique.
- Le choix du seuil alpha à utiliser doit aussi tenir compte des ajustements pour les comparaisons multiples s'il y a lieu.



Comment | Published: 26 February 2018

Justify your alpha

Daniel Lakens, Federico G. Adolfi, Casper J. Albers, Farid Anvari, Matthew A. J. Apps, Shlomo E. Argamon, Thom Baguley, Raymond B. Becker, Stephen D. Benning, Daniel E. Bradford, Erin M. Buchanan, Aaron R. Caldwell, Ben Van Calster, Rickard Carlsson, Sau-Chin Chen, Bryan Chung, Lincoln J. Colling, Gary S. Collins, Zander Crook, Emily S. Cross, Sameera Daniels, Henrik Danielsson, Lisa DeBruine, Daniel J. Dunleavy, ... Rolf A. Zwaan + Show authors

Nature Human Behaviour **2**, 168–171 (2018) | Cite this article

6947 Accesses | **209** Citations | **216** Altmetric | Metrics

Puissance statistique ~ (Seuil alpha, Taille d'effet, Taille d'échantillon)

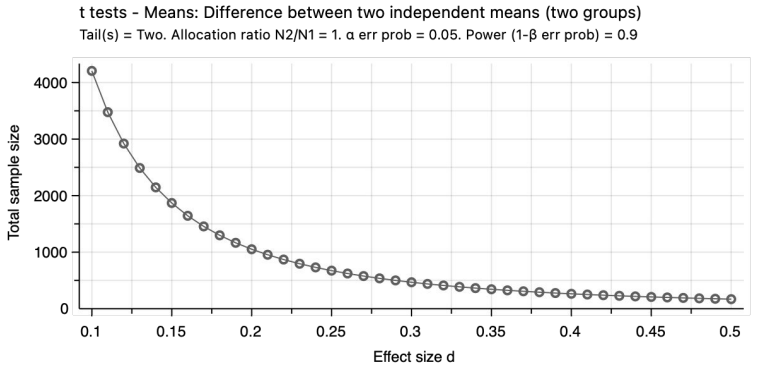
0.90

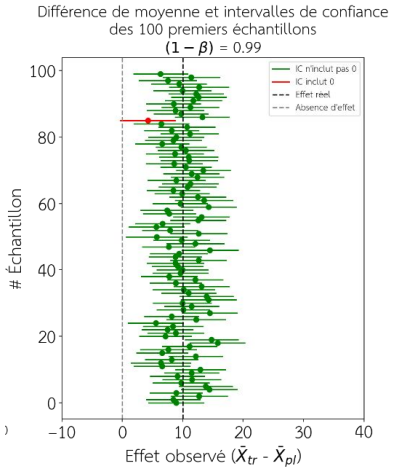
0.05, bilatéral

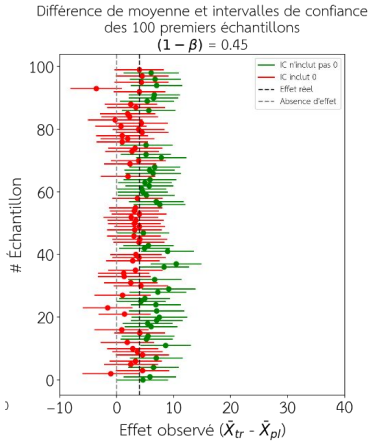
?

Analyse de puissance *a priori* - Choix de la taille d'effet

- Une plus grande taille d'effet permet de rejeter plus facilement l'hypothèse nulle et augmente donc la puissance statistique.
- L'équipe de recherche a un contrôle limité sur la taille d'effet d'une variable indépendante.
  - Elle peut toutefois être augmentée en utilisant un devis qui:
    - Optimise la mesure de la variable dépendante (réduit le bruit/écart-type)
    - Augmente la différence entre les conditions (plus sensible)
- **La taille d'effet** qui est l'objet de l'analyse de puissance doit être estimée par l'équipe de recherche et dépend du phénomène étudié.







Analyse de puissance *a priori* - Choix de la taille d'effet

- Le principal défi dans l'analyse de puissance *a priori* est le choix d'une taille d'effet plausible sur laquelle calibrer l'analyse. Plusieurs stratégies sont possibles:
  a. Se baser sur la taille d'effet observée dans une expérience pilote.
     - Peut être problématique, étant donné que les études pilotes sont généralement faites avec de petits échantillons. L'estimation de la taille d'effet sera peu précise.
  b. Se baser sur une méta-analyse ou sur les études antérieures.
     - Peut être problématique, étant donné le biais de publication bien connu (études non significatives, donc avec de plus petits effets, sont moins fréquemment publiées).
     - Aussi problématique si études antérieures sont petites puisque l'estimation de la taille d'effet sera peu précise.
     - Taille d'effets des études publiées peut aussi être influencée par les pratiques de recherches douteuses (QRP, Munafo et al., 2017)
  c. Utiliser la plus petite taille d'effet qui est considérée comme pertinente sur le plan scientifique/pratique/clinique.
     - Stratégie généralement recommandée (Lakens, 2022)

Analyse de puissance *a priori* - Choix de la taille d'effet

- Avec une taille d'échantillon assez grande, nous pourrions détecter n'importe quelle taille d'effet.
- Toutefois, à partir d'un certain seuil, les tailles d'effets peuvent ne pas avoir d'intérêt pratique ou clinique ou être inférieures à l'erreur de mesure.
  - Un traitement qui diminue en moyenne les symptômes dépressifs par rapport au groupe contrôle de 0.01 sur une échelle de 100.
  - Un nouveau type d'entraînement à la course qui diminue en moyenne de 10s le temps d'un marathon par rapport à l'entraînement standard.
- **Il revient donc à l'équipe de recherche de choisir la taille d'effet minimale qui a un intérêt scientifique, clinique ou pratique et d'utiliser cette taille d'effet pour l'analyse de puissance *a priori.***

Analyse de puissance *a priori* - Choix de la taille d'effet

- Comparer la taille d'effet attendue à celles d'autres phénomènes peut permettre d'évaluer sa plausibilité.

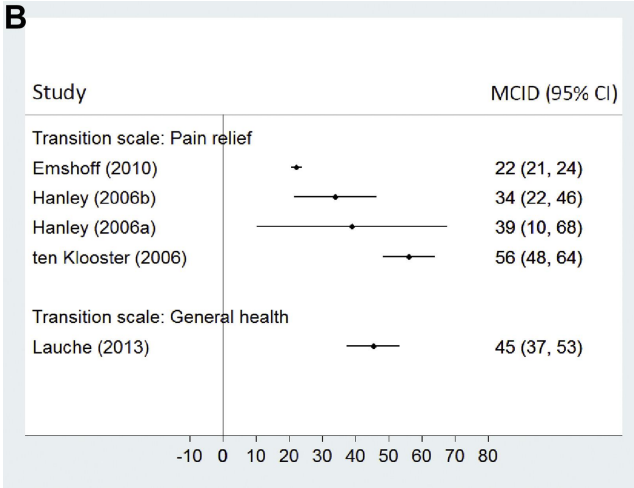
Description typique	$\|d_s\|$	Exemple
Très petit	~0.05	Effet du moment de la journée sur les résultats à un examen (0.03)
Petit	0.2	Différence de taille entre filles de 15 et 16 ans. Différence entre antidépresseurs et placebo pour dépression.
Moyen	0.5	Différences entre les habiletés visuo-spatiales/verbales entre hommes et femmes.
Large	0.8	Différence de taille entre filles de 13 et 18 ans.
Énorme	~ 2	Différence taille entre hommes et femmes.
Gigantesque	~ 5	Effet du sexe biologique sur l'identité de genre (~12); Effet du sexe sur l'orientation sexuelle (~6-7)

https://www.psy.gla.ac.uk/~steve/best/effect.html

Analyse de puissance *a priori* - Choix de la taille d'effet

Différence minimale cliniquement importante

- Lorsqu'une étude vise à déterminer l'effet d'un traitement sur une variable clinique, il peut être pertinent d'utiliser une estimation de la **différence minimale cliniquement importante** comme taille d'effet dans l'analyse de puissance.
- **Différence minimale cliniquement importante:** *"Le plus petit changement de score dans le domaine d'intérêt que le patient perçoit comme bénéfique et qui induirait pour le patient, en l'absence d'effets secondaires et d'un coût excessif, une modification dans la gestion de sa maladie"* (Jayadevappa et al., 2017).



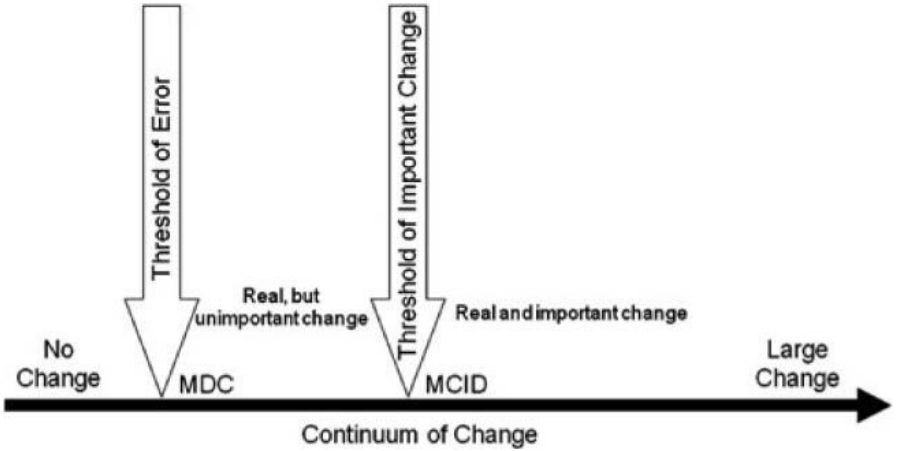
Relative MCID in chronic pain relief assessed as a mean change (five studies, 1,200 patients). MCID assessed as the mean change in pain score among patients with minimum improvement of pain, MCID, Minimum Clinically Important Difference (% reduction from baseline).

Olsen et al., 2018

Analyse de puissance *a priori* - Choix de la taille d'effet

Différence minimalement détectable

- Une variable dépendante est toujours mesurée avec une certaine erreur.
- Il est impossible de détecter une taille d'effet plus petite que l'erreur de mesure (i.e. fidélité test-retest, accord inter-juges, corrélation intra-classe).
- L'analyse de puissance ne devrait pas être calibrée pour détecter une taille d'effet plus petite que l'erreur de mesure de l'instrument utilisé.
- **Différence minimalement détectable:** *"Le plus petit changement qui dépasse l'erreur de mesure de l'instrument utilisé pour mesurer la variable dépendante."* (Kovacs et al., 2008)
- En l'absence de justification pratique/clinique, la taille d'effet minimalement détectable peut être une taille d'effet pertinente à utiliser pour l'analyse de puissance.



Beninato et al., 2011

Puissance statistique ~ (Seuil alpha, Taille d'effet, Taille d'échantillon)

0.90 | 0.05, bilatéral | d = 0.5 | ?

Analyse de puissance *a priori* - Calculs

- **Solution analytique:** Une fois les paramètres choisis, la taille d'échantillon peut être calculée pour la plupart des tests paramétriques de base en utilisant les distributions de probabilités théoriques. Plusieurs des logiciels statistiques permettent de faire ce genre d'analyse:
  - Gpower*
  - Python -> statsmodels.stats.power
  - R - > pwr
  - Matlab -> sampsizepwr
  - SPSS -> Analysis -> Power (à partir de la version 27)
  - JAMOVI -> Jpower
  - …
- **Solution par simulation:** Pour les devis plus complexes et ceux utilisant une approche non-paramétrique, il est généralement nécessaire d'effectuer des simulations de plusieurs expériences afin d'estimer la puissance statistique d'un devis (voir Lakens et al., 2021)
  - Certains outils de simulation spécifiques:
    - Plans factoriels complexes - https://shiny.ieis.tue.nl/anova_power/ - https://jakewestfall.shinyapps.io/pangea/
    - Modèles linéaires mixtes - https://debruine.github.io/lmem_sim/
    - IRMf - https://brainpower.readthedocs.io/en/latest/software_tools.html

Analyse de puissance *a priori* - Résumé

Puissance statistique ~ (Seuil alpha, Taille d'effet, Taille d'échantillon)

Probabilité pour notre étude de rejeter l'hypothèse nulle s'il existe un véritable effet de taille *d*.

Traditionnellement fixé à 0.8 mais souvent recommandé de viser > 0.9.

Augmenter le seuil alpha augmente la puissance.

Traditionnellement fixé à 0.05. Toutefois, rien n'empêche de choisir un seuil plus libéral ou conservateur en fonction des objectifs de recherche et des conséquences de faire une erreur de type 1 ou 2.

Augmenter la taille d'effet à détecter augmente la puissance.

La taille d'effet qui est l'objet de l'analyse de puissance doit être estimée par l'équipe de recherche et dépend du phénomène étudié.

Il est généralement recommandé d'utiliser la plus petite taille d'effet qui est considérée comme pertinente sur le plan scientifique/pratique/clinique.

Augmenter la taille d'échantillon augmente la puissance.

La taille de l'échantillon est généralement le paramètre sur lequel l'équipe de recherche a le plus de contrôle et qui est donc choisi pour obtenir la puissance statistique voulue.

Analyse de puissance *a priori* - Conclusion

- Le choix d'une taille d'échantillon en recherche quantitative doit être justifié en fonction de critères scientifiques, éthiques et pratiques.
- Un manque de puissance limite la capacité d'une expérience à répondre adéquatement à une question de recherche et est un problème persistant dans la littérature scientifique.
- L'analyse de puissance a priori est une excellente façon de justifier une taille d' échantillon afin d'atteindre une puissance adéquate.
- L'analyse de puissance repose sur le choix d'un niveau de puissance, d'un seuil alpha et d'une taille d'effet plausible pour calculer la taille d'échantillon nécessaire.

Ressources supplémentaires

Articles pertinents:

Lakens, D. (2022). Sample size justification. *Collabra: Psychology*, *8*(1), 33267.

Lakens, D., Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M. A., Argamon, S. E., ... & Zwaan, R. A. (2018). Justify your alpha. Nature human behaviour, 2(3), 168-171.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Frontiers in psychology*, *4*, 863.

Lakens, D., & Caldwell, A. R. (2021). Simulation-based power analysis for factorial analysis of variance designs. Advances in Methods and Practices in Psychological Science, 4(1), 2515245920951503.

Cours en ligne gratuits:

Improving your statistical questions (https://www.coursera.org/learn/improving-statistical-questions)

- Module 3: Designing Informative Studies

Improving your statistical inferences (https://www.coursera.org/learn/statistical-inferences)

- Module 3 - Multiple Comparisons, Statistical Power, Pre-Registration:

Références

Button, K. S., Ioannidis, J., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature reviews neuroscience*, *14*(5), 365-376.

Chambers, C. D. (2013). Registered reports: a new publishing initiative at Cortex. Cortex, 49(3), 609-610.

Jayadevappa, R., Cook, R., & Chhatre, S. (2017). Minimal important difference to infer changes in health-related quality of life—a systematic review. Journal of Clinical Epidemiology, 89, 188-198.

Kovacs, F. M., Abraira, V., Royuela, A., Corcoll, J., Alegre, L., Tomás, M., ... & Mufraggi, N. (2008). Minimum detectable and minimal clinically important changes for pain in patients with nonspecific neck pain. BMC musculoskeletal disorders, 9(1), 1-9.

Lakens, D. (2022). Sample size justification. *Collabra: Psychology*, *8*(1), 33267.

Lakens, D., Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M. A., Argamon, S. E., ... & Zwaan, R. A. (2018). Justify your alpha. Nature human behaviour, 2(3), 168-171.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Frontiers in psychology*, *4*, 863.

Lakens, D., & Caldwell, A. R. (2021). Simulation-based power analysis for factorial analysis of variance designs. Advances in Methods and Practices in Psychological Science, 4(1), 2515245920951503.

Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., Percie du Sert, N., ... & Ioannidis, J. (2017). A manifesto for reproducible science. Nature human behaviour, 1(1), 1-9.

Olsen, M. F., Bjerre, E., Hansen, M. D., Tendal, B., Hilden, J., & Hróbjartsson, A. (2018). Minimum clinically important differences in chronic pain vary considerably by baseline pain and methodological factors: systematic review of empirical studies. Journal of clinical epidemiology, 101, 87-106.

Zhang, Y., Hedo, R., Rivera, A., Rull, R., Richardson, S., & Tu, X. M. (2019). Post hoc power analysis: is it an informative and meaningful analysis?. General psychiatry, 32(4).