Statistical and sociological components of reproducibility

Jean-Baptiste Poline

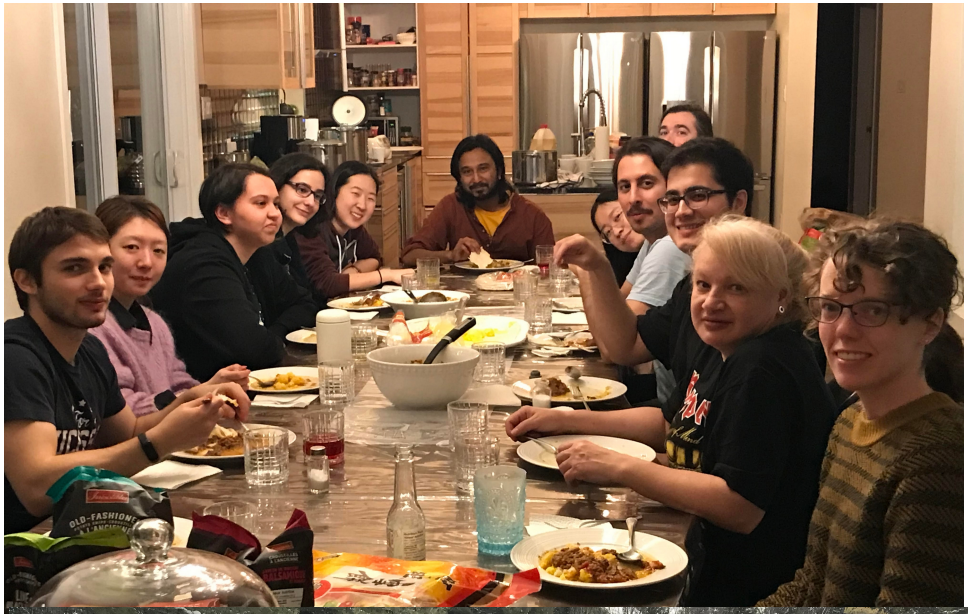
Neurohackademy 2024

ORIGAMI lab
MNI, Brain Imaging Centre,
McGill, Montreal
HWNI, UC Berkeley

neuro
Institut-Hôpital neurologique de Montréal
Montreal Neurological Institute-Hospital

The ORIGAMI Lab

ORIGAMI Lab



- A reproducibility story
- Some statistical aspects
- What should we solve – and how?

Forensic Analysis

- Potti et al., Nat. Med. 2006, 2008 vs Baggerly and Coombes, "Forensic analysis", Annals of applied Stat., 2009
- Choose cell lines that are most sensitive / resistant to a drug
- Use patients expression profiles to build a model that predicts patient response

Forensic Analysis

- Potti et al., Nat. Med. 2006, 2008 vs Baggerly and Coombes, "Forensic analysis", Annals of applied Stat., 2009
- Choose cell lines that are most sensitive / resistant to a drug
- Use patients expression profiles to build a model that predicts patient response

Baggerly and Coombes Forensic:

- Tried to replicate the results (data were available)
- Tried for several months and finally were able to obtain what was in the Potti article

Forensic Analysis

- Potti et al., Nat. Med. 2006, 2008 vs Baggerly and Coombes, "Forensic analysis", Annals of applied Stat., 2009
- Choose cell lines that are most sensitive / resistant to a drug
- Use patients expression profiles to build a model that predicts patient response

Baggerly and Coombes Forensic:

- Tried to replicate the results (data were available)
- Tried for several months and finally were able to obtain what was in the Potti article

- Finally – tried to publish the results in the journals where Potti published – at the time when a clinical trial was on its way

Forensic Analysis

Baggerly and Coombes Forensic:

"with poor documentation and irreproducibility even well meaning investigator may argue for drug that are contraindicated to some patients"

Forensic Analysis

Baggerly and Coombes Forensic:

"with poor documentation and irreproducibility even well meaning investigator may argue for drug that are contraindicated to some patients"

"**the most common errors are simple**
(e.g.,row or column offsets); conversely,
the most simple errors are common."

Amgen replication

Begley and Ellis, Nature, **2012**

- 53 papers examined at Amgen in preclinical cancer research
- Papers were selected that described something completely new and in very high impact factor journals
- **Scientific findings were confirmed in only 6 (11%)**

Errington et al Elife, **2021**
~30% completed with modifications

REPRODUCIBILITY IN CANCER BIOLOGY

Challenges for assessing replicability in preclinical cancer biology

Funding agencies reaction

NIH plans to enhance reproducibility

Francis S. Collins and **Lawrence A. Tabak** discuss initiatives that the US National Institutes of Health is exploring to restore the self-correcting nature of preclinical research.

Collins and Tabak. 2014. Nature 505: 612–13.

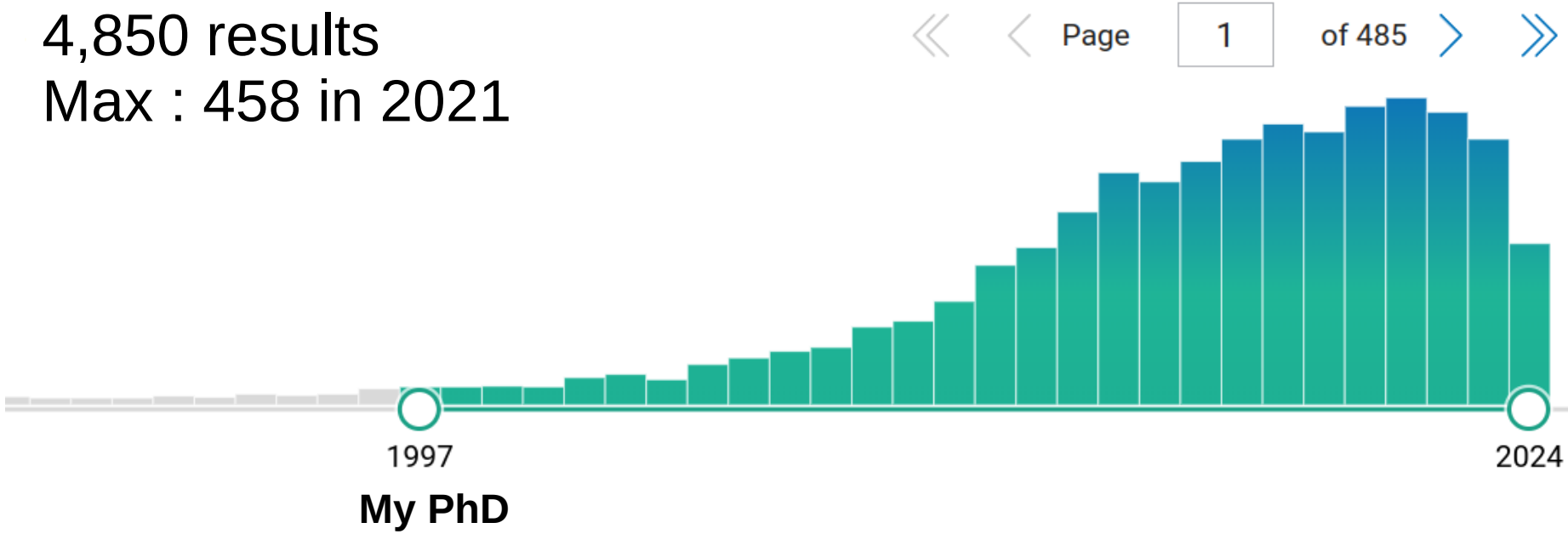

vevo
IT'S BEEN
10 YEARS

- A reproducibility story
- Some statistical aspects
- What should we solve – and how?

The end of classical statistics ?

(neuroimaging[Title/Abstract] OR (brain[Title/Abstract] AND imaging[Title/Abstract])) AND (statistics[Title/Abstract] OR (statistical[Title/Abstract] AND method[Title/Abstract]))

4,850 results
Max : 458 in 2021



My PhD

The Big Split

The Big Split

- How many good techniques … **are not used**
  - ISBI / MICCAI / NIPS/NeurIPS / ASA Statistical Methods in Imaging / SMI / COSYN
- How many are **not packaged** such that they are reusable by others ?
- Monti 2011:

HUMAN NEUROSCIENCE
published: 18 March 2011
doi: 10.3389/fnhum.2011.00028

Statistical analysis of fMRI time-series: a critical review of the GLM approach

*Martin M. Monti**

Deluge of papers do not help

Slowed canonical progress in large fields of science

Johan S. G. Chu[a,1] and James A. Evans[b,c,d] PNAS 2021

Examining 1.8 billion citations among 90 million papers across 241 subjects, we find a deluge of papers does not lead to turnover of central ideas in a field, but rather to ossification of canon.

THE LANCET

COMMENT | VOLUME 385, ISSUE 9976, P1380, APRIL 11, 2015

Offline: What is medicine's 5 sigma?

Richard Horton

Published: April 11, 2015 . DOI: https://doi.org/10.1016/S0140-6736(15)60696-1

Who said that ...

- "*We are inclined to think that as far as a particular hypothesis is concerned, no test based upon a theory of probability can by itself provide any valuable evidence of the truth or falsehood of a hypothesis*"

Who said that ...

- "*We are inclined to think that as far as a particular hypothesis is concerned, no test based upon a theory of probability can by itself provide any valuable evidence of the truth or falsehood of a hypothesis*"

Neyman J, Pearson E. On the problem of the most efficient tests of statistical hypotheses. *Philos Trans R Soc Lond A*. 1933;231: 289–337.

"In so far as a scientific statement speaks about reality, it must be falsifiable; and in so far as it is not falsifiable, it does not speak about reality."

Karl Popper, 1959

What are we testing?

- Most – or a very large part – of our hypotheses are "null" – those our tests reject

- Most of the interesting hypotheses are H1
  - H1: TBI is decreasing connectivity between brain regions
  - H0: TBI is not decreasing connectivity

- We are not designing tests to reject H1s
- Knowing that a parameter is not zero doesn't generally solidify a precise theory

TBI example

Hypothesis 1

A number of authors observed that : TBI physical disruption of pathways due to focal and diffuse injury results in regional expansion (increase) in strength or number of functional connections.

TBI example

Hypothesis 1

A number of authors observed that : TBI physical disruption of pathways due to focal and diffuse injury results in regional expansion (increase) in strength or number of functional connections.

Hypothesis 2

Other studies of moderate to severe brain injury found that white matter disruption during injury resulted in structural and functional disconnection of networks.

TBI example

Hypothesis 1

A number of authors observed that : TBI physical disruption of pathways due to focal and diffuse injury results in regional expansion (increase) in strength or number of functional connections.

Hypothesis 2

Other studies of moderate to severe brain injury found that white matter disruption during injury resulted in structural and functional disconnection of networks.

Even with two apparently contradictory hypotheses in place, there has been no direct testing of these positions against one-another to determine the scenarios where either may have merit. Instead, each of these hypotheses remained unconditionally intact and served to support distinct sets of outcomes.

Hypothesis quality

Hypothesis Quality	Example
Exploratory	"We examine the neural correlates of cognitive deficit after brain injury implementing graph theoretical measures of whole brain neural networks"
Testable Association	"We hypothesize that graph theoretical measures of whole brain neural networks predict cognitive deficit after brain injury"
Testable/Falsifiable Position *(offers possible mechanism and direction/magnitude of expected finding)*	"We hypothesize that memory deficits during the first 6-months post injury are due to white matter connection loss and maintain a linear and positive relationship with increased global network path length"
Testable/Falsifiable with Alternative Finding *(indicates how the hypothesis would and would not be supported)*	"We hypothesize that memory deficits during the first 6-months post injury are due to white matter connection loss and maintain a linear and positive relationship with increased global network path length. Diminished global path length in individuals with greatest memory impairment would challenge this hypothesis"

Ratmajer et al., eLife, 2023

Hypothesis quality

Hypothesis Quality	Example
Exploratory	"We examine the neural correlates of cognitive deficit after brain injury implementing graph theoretical measures of whole brain neural networks"
Testable Association	"We hypothesize that graph theoretical measures of whole brain neural networks predict cognitive deficit after brain injury"
Testable/Falsifiable Position *(offers possible mechanism and direction/magnitude of expected finding)*	"We hypothesize that memory deficits during the first 6-months post injury are due to white matter connection loss and maintain a linear and positive relationship with increased global network path length"
Testable/Falsifiable with Alternative Finding *(indicates how the hypothesis would and would not be supported)*	"We hypothesize that memory deficits during the first 6-months post injury are due to white matter connection loss and maintain a linear and positive relationship with increased global network path length. Diminished global path length in individuals with greatest memory impairment would challenge this hypothesis"

Ratmajer et al., eLife, 2023

Hypothesis quality

Hypothesis Quality	Example
Exploratory	"We examine the neural correlates of cognitive deficit after brain injury implementing graph theoretical measures of whole brain neural networks"
Testable Association	"We hypothesize that graph theoretical measures of whole brain neural networks predict cognitive deficit after brain injury"
Testable/Falsifiable Position *(offers possible mechanism and direction/magnitude of expected finding)*	"We hypothesize that memory deficits during the first 6-months post injury are due to white matter connection loss and maintain a linear and positive relationship with increased global network path length"
Testable/Falsifiable with Alternative Finding *(indicates how the hypothesis would and would not be supported)*	"We hypothesize that memory deficits during the first 6-months post injury are due to white matter connection loss and maintain a linear and positive relationship with increased global network path length. Diminished global path length in individuals with greatest memory impairment would challenge this hypothesis"

Ratmajer et al., eLife, 2023

Hypothesis quality

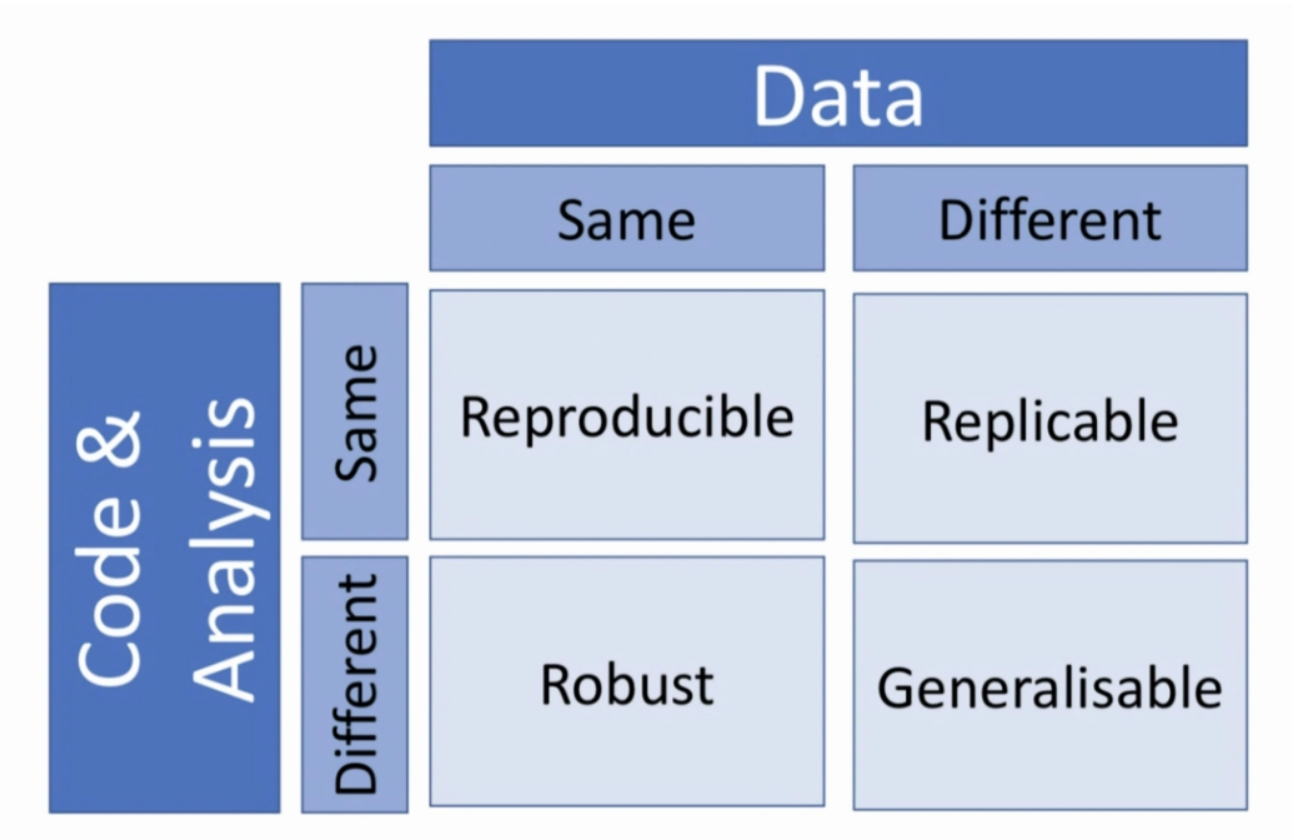
Hypothesis Quality	Example
Exploratory	"We examine the neural correlates of cognitive deficit after brain injury implementing graph theoretical measures of whole brain neural networks"
Testable Association	"We hypothesize that graph theoretical measures of whole brain neural networks predict cognitive deficit after brain injury"
Testable/Falsifiable Position *(offers possible mechanism and direction/magnitude of expected finding)*	"We hypothesize that memory deficits during the first 6-months post injury are due to white matter connection loss and maintain a linear and positive relationship with increased global network path length"
Testable/Falsifiable with Alternative Finding *(indicates how the hypothesis would and would not be supported)*	"We hypothesize that memory deficits during the first 6-months post injury are due to white matter connection loss and maintain a linear and positive relationship with increased global network path length. Diminished global path length in individuals with greatest memory impairment would challenge this hypothesis"

Ratmajer et al., eLife, 2023

Hypothesis testing and reproducibility

1. **P-value and the null hypothesis statistical testing (NHST)**
2. P-hacking
3. File drawer
3. Winner's curse
4. Effect sizes
5. Power
6. PPV
7. Statistical generalizability

Some Definitions



- Reproducibility / Robustness: mostly a computer science issue / a scientific field methodology issue
- **Replicability** / **generalizability** : also a statistical issue !

Credit: adapted from the Turing way

P-value to address the solidity of a result ?

- P-values are still **very** dominant as a tool to state that a result should be trusted
- They are simple to compute and every statistical software will implement them
- Originally developed by Fisher (1920)
- The “null hypothesis statistical significance tests” (the NHST framework) was developed by Neyman and Pearson to “make decision”

Quick reminder:

	Null hypothesis is true	Null hypothesis is false
Null hypothesis is not rejected	True negative	Type II error (β) (false negative)
Null hypothesis is rejected	Type I error (α) (false positive)	True positive

Understanding statistics

Consider a typical medical research study, for example designed to test the efficacy of a drug, in which a null hypothesis H_0 ('no effect') is tested against an alternative hypothesis H_1 ('some effect'). Suppose that the study results pass a test of statistical significance (that is P-value <0.05) in favor of H_1. What has been shown?

1. H_0 is false.
2. H_1 is true.
3. H_0 is probably false.
4. H_1 is probably true.
5. Both (1) and (2).
6. Both (3) and (4).
7. None of the above.

Results

Table 1 Quiz answer profile

Answer	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Number	8	0	58	37	6	69	12
Percent	4.2	0	30.5	19.5	3.2	36.3	6.3

Westover, M.B., Westover, K., Bianchi, M., 2011. Significance testing as perverse probabilistic reasoning. BMC medicine 9, 20.

Results

Table 1 Quiz answer profile

Answer	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Number	8	0	58	37	6	69	12
Percent	4.2	0	30.5	19.5	3.2	36.3	6.3

Westover, M.B., Westover, K., Bianchi, M., 2011. Significance testing as perverse probabilistic reasoning. BMC medicine 9, 20.

Definition of a p-value

Probability of observing a statistic, equal or more "extreme" to the one seen in the data, when the null hypothesis is true

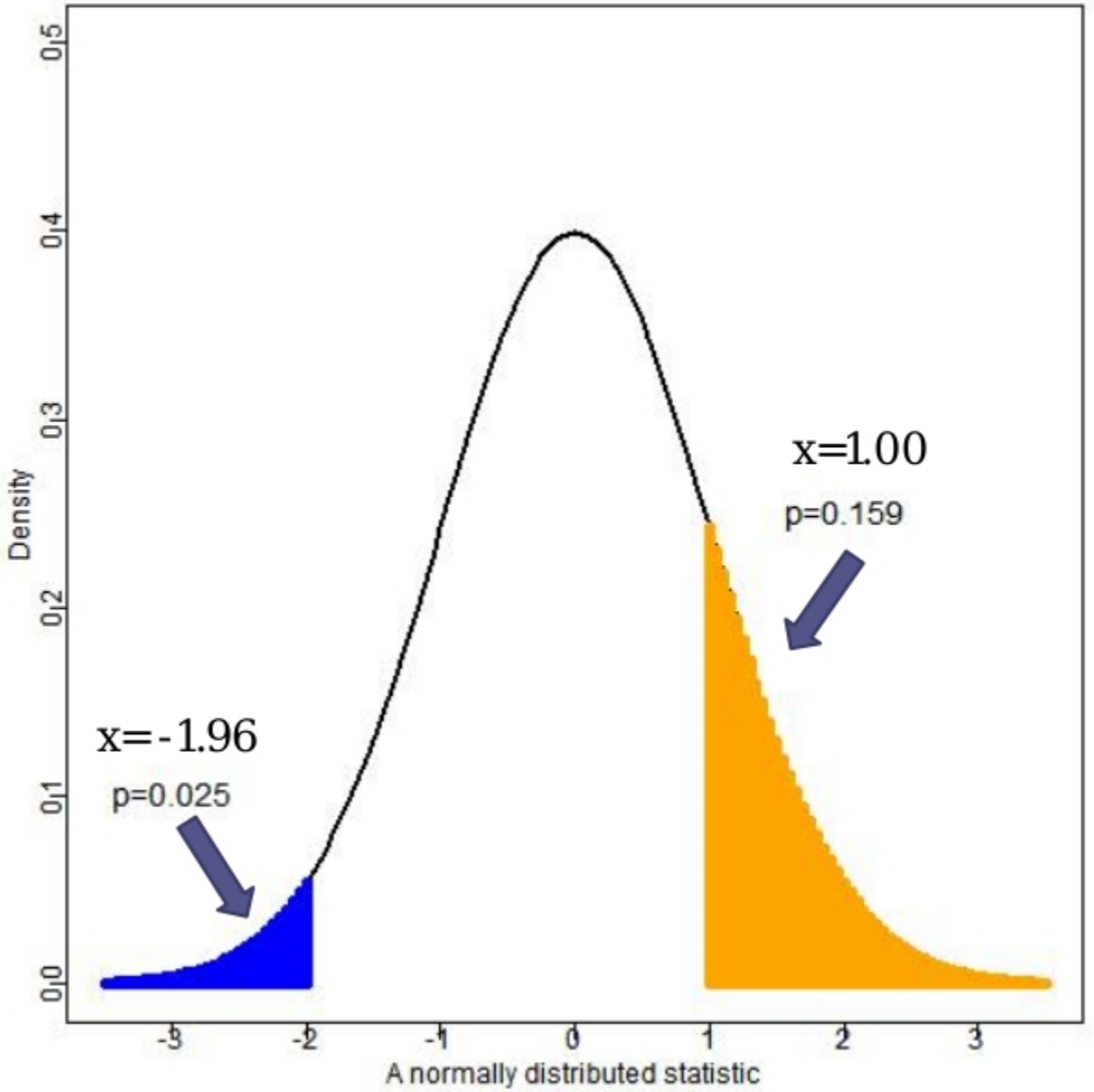
- What is a "statistic" ?
  - Any function of the data
    - the mean,
    - the SD,
    - the mean/SD,
    - the t statistics,
    - the z statistics,
    - etc

Under the null hypothesis

- Knowledge of the null hypothesis
- Choice of a statistic
- Concept of repeating the whole study in the same way
    - Same study design
    - Same sampling scheme
    - Same definition of the statistics
    - Same population sampled

Examples of null distributions

A normally distributed statistic

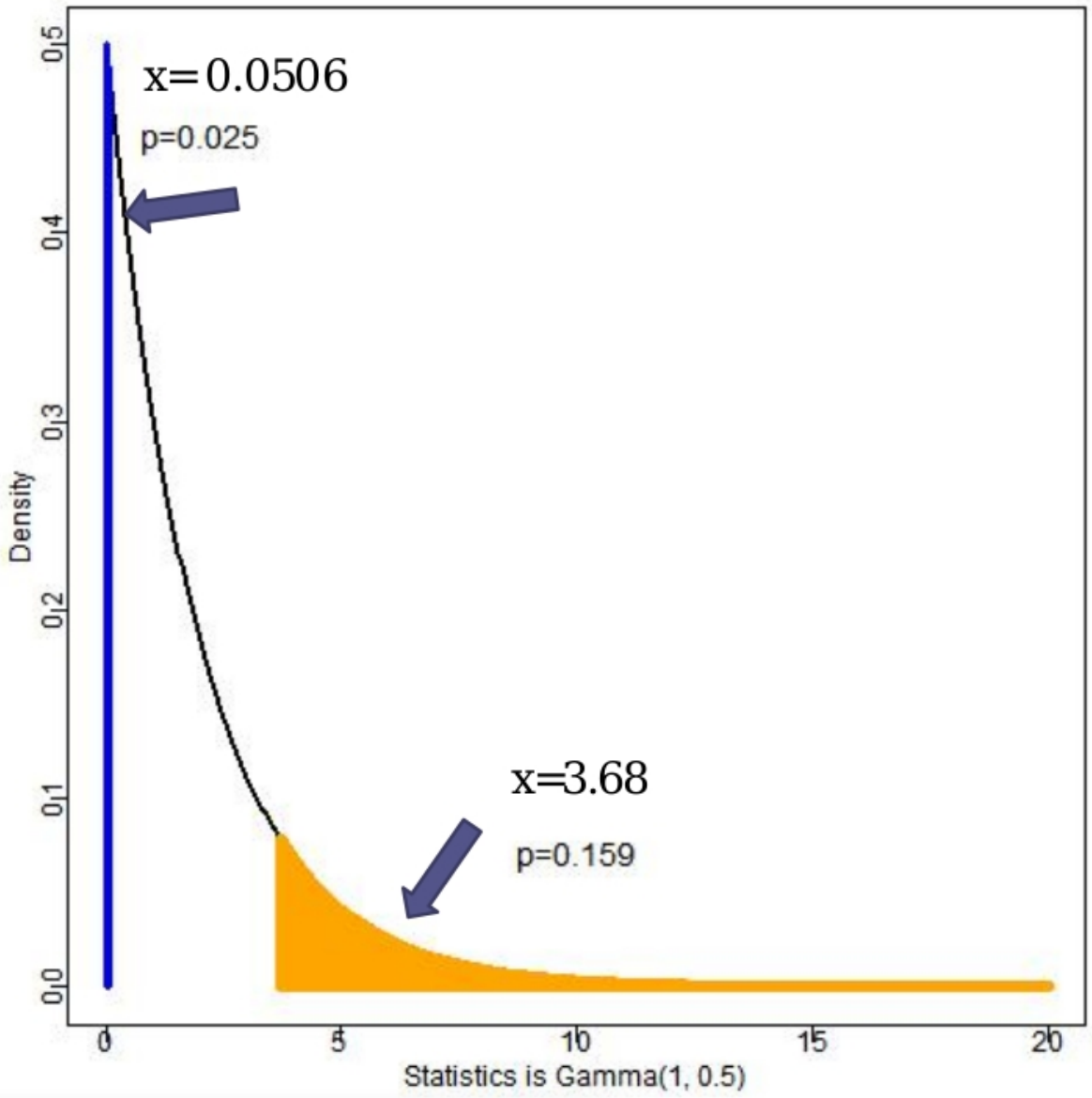


Credit: Celia Greenwood McGill

Gamma

A gamma-distributed statistic

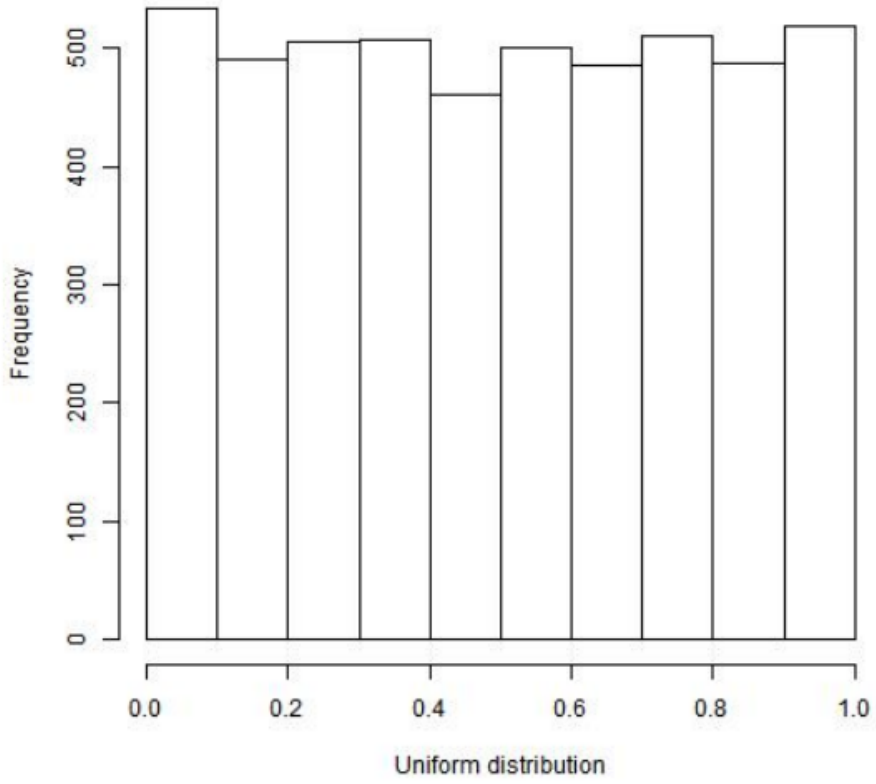
Gamma(1, 0.5)



Credit: Celia Greenwood McGill

Uniform distribution

- P-values have a uniform distribution when the null hypothesis is true



Why p-value are uniform under H0 ?

- We want the probability of rejecting the null to be alpha : P(*observed-stat* > *quantile-*α) = α
- P(*p-value is* < *5%*) = P(*observed-stat* > 95th percentile of the null) = 5%
- Hence: P(*p* <= *x*) = *x*
  - *Fact used in the "p-hacking test"*
- A p-value is a statistic : function of the data - therefore random

Credit: Jerôme Dockes

What are the common problems?

- "The most common and certainly most serious error made is to consider the p value as the probability that the null hypothesis is true."
  - It is the probability of observing these data, or more extreme data, if the null is true
- if trials are conducted with a controlled Type I error, say 5%, and adequate power, say 80%, then significant results almost always are corresponding to a true difference between the treatments compared.
  - How many true positive ?

Biau, D.J., Jolles, B.M., Porcher, R., 2010. P Value and the Theory of Hypothesis Testing: An Explanation for New Researchers. Clin Orthop Relat Res 468, 885–892. https://doi.org/10.1007/s11999-009-1164-4

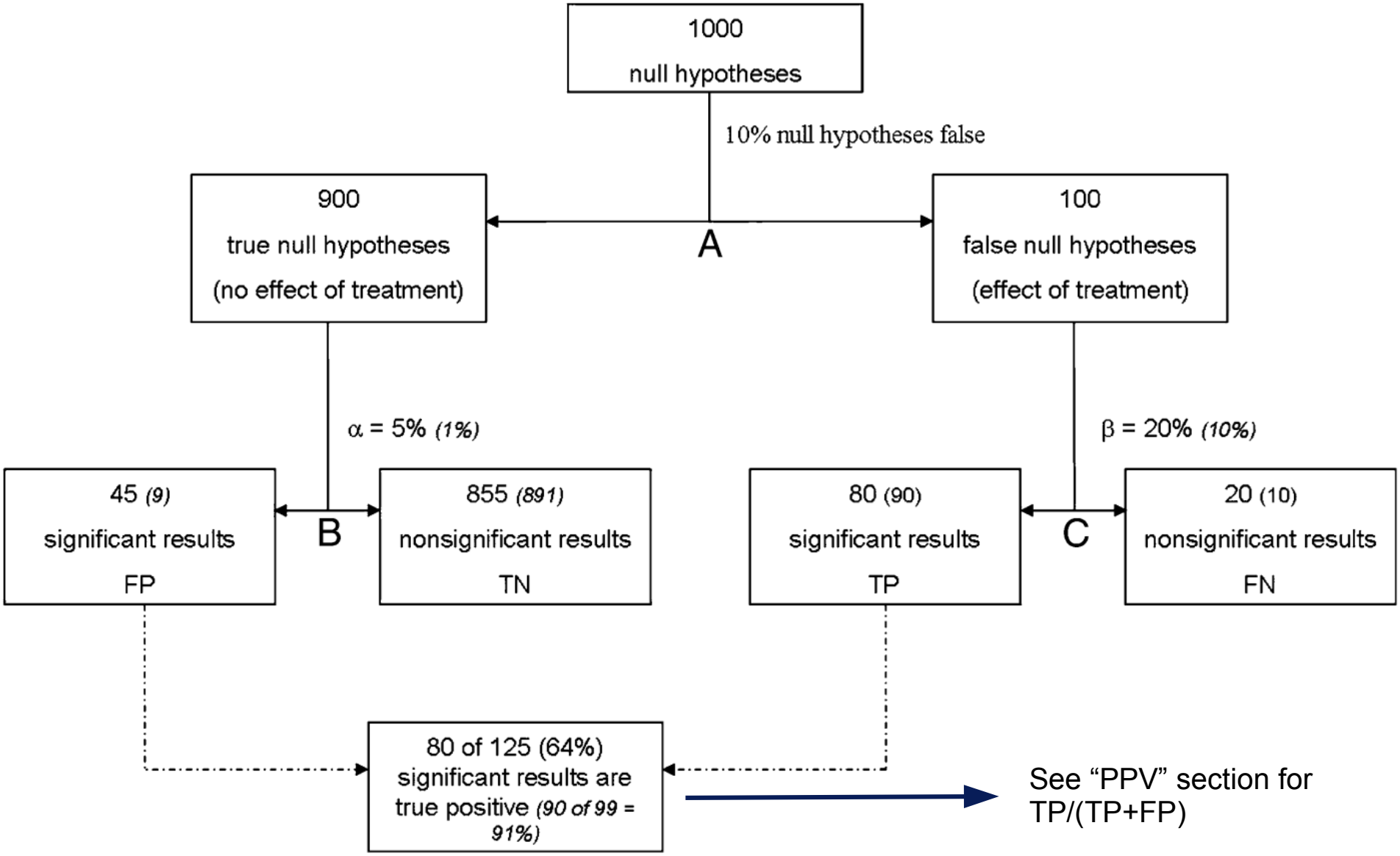
What are the common problems?

- "The most common and certainly most serious error made is to consider the p value as the probability that the null hypothesis is true."
    - It is the probability of observing these data, or more extreme data, if the null is true
- If trials are conducted with a controlled Type I error, say 5%, and adequate power, say 80%, then significant results almost always are corresponding to a true difference between the treatments compared.
    - Proportion of true positive / positive ?

A: >95% B: 85-95% C: 75-85% D: 65-75% E: <65%

Biau, D.J., Jolles, B.M., Porcher, R., 2010. P Value and the Theory of Hypothesis Testing: An Explanation for New Researchers. Clin Orthop Relat Res 468, 885–892. https://doi.org/10.1007/s11999-009-1164-4

What are the common problems?



Biau, D.J., Jolles, B.M., Porcher, R., 2010. P Value and the Theory of Hypothesis Testing: An Explanation for New Researchers. Clin Orthop Relat Res 468, 885–892.

P-values ...

- P-values can indicate how incompatible the data are with a specified statistical model.
- P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
- Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.

P-values ...

- P-values can indicate how incompatible the data are with a specified statistical model.
- P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
- Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.

P-values ...

- P-values can indicate how incompatible the data are with a specified statistical model.
- P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
- Scientific / clinical conclusions and policy decisions should not be based only on whether a p-value passes a specific threshold.

- Proper inference requires full reporting and transparency. P-values and related analyses should not be reported selectively.
- A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.
- By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

- Proper inference requires full reporting and transparency. P-values and related analyses should not be reported selectively.
- A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.
- By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

- Proper inference requires full reporting and transparency. P-values and related analyses should not be reported selectively.
- A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.
- By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

1. P-value and the null hypothesis statistical testing (NHST)
2. **P-hacking**
3. File drawer
3. Winner's curse
4. Effect sizes
5. Power
6. PPV
7. Statistical generalizability

P-Hacking

- Simmons and Simonsohn 2011
- Most often not intentional - and can be difficult to detect
  - As soon as some summary are seen ?
  - Necessity to “visualize data”
- P-hacking test
  - Based on a well known fact: p-values are uniformly distributed under the null
  - P-curves : Simonsohn et al, 2014
- Evil P-value

  http://www.repronim.org/module-stats/03-p-values/

  github.com/repronim/module-stats/notebooks/evil-p.ipynb
- P-hacking exercise

  github.com/repronim/module-stats/notebooks/P-value-exercise.ipynb

P-hacking - solutions ?

- **Pre-registration**
- Ban p-values
- Change \alpha
- Complement with other statistics !

Significance

The lack of reproducibility of scientific research undermines public confidence in science and leads to the misuse of resources when researchers attempt to replicate and extend fallacious research findings. Using recent developments in Bayesian hypothesis testing, a root cause of nonreproducibility is traced to the conduct of significance tests at inappropriately high levels of significance. Modifications of common standards of evidence are proposed to reduce the rate of nonreproducibility of scientific research by a factor of 5 or greater.

Johnson, V.E. (2013). Revised standards for statistical evidence. PNAS 110, 19313–19317.

1. P-value and the null hypothesis statistical testing (NHST)
2. P-hacking
3. **File drawer**
3. Winner's curse
4. Effect sizes
5. Power
6. PPV
7. Statistical generalizability

File drawer effect

The "File Drawer Problem" and Tolerance for Null Results

Robert Rosenthal
Harvard University

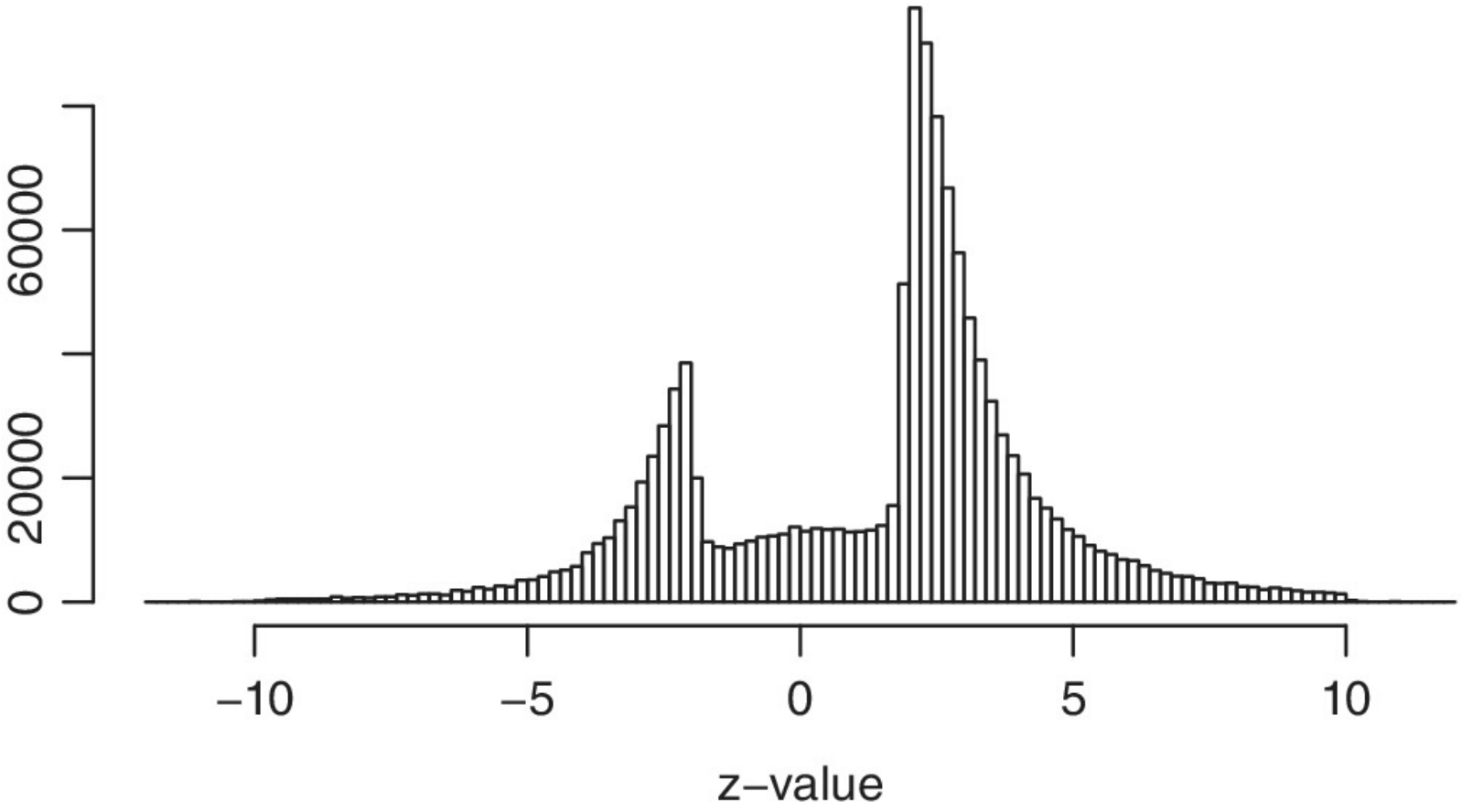
For any given research area, one cannot tell how many studies have been conducted but never reported. The extreme view of the "file drawer problem" is that journals are filled with the 5% of the studies that show Type I errors, while the file drawers are filled with the 95% of the studies that show non-significant results. Quantitative procedures for computing the tolerance for filed and future null results are reported and illustrated, and the implications are discussed.

Rosenthal, R., 1979. The file drawer problem and tolerance for null results. Psychological bulletin 86, 638.

File drawer effect

- Most journals will ask for a “new finding”
- The finding must survive some statistical threshold (ie, have some evidence of being likely “True”)
- P-values are used (and abused) to set this threshold
- A finding with P-value not surviving the 5% threshold will not be considered “statistically significant”
- The finding will not be published. The literature will only contain the “significant” results.

Z-values reported in Medline



The distribution of more than one million z-values from Medline (1976–2019)

Zwet, E.W. van, Cator, E.A., n.d. The significance filter, the winner's curse and the need to shrink. Statistica Neerlandica n/a. https://doi.org/10.1111/stan.12241

Remedies?

- **Pre-registration** !
- Convince journals that null results need to be published
- Can the importance of a null result be stated? ie. Can we assess power ?
    - A null result in a strong powered study is very valuable and interpretable
- Change statistical framework
    - Can the result be framed in a prediction setting ?

1. P-value and the null hypothesis statistical testing (NHST)
2. P-hacking
3. File drawer
3. **Winner's curse**
4. Effect sizes
5. Power
6. PPV
7. Statistical generalizability

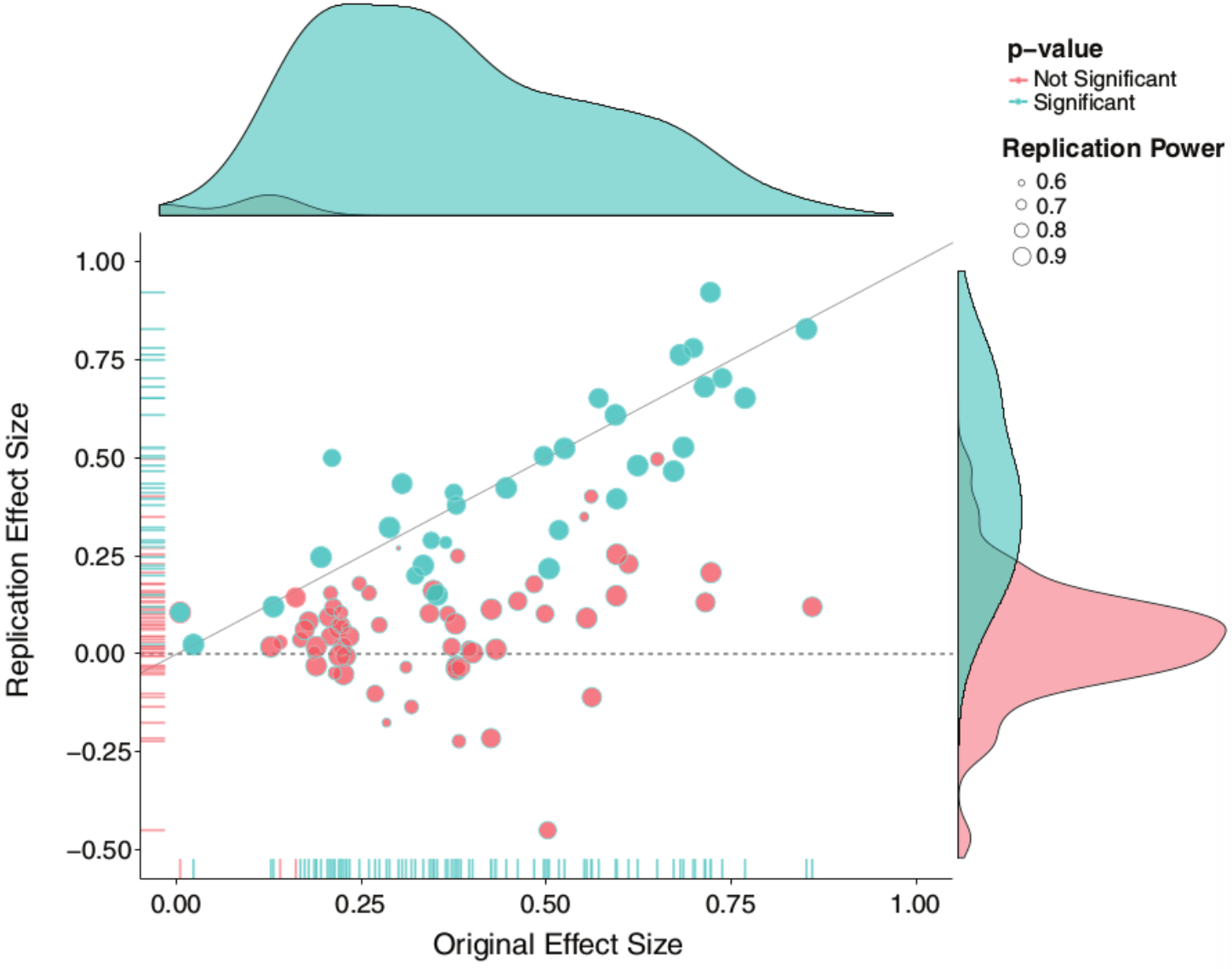
Winner's curse

- What is it ?
  - **Associations passing predetermined thresholds of statistical significance tend to overestimate the size of the effect**,
  - Especially when the sample size of the study is small and the threshold is stringent in multiple testing situations
- When does it occur ?
  - Predefined threshold
  - Small sample sizes
  - Stringent type I threshold (eg in multiple comparison)
- What's the impact on the literature?
  - Effect sizes reported are often going to be overestimated

Winner's curse

- What is it ?
  - **Associations passing predetermined thresholds of statistical significance tend to overestimate the size of the effect**,
  - Especially when the sample size of the study is small and the threshold is stringent in multiple testing situations
- When does it occur ?
  - Predefined threshold
  - Small sample sizes
  - Stringent type I threshold (eg in multiple comparison)
- What's the impact on the literature?
  - Effect sizes reported are often going to be overestimated

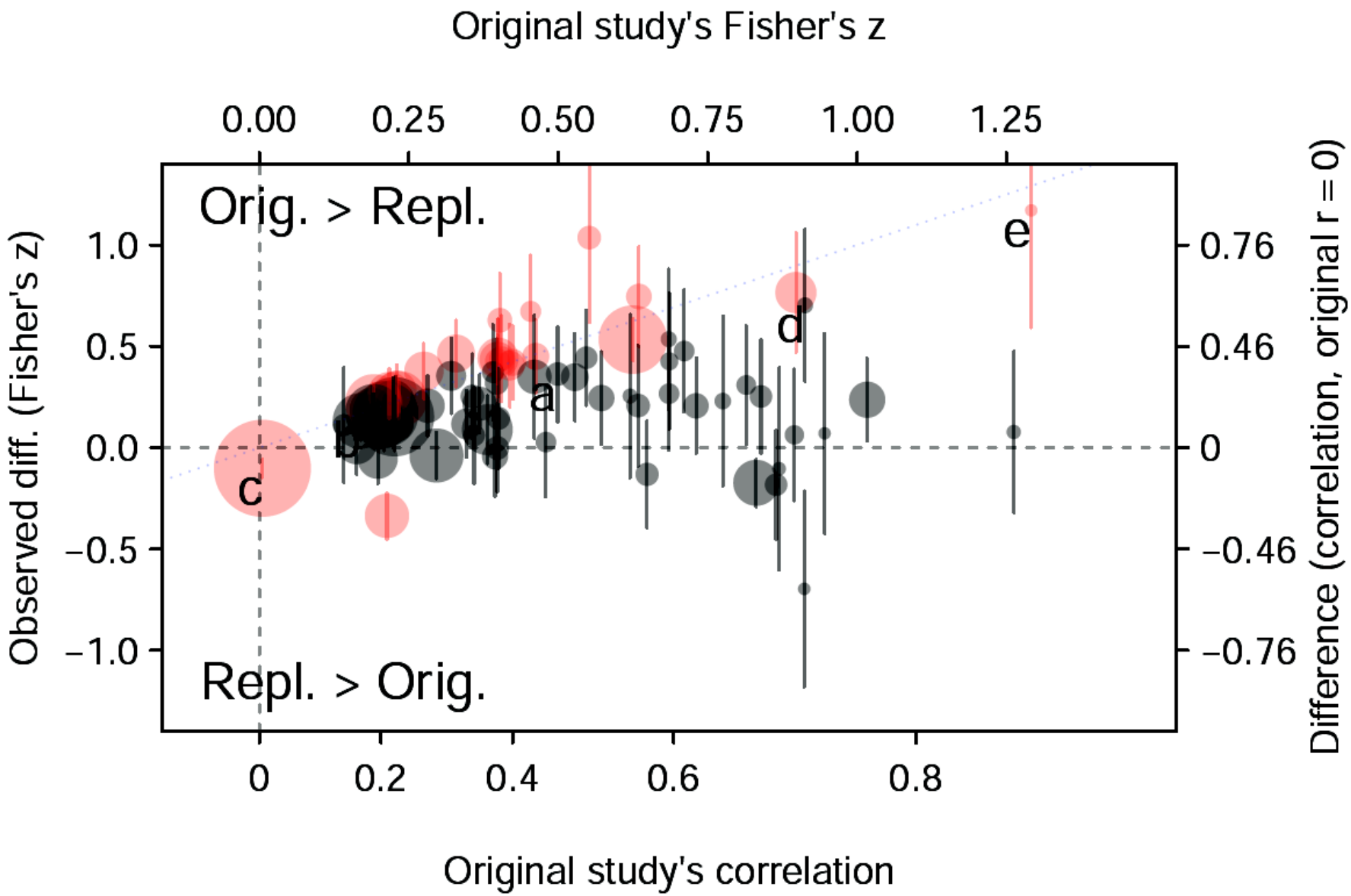
B. Nosek, Estimating the reproducibility of psychological science, Science 2015



* The mean **effect size** (r) of the replication effects (M r = 0.197, SD = 0.257) **was half the magnitude** of the mean effect size of the original effects (M r = 0.403, SD = 0.188)

* **39%** of effects were rated to have replicated the original effect

“Unfalsifiable”



Morey and Lakens, 2016

Effect size corrections

- The relation between power, sample size, and biais of effect size for a normally distributed effect size d

$$E[\bar{d}^*] = d + \frac{\sigma \cdot \phi(\Phi_\beta^{-1})}{\sqrt{n}(1-\beta)}$$

1. P-value and the null hypothesis statistical testing (NHST)
2. P-hacking
3. File drawer
3. Winner's curse
4. **Effect sizes**
5. Power
6. PPV
7. Statistical generalizability

Effect size - Example

What is the non standardized effect ?

Imagine 2 groups (1 and 2):

$$\mu = \bar{x_1} - \bar{x_2}$$

What is the standardized effect ? (eg Cohen's d)

$$d = \frac{\bar{x_1} - \bar{x_2}}{\sigma} = \frac{\mu}{\sigma}$$

"Z" : Effect accounting for the sample size

$$Z = \frac{\mu}{\sigma/\sqrt{n}}$$

Effect size ?

- A vague notion because any “measure of interest” can be an effect size
- Good quality for an “effect size” :
  - Simple
  - Interpretable - with units !
  - Standard in your field of science
- Examples:
  - Percentage of variance explained by a model
  - Correlation
  - Difference between means of two groups
    - Standardized / normalized ?
- What should be reported ?

Effect size ?

- A vague notion because any “measure of interest” can be an effect size
- Good quality for an “effect size” :
  - Simple
  - Interpretable - with units !
  - Standard in your field of science
- Examples:
  - Percentage of variance explained by a model
  - Correlation
  - Difference between means of two groups
    - Standardized / normalized ?
- What should be reported ?

Effect size ?

- A vague notion because any “measure of interest” can be an effect size
- Good quality for an “effect size” :
  - Simple
  - Interpretable - with units !
  - Standard in your field of science
- Examples:
  - Percentage of variance explained by a model
  - Correlation
  - Difference between means of two groups
    - Standardized / normalized ?
- What should be reported ?

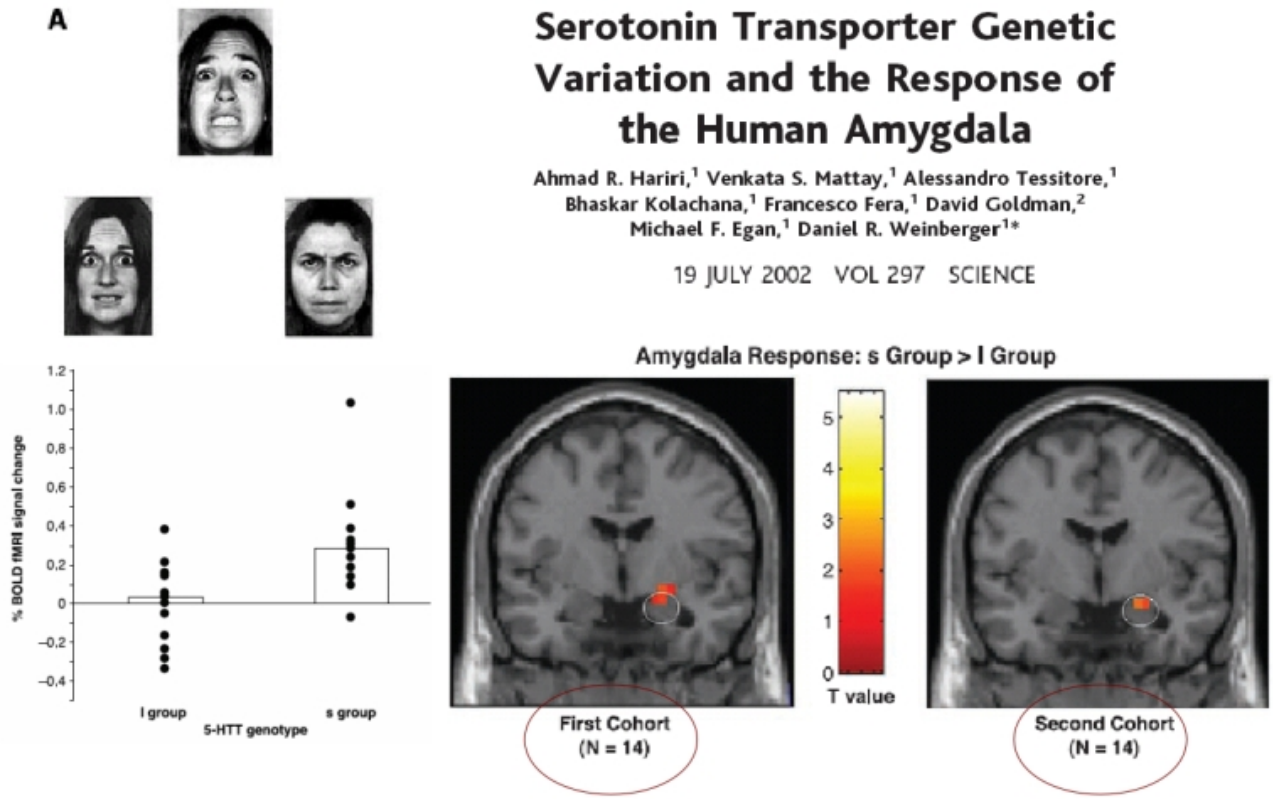
Guidelines for reporting

- It is hard to change reviewers, editors, and scientists habits - It is not the mission of publishing companies

- Almost every journals now require basic reporting:
  - Effect size, normalized and unormalized
  - Standard deviations and standard deviation of the means
  - Confidence / Credible intervals
  - (not generally in guidelines) : Some bootstraping of the data if possible to give an idea of the “results distribution”

- In neuroimaging : COBIDAS

An Example



- Authors report
  $m_1 = .28, m_2 = .03, \mathrm{SDM}_1 = 0.08, \mathrm{SDM}_2 = 0.05, N_1 = N_2 = 14$
- How do we compute the effect size ?

Effect size: in practice ?

First, compute the standard deviation of the data from the SDM

- get σ from SDM : $\sigma = \sqrt{14-1} \times \mathrm{SDM}$
- Combine the σ to have one estimation across the groups
  - formula easy to recompute or find
- $\sigma = \sqrt{14-1} \times \mathrm{SDM}$, $d = \frac{m_1 - m_2}{\sigma} = 1.05$
- What is the percentage of variance explained ?
- Write the estimated model: $Y = [1 \dots 1]^t [m_1 - m_2] + \mathrm{residual}$
- Compute the total sum of square $Y^t Y$, then the proportion:
- $V_e = \frac{(n_1+n_2)(m_1-m_2)^2}{n_1 s_1^2 + n_2 s_2^2 + (n_1+n_2)(m_1-m_2)^2} > 40\%$

Effect size: in practice ?

First, compute the standard deviation of the data from the SDM

- get σ from SDM : $\sigma = \sqrt{14-1} \times \mathrm{SDM}$
- Combine the σ to have one estimation across the groups
  - formula easy to recompute or find

m1=.28, m2=.03 $\quad d = \frac{m_1 - m_2}{\sigma} = 1.05$

- What is the percentage of variance explained ?

- Write the estimated model: $Y = [1 \ldots 1]^t [m_1 - m_2] + \mathrm{residual}$
- Compute the total sum of square $Y^t Y$, then the proportion:
- $V_e = \frac{(n_1+n_2)(m_1-m_2)^2}{n_1 s_1^2 + n_2 s_2^2 + (n_1+n_2)(m_1-m_2)^2} > 40\%$

Effect size: in practice ?

First, compute the standard deviation of the data from the SDM

- get σ from SDM : $\sigma = \sqrt{14-1} \times \mathrm{SDM}$
- Combine the σ to have one estimation across the groups
  - formula easy to recompute or find
- $\sigma = \sqrt{14-1} \times \mathrm{SDM}$, $d = \frac{m_1 - m_2}{\sigma} = 1.05$

- What is the percentage of variance explained ?

- Write the estimated model: $Y = [1 \dots 1]^t [m_1 - m_2] + \mathrm{residual}$
- Compute the total sum of square $Y^t Y$, then the proportion:
- $V_e = \frac{(n_1+n_2)(m_1-m_2)^2}{n_1 s_1^2 + n_2 s_2^2 + (n_1+n_2)(m_1-m_2)^2} > 40\%$

Effect size: in practice ?

First, compute the standard deviation of the data from the SDM

- get σ from SDM : $\sigma = \sqrt{14-1} \times \mathrm{SDM}$
- Combine the σ to have one estimation across the groups
  - formula easy to recompute or find
- $\sigma = \sqrt{14-1} \times \mathrm{SDM}$, $d = \frac{m_1 - m_2}{\sigma} = 1.05$
- What is the percentage of variance explained ?
- Write the estimated model: $Y = [1 \dots 1]^t [m_1 - m_2] + \mathrm{residual}$
- Compute the total sum of square $Y^t Y$, then the proportion:
- $V_e = \frac{(n_1+n_2)(m_1-m_2)^2}{n_1 s_1^2 + n_2 s_2^2 + (n_1+n_2)(m_1-m_2)^2} > 40\%$

Effect size: in practice ?

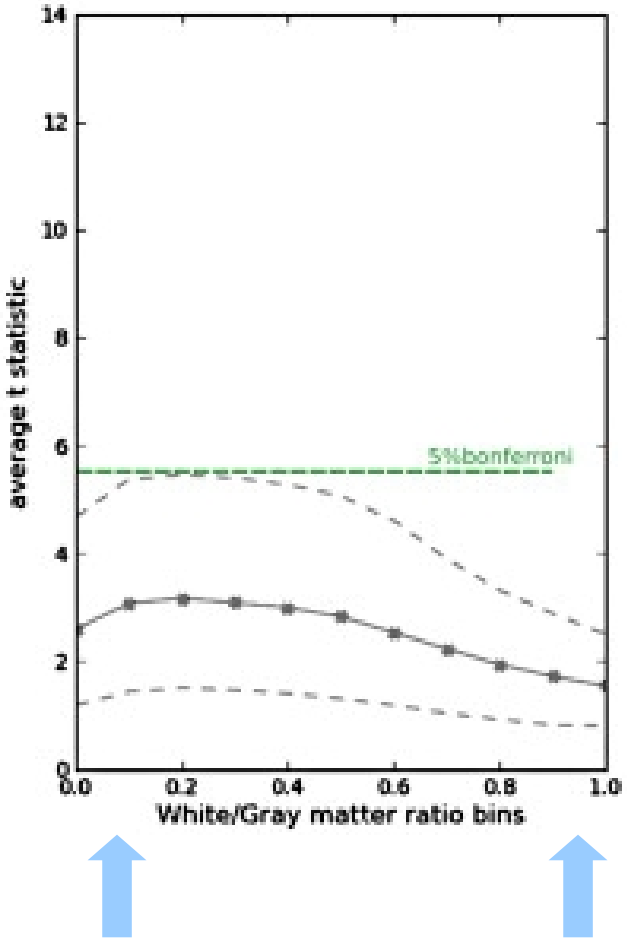
First, compute the standard deviation of the data from the SDM

- get σ from SDM : $\sigma = \sqrt{14-1} \times \mathrm{SDM}$
- Combine the σ to have one estimation across the groups
  - formula easy to recompute or find
- $\sigma = \sqrt{14-1} \times \mathrm{SDM}$, $d = \frac{m_1 - m_2}{\sigma} = 1.05$
- What is the percentage of variance explained ?
- Write the estimated model: $Y = [1 \ldots 1]^t [m_1 - m_2] + \mathrm{residual}$
- Compute the total sum of square $Y^t Y$, then the proportion:
- $$V_e = \frac{(n_1+n_2)(m_1-m_2)^2}{n_1 s_1^2 + n_2 s_2^2 + (n_1+n_2)(m_1-m_2)^2} > 40\%$$

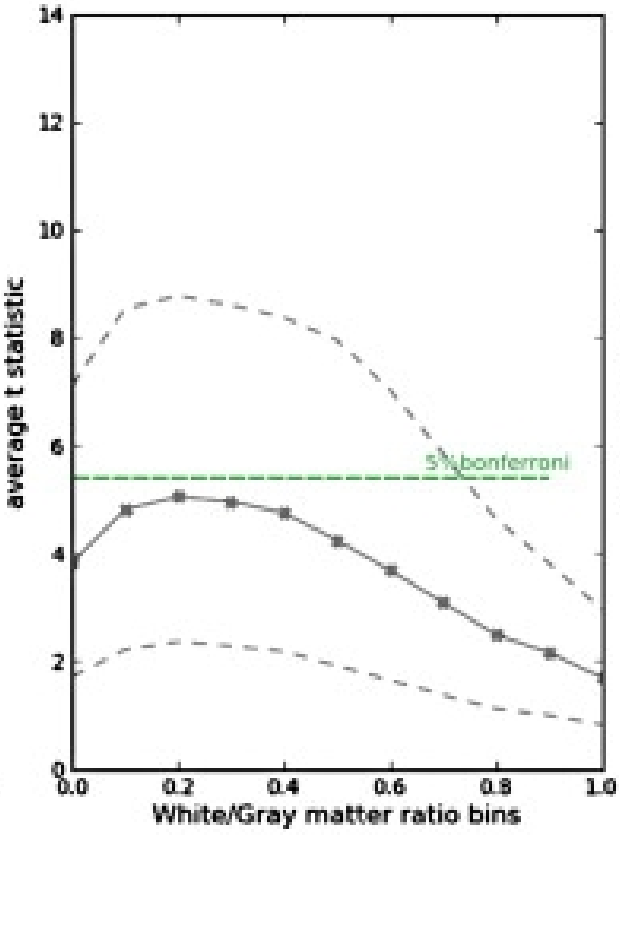
If N is very large ?

- The other issue: anything can be significant with large N
  - Eg, Thyreau et al., NeuroImage : most of the brain is significant with very large N
- “interesting” or “protected” inference
  - Choose a null that makes sense !
  - Find the minimal effect size for which a finding would be of biological interest
  - Apply statistical test with new null

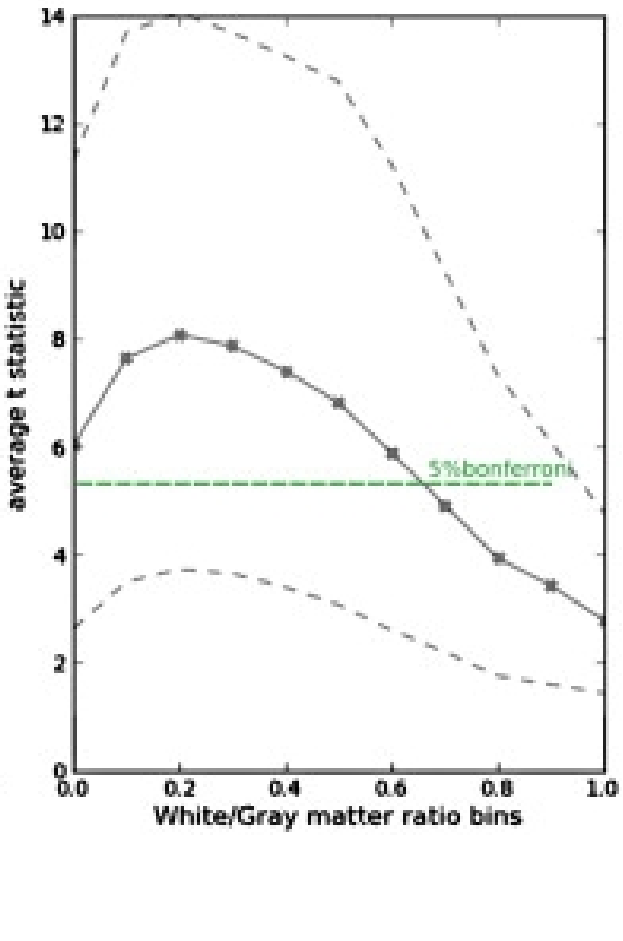
N = 200



N = 500



N = 1300



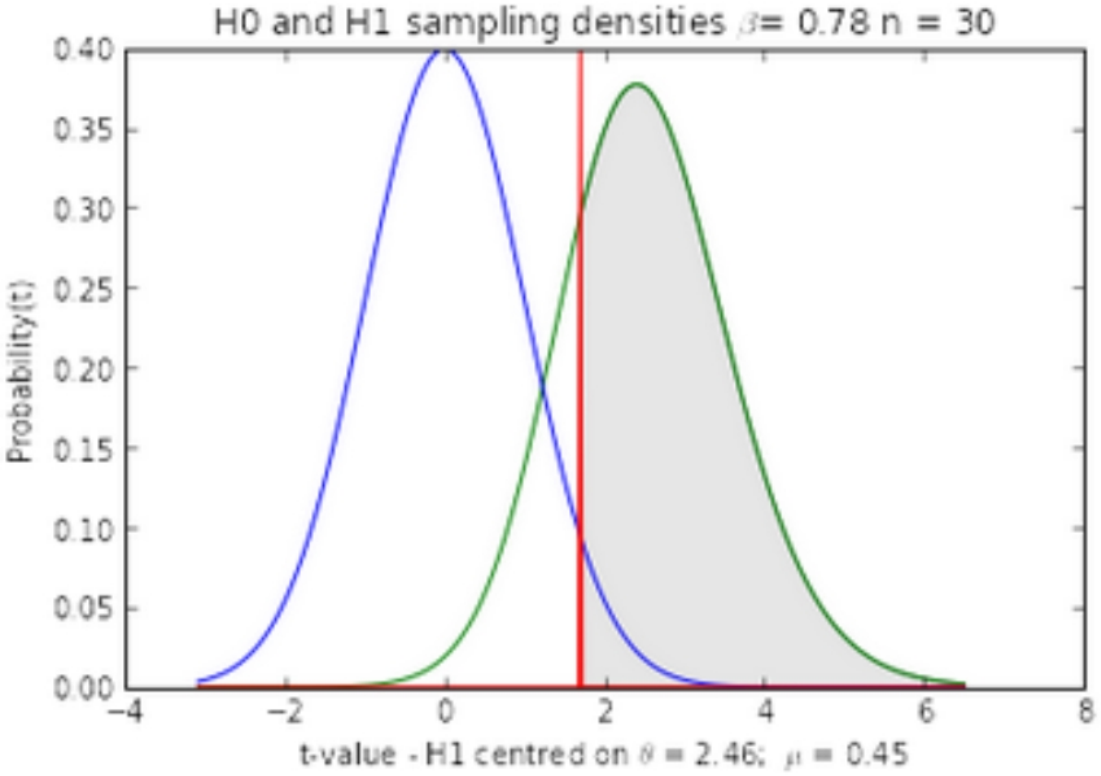
Grey matter

White matter

1. P-value and the null hypothesis statistical testing (NHST)
2. P-hacking
3. File drawer
3. Winner's curse
4. Effect sizes
5. **Power**
6. PPV
7. Statistical generalizability

Power - definition

Decision/H	H0 True	H1 True
reject	α (type I)	$1 - \beta$ (Power)
not reject	1 - α	β (type II)

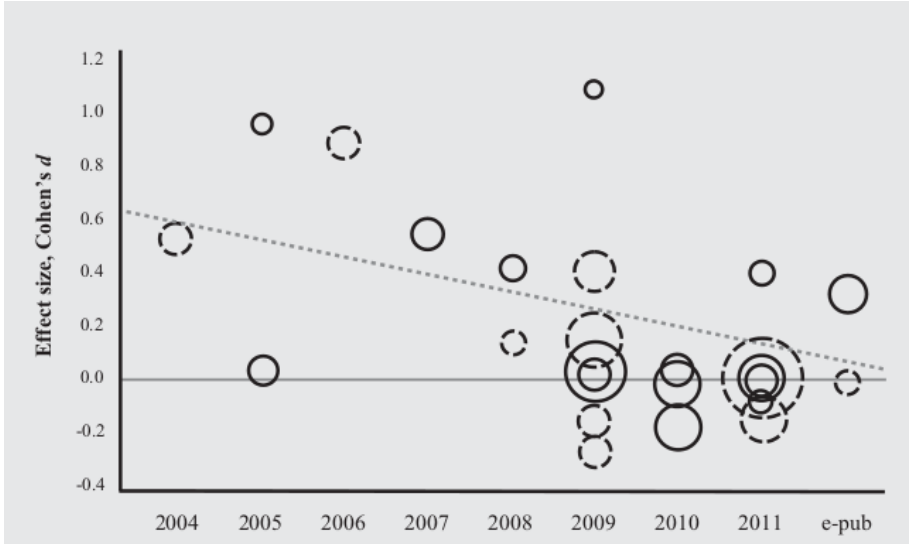


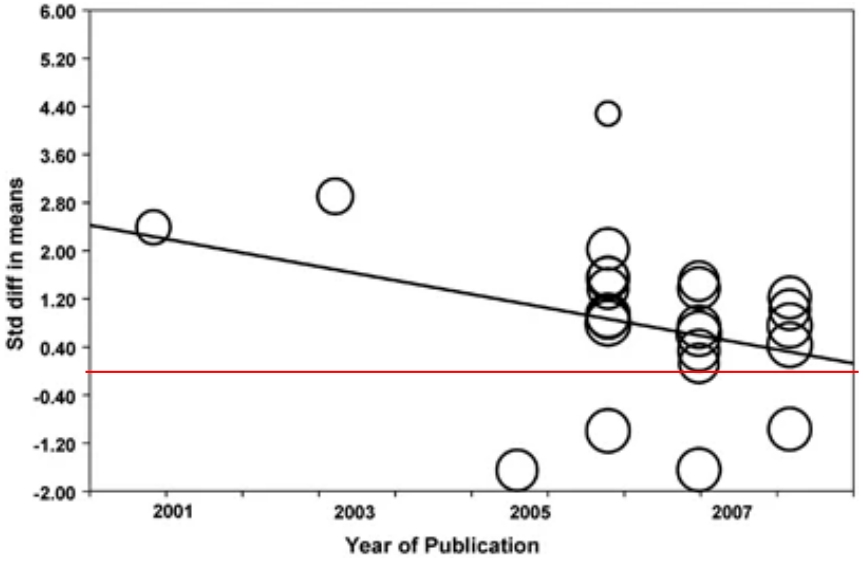
Power (W)

1. Power is a key measure because
   - Without good power, the study is not worth doing
   - Without good power, the study results are doubtful (see PPV)
2. Power is hard - or very hard to measure
   - You have to know H0, H1, effect size
3. Despite (2.), attempting to estimate power is important !

https://github.com/ReproNim/module-stats/blob/gh-pages/notebooks/Misconceptions-Confidence-Intervals.ipynb

Low power - high effect size

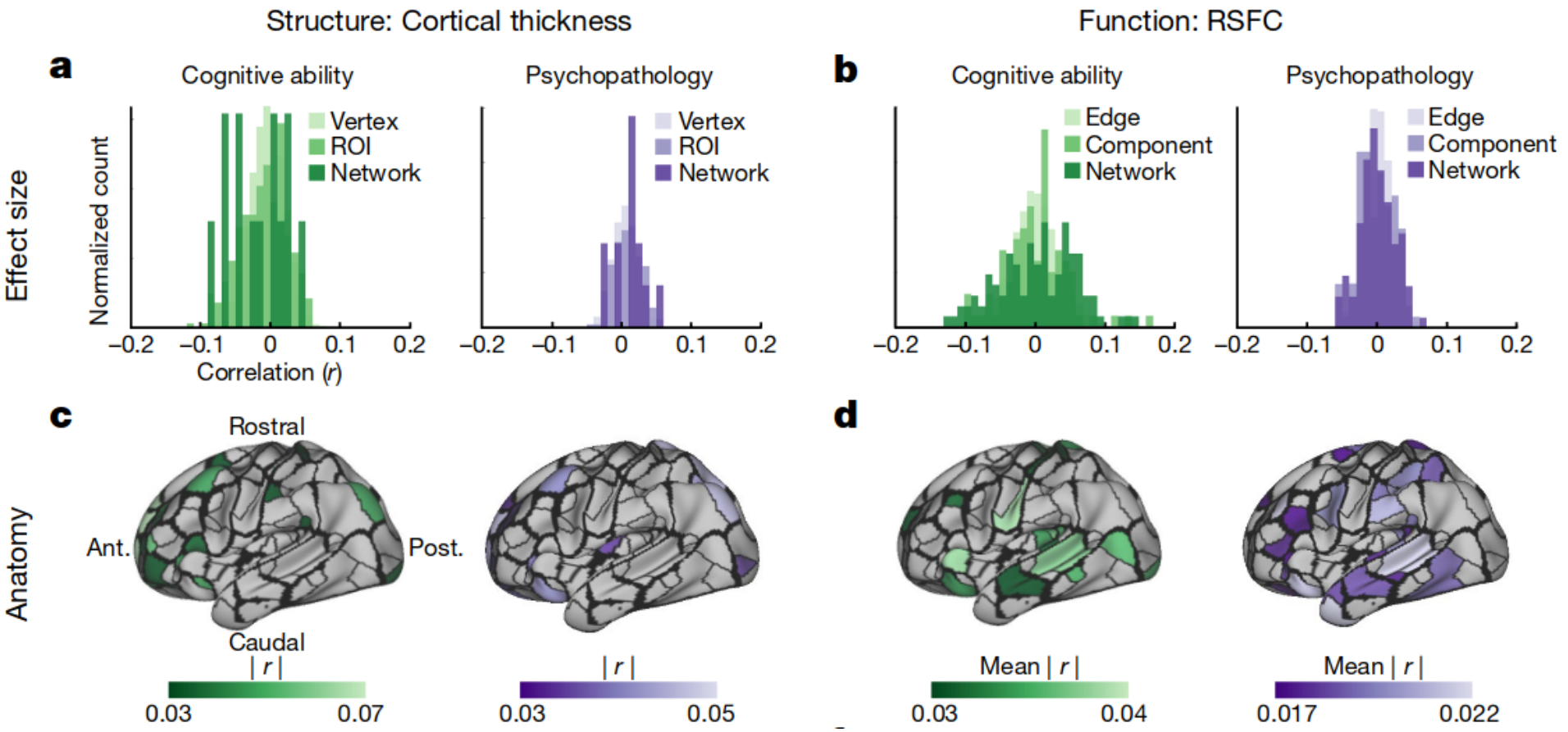




Molendijk, 2012: BDNF and hippocampal volume

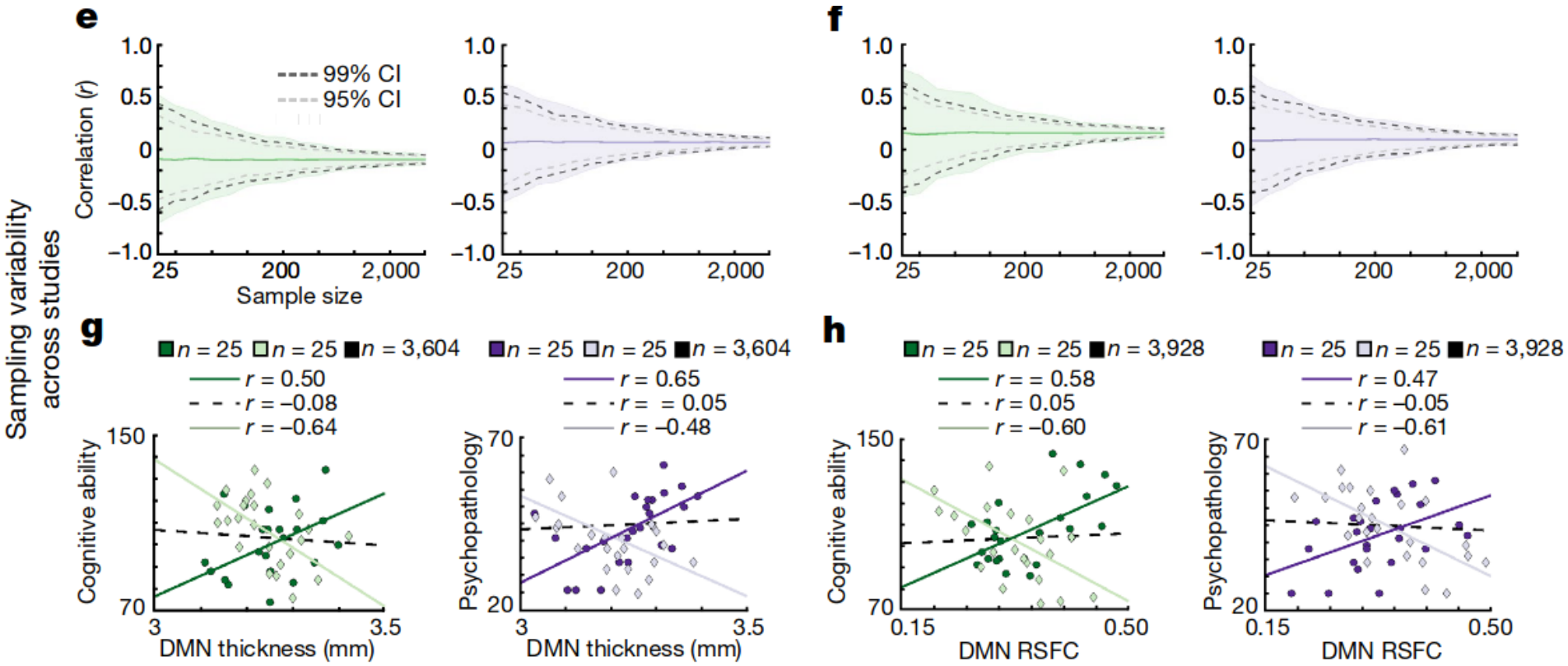
Mier, 2009: COMT and DLPFC

BWAS : Effect size are small



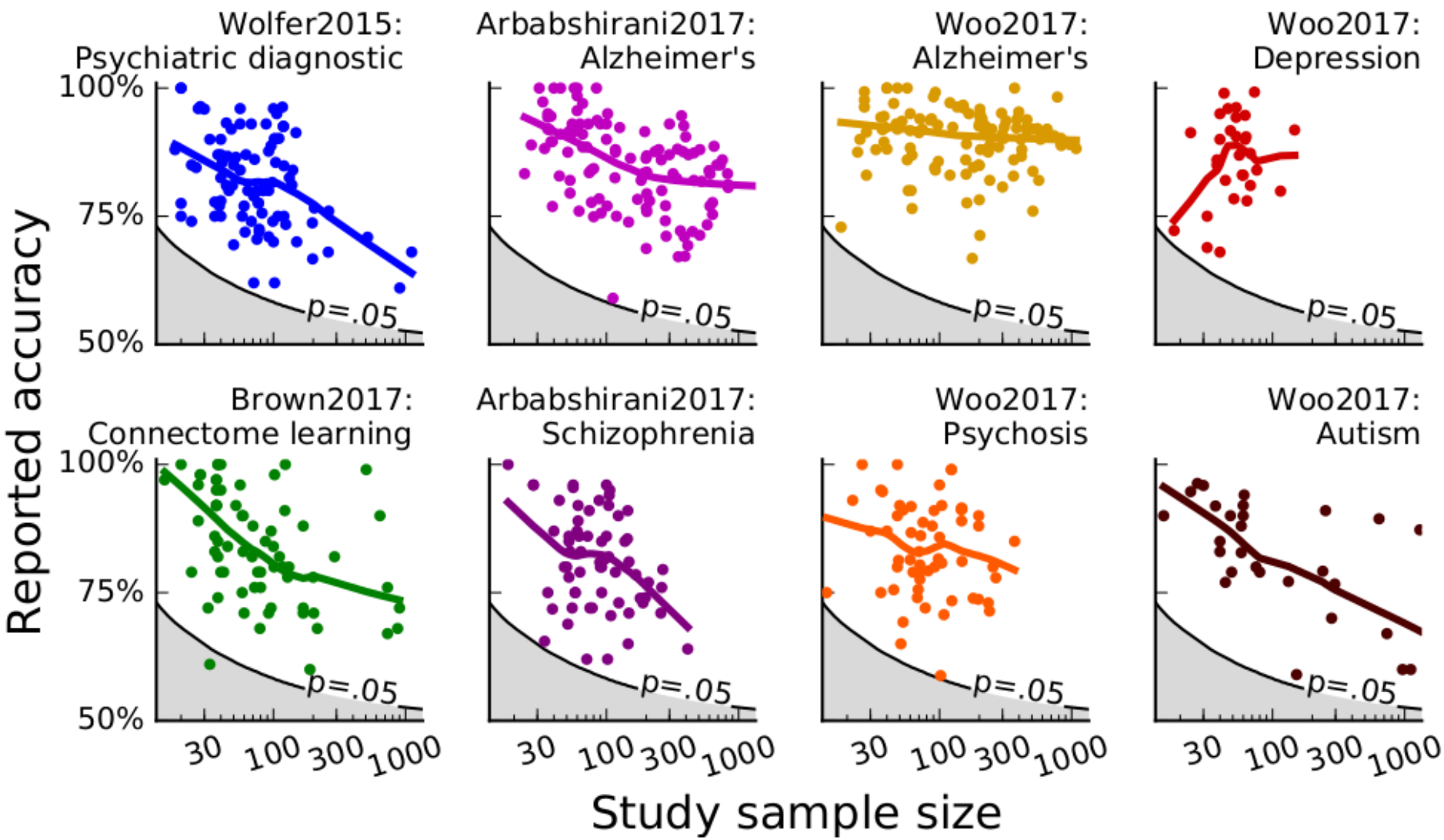
Reproducible brain-wide association studies require thousands of individuals
Marek et al., 2022

BWAS : small N effect size unreliable



Reproducible brain-wide association studies require thousands of individuals
Marek et al., 2022

Sample size issue in ML



Varoquaux, 2017, PsyArXiv, Neuroimage, Cross validation failure

1. P-value and the null hypothesis statistical testing (NHST)
2. P-hacking
3. File drawer
3. Winner's curse
4. Effect sizes
5. Power
6. **Positive Predictive Value**
7. Statistical generalizability

Positive Predictive Value

Open access, freely available online

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

- Positive Predictive Value : The probability that the alternative hypothesis is true knowing that the test is significant
- Requires that “hypothesis” has a probability !

Positive Predictive Value

Open access, freely available online

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

- Positive Predictive Value : The probability that the alternative hypothesis is true knowing that the test is significant
- Requires that “hypothesis” has a probability !

$$PPV = \frac{WR}{WR + \alpha} \qquad R = \frac{P(H_A)}{P(H_0)}$$

W: Power R: odd ratio α :Type I error

But what is a "probability"?

- Objective / Physical : property of the nature or system
  - Associated with a **collective**
- Subjective: a degree of belief ("Evidential probability")
- Frequentist: limit of frequency across random trials
- Bayesian: as reasonable expectation representing a state of knowledge

  Often part of the definition:

  P(A and B) = P(A) P(B given A) = P(A) P(B|A)

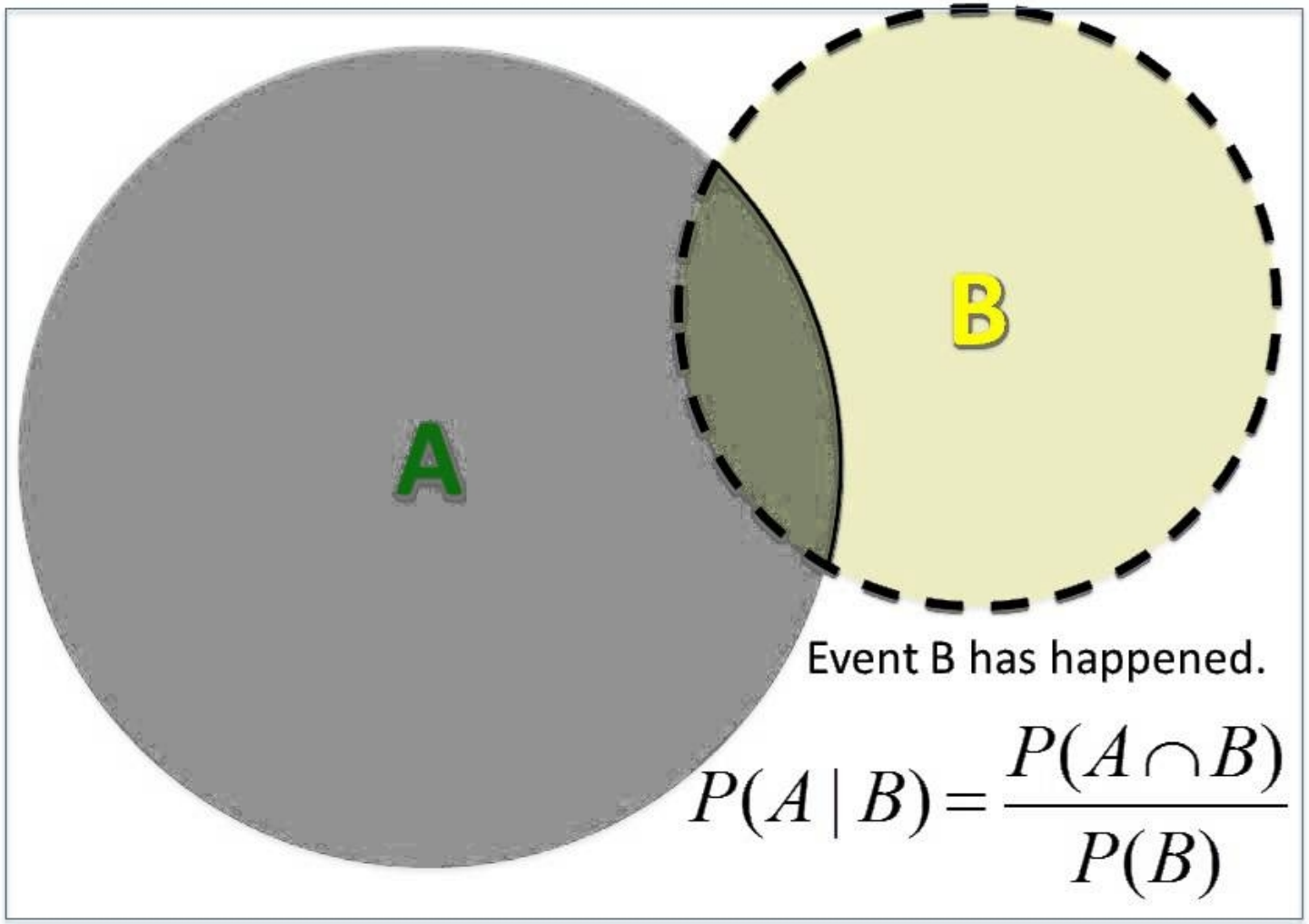
Why is PPV important?

- PPV measure the probability of the alternative hypothesis to be true, knowning that the test is significant:

- Is it worth concluding anything if this number is small?

- This probability depends on
  - Prior probabilities of P(H_A) and of P(H_0) or their ratio
  - Power W
  - Risk of error α (type I error)

http://www.repronim.org/module-stats/05-PPV/



Given that B has happened, the probability that A will happen, too, is just the area ratio of the banana-shaped region to the B circle.

De-mystification

$$P(T_S) = P(T_S, H_A) + P(T_S, H_0)$$

$$P(T_S) = P(T_S \mid H_A)P(H_A) + P(T_S \mid H_0)P(H_0)$$

De-mystification

$$P(T_S) = P(T_S, H_A) + P(T_S, H_0)$$

$$P(T_S) = P(T_S \mid H_A)P(H_A) + P(T_S \mid H_0)P(H_0)$$

$$P(H_A \mid T_S) = \frac{P(T_S \mid H_A)P(H_A)}{P(T_S)}$$

De-mystification

$$P(T_S) = P(T_S, H_A) + P(T_S, H_0)$$

$$P(T_S) = P(T_S \mid H_A)P(H_A) + P(T_S \mid H_0)P(H_0)$$

$$P(H_A \mid T_S) = \frac{P(T_S \mid H_A)P(H_A)}{P(T_S)}$$

$$= \frac{P(T_S \mid H_A)P(H_A)}{P(T_S \mid H_A)Pr(H_A) + Pr(T_S \mid H_0)Pr(H_0)}$$

De-mystification

$$P(T_S) = P(T_S, H_A) + P(T_S, H_0)$$

$$P(T_S) = P(T_S \mid H_A)P(H_A) + P(T_S \mid H_0)P(H_0)$$

$$P(H_A \mid T_S) = \frac{P(T_S \mid H_A)P(H_A)}{P(T_S)}$$

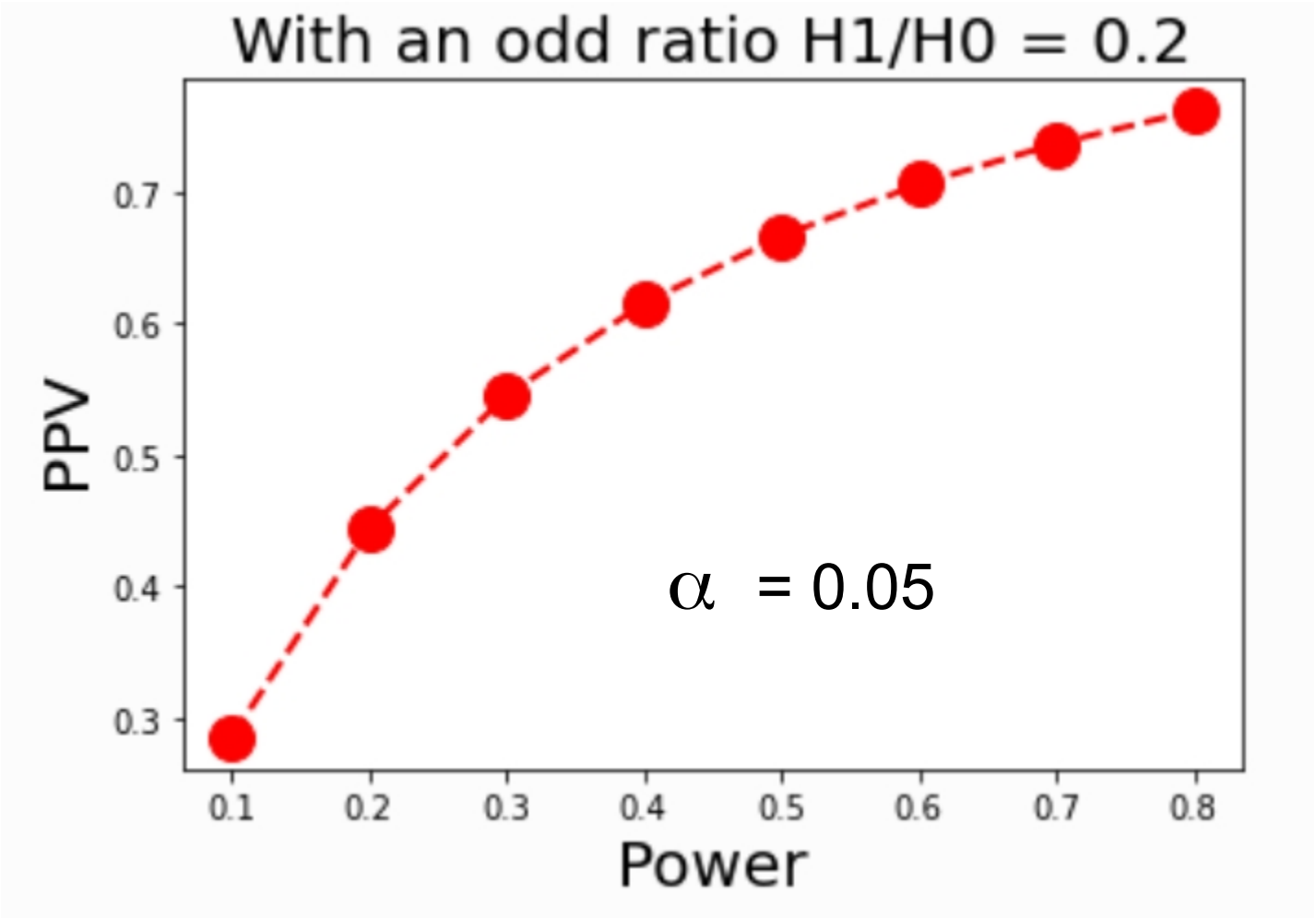
$$= \frac{P(T_S \mid H_A)P(H_A)}{P(T_S \mid H_A)Pr(H_A) + Pr(T_S \mid H_0)Pr(H_0)}$$

$$P(T_S \mid H_A) = W$$

$$P(H_A \mid T_S) = \frac{WP(H_A)}{WP(H_A) + \alpha P(H_0)} = \frac{WR}{WR + \alpha}$$

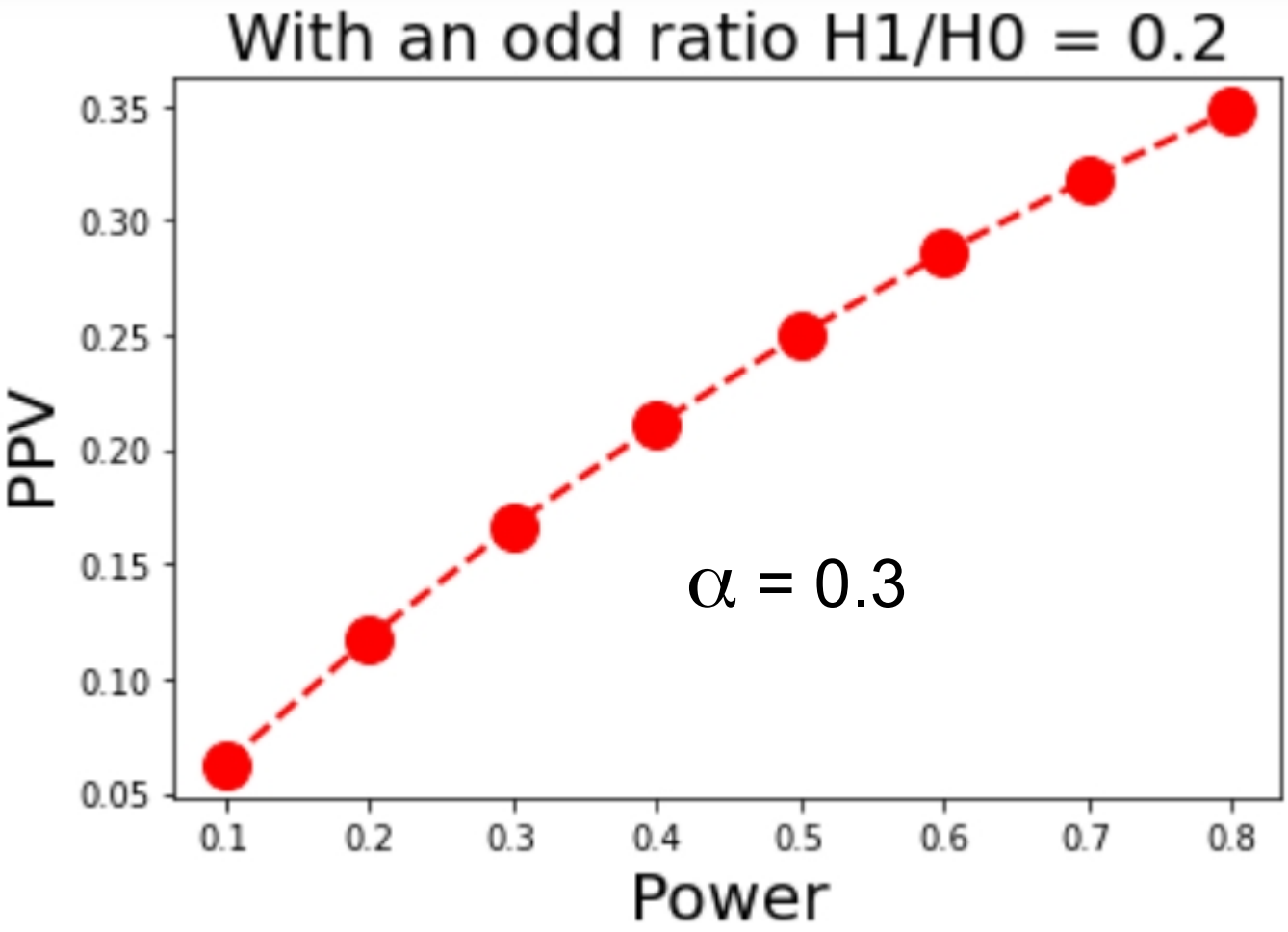
Deriving PPV

$$\text{PPV} = \frac{WR}{WR + \alpha} \qquad R = \frac{P(H_A)}{P(H_0)}$$

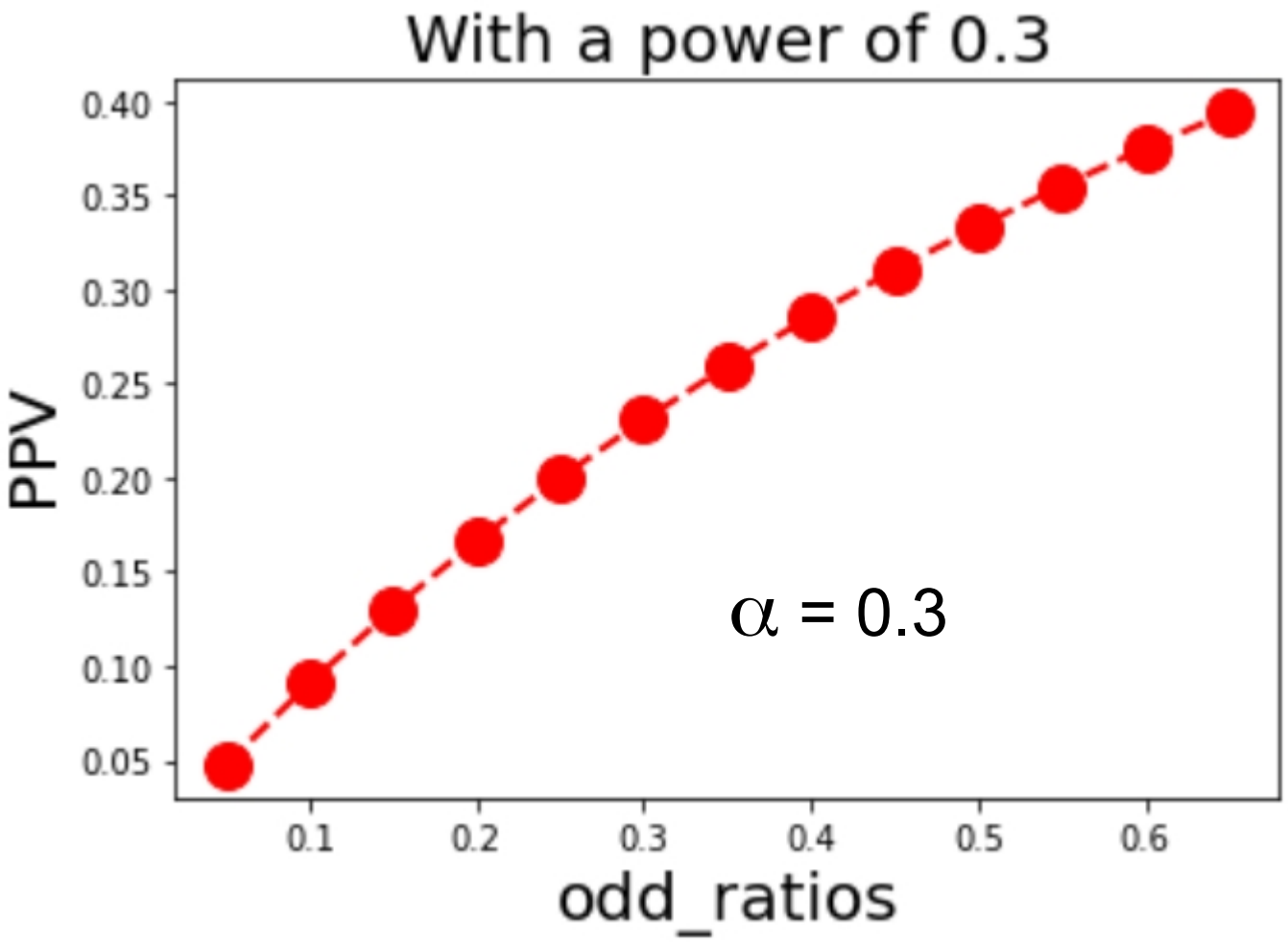


Deriving PPV

$$\text{PPV} = \frac{WR}{WR + \alpha} \qquad R = \frac{P(H_A)}{P(H_0)}$$



$$\text{PPV} = \frac{WR}{WR + \alpha} \qquad R = \frac{P(H_A)}{P(H_0)}$$

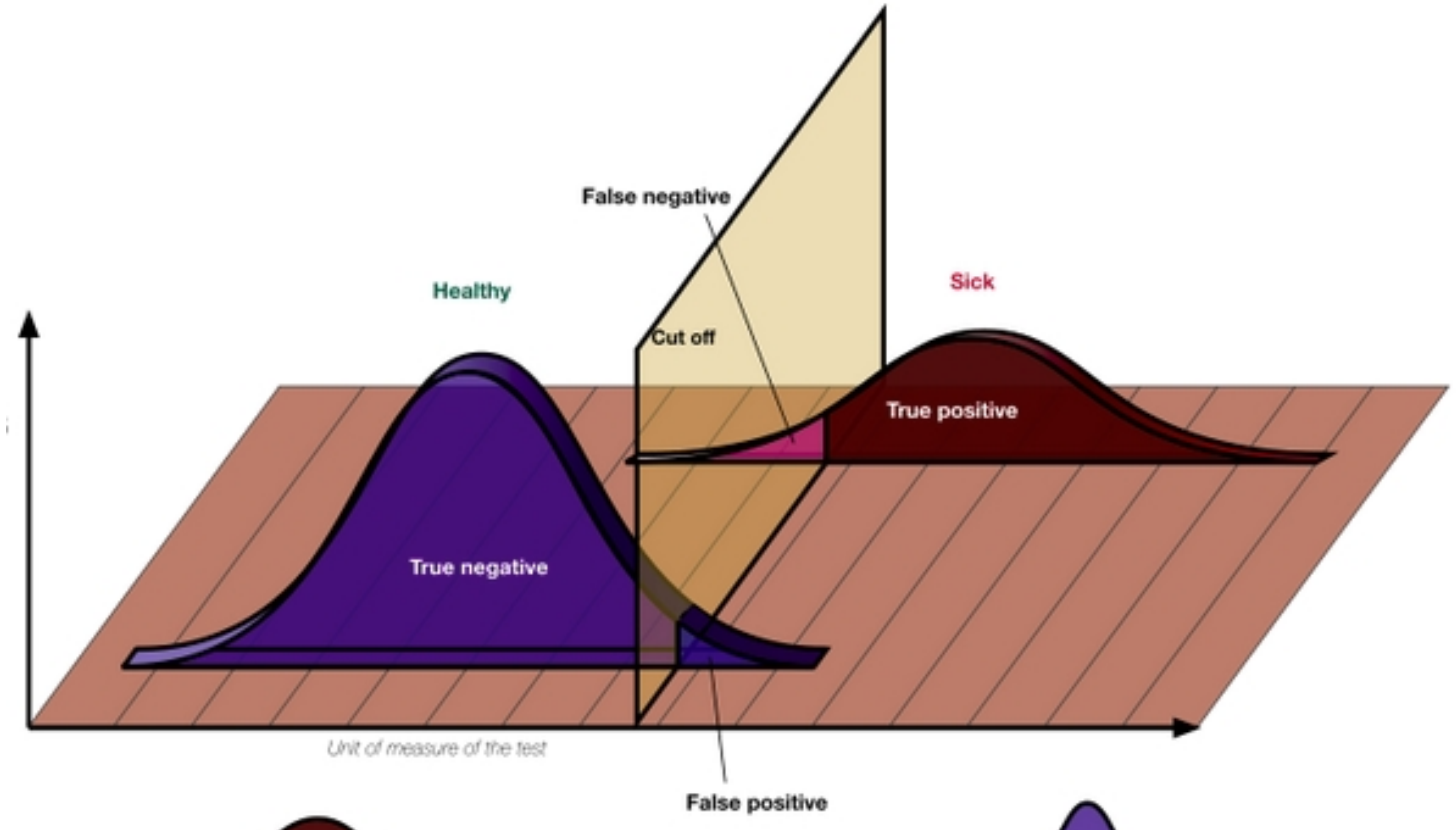


PPV is also “precision” in ML

	Sample is “TRUE”	Sample is “FALSE”
Test positive	True positive	False positive
Test negative	False negative	True negative

$$\text{PPV} = \frac{WR}{WR + \alpha} = \frac{W\,P_1}{W\,P_1 + \alpha P_0} = \frac{TP}{TP + FP}$$

PPV, NPV, Sensitivity, Specificity



PPV =

NPV =

Sensitivity =

Specificity =

Summary

- It is **hard** to estimate well PPV, we need:
  - Power
  - Prior odds $P(H_A) / P(H_0)$
  - Type I error

- These are usually unknown, but can be estimated
  - The process of estimating this quantities helps assess the solidity of the result
  - PPV may be in general quite small

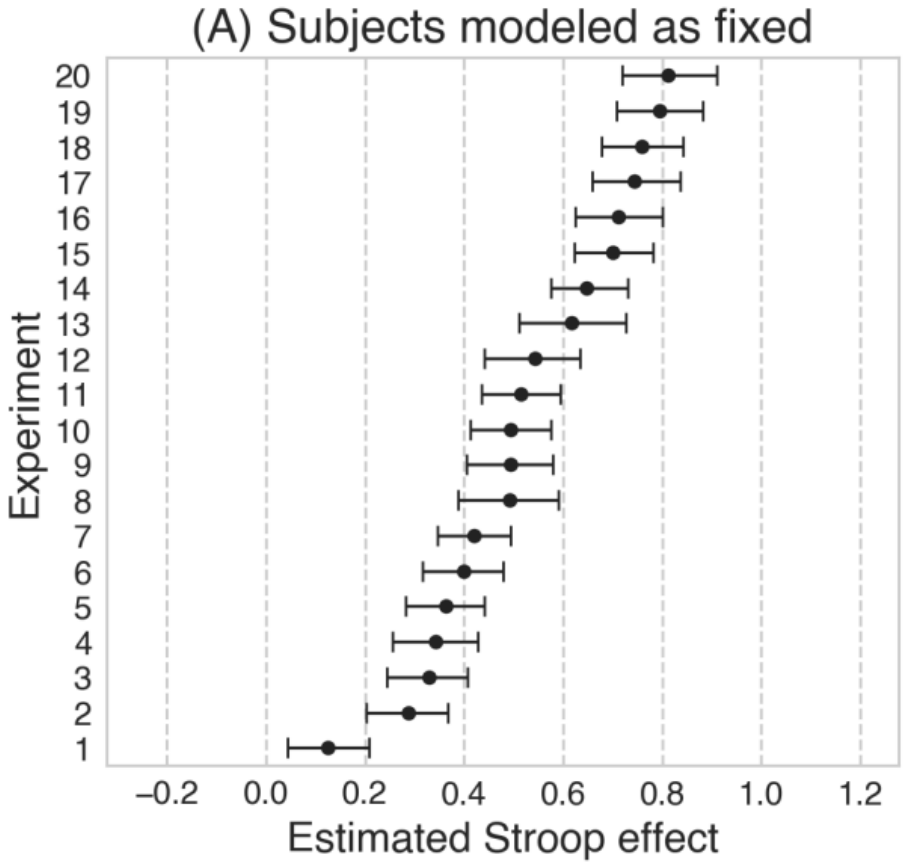
1. P-value and the null hypothesis statistical testing (NHST)
2. P-hacking
3. File drawer
3. Winner's curse
4. Effect sizes
5. Power
6. Positive Predictive Value
7. **Statistical generalizability**

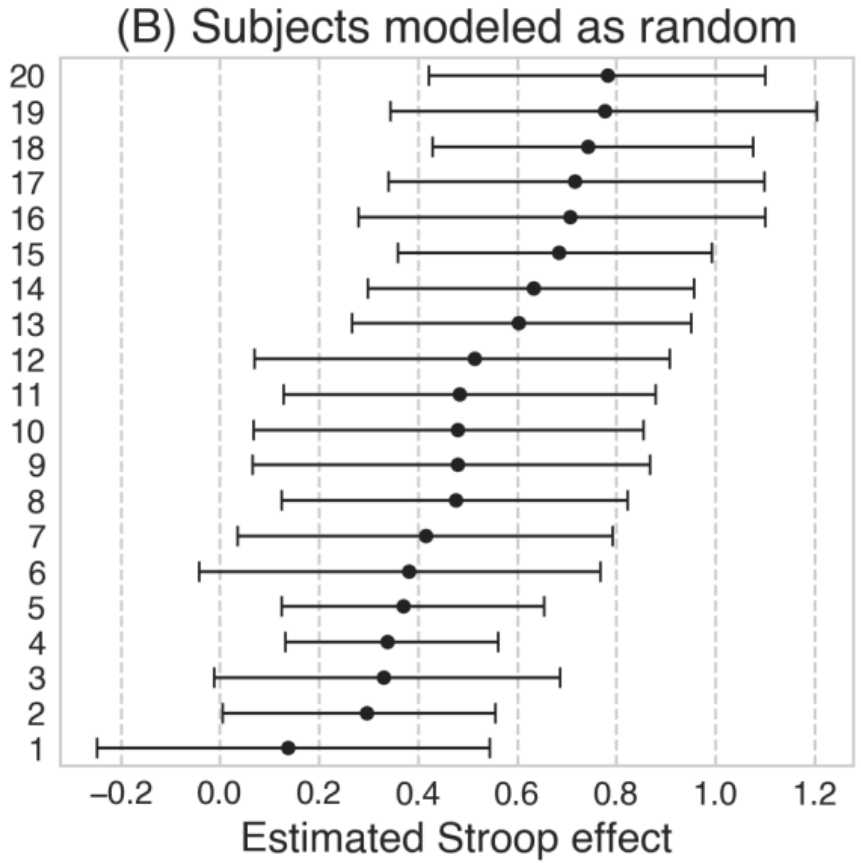
Statistical generalizability

- What is the problem?
  - We consider participants in studies as “random” - but other effects should be “random”
  - Yarkoni et al, the “generalizability crisis”

- Recall what is a “random” versus a “fixed” effect
  - $Y = X\underline{b} + Z\mathbf{g} + \mathbf{e}$
  - Random: consider another source of variance
  - Fixed: noise is source of randomness
  - Example: several observation per groups, many groups, linear regression where slope and intercepts vary with groups

Statistical generalizability





Each row is a simulated Stroop experiment with n = 20 new subjects drawn from the same global population (constant over all experiments). Estimated Bayesian 95% highest posterior density (HPD) intervals for the (fixed) condition effect of interest in each experiment.
(A) The fixed-effects model specification does not account for random subject sampling
(B) The random-effects specification produces appropriately calibrated uncertainty estimates.

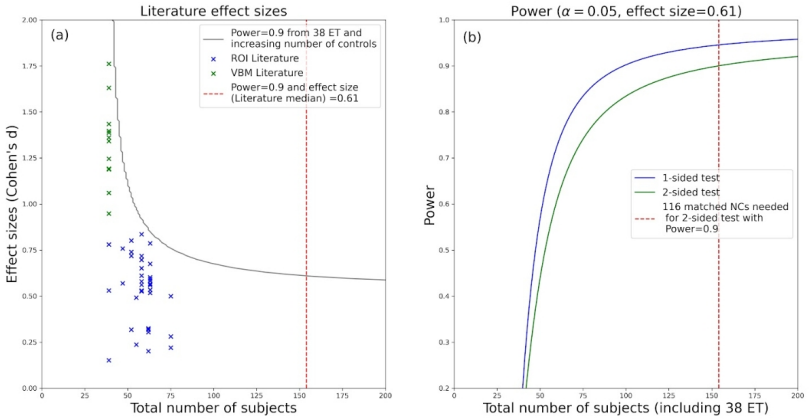
Consequences ?

- Small power —> biased effect sizes
- Winner's curse effect —> biased effect sizes
- P-hacking —> uncontrolled type one error
- File drawer effect —> biased effect sizes

Individual results biases will be reflected in Meta analyses

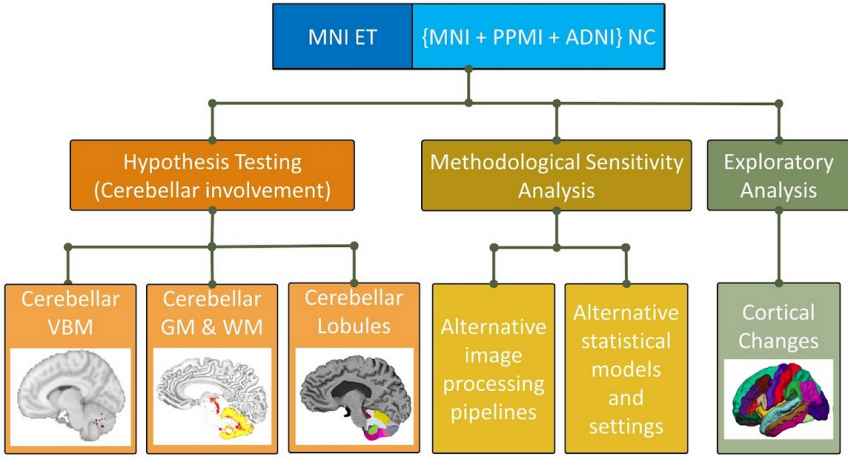
Example of a Meta-analysis

Mavroudis, I. et al. A voxel-wise **meta-analysis on the cerebellum** in essential tremor. Medicina (Mex.) 57, 264 (2021).

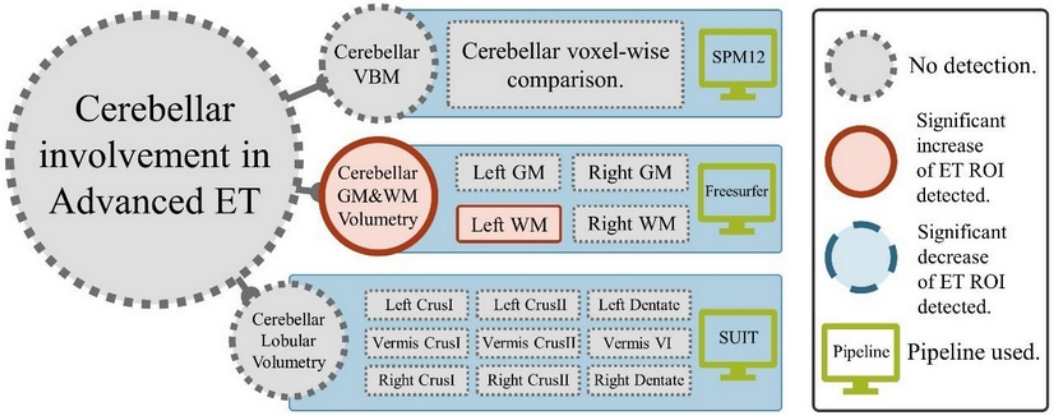


Reproducibility of cerebellar involvement as quantified by consensus structural MRI biomarkers in advanced essential tremor

Qing Wang[1↓], Meshal Aljassar[2↓], Nikhil Bhagwat[1↓], Yashar Zeighami[3], Alan C Evans[3], Alain Dagher[3], G. Bruce Pike[4], Abbas F. Sadikot[2*], Jean-Baptiste Poline[1*]



Meta-analysis not replicated



Pipeline	Cerebellar GrayMatter		Cerebellar WhiteMatter		CrusI (GrayMatter)		CrusII (GrayMatter)	
	L	R	L	R	L	R	L	R
Freesurfer	⇄	↓ 2	↑ 10	↑	–	–	–	–
SUIT	↑ 1	↑ 1	–	–	↑ 2	↑ 1	↑	↑
MAGeT	⇄	↓ 2	↑	⇄	↑	⇄	⇄	⇄

Committee on Reproducibility and Replicability in Science (National Academies of Sciences, Engineering, and Medicine, 2019)

- Non-replicability is a normal part of the scientific process and can be due to the intrinsic variation and complexity of nature, the scope of current scientific knowledge, and the limits of current technologies. Highly surprising and unexpected results are often not replicated by other researchers . . .

- Non-replicability of results is a normal consequence of studying complex systems with imperfect knowledge and tools [emphasis added]. (pp. 85–86)

Committee on Reproducibility and Replicability in Science (National Academies of Sciences, Engineering, and Medicine, 2019)

- Non-replicability is a normal part of the scientific process and can be due to the intrinsic variation and complexity of nature, the scope of current scientific knowledge, and the limits of current technologies. Highly surprising and unexpected results are often not replicated by other researchers . . .

- Non-replicability of results is a normal consequence of studying complex systems with imperfect knowledge and tools [emphasis added]. (pp. 85–86)

- A reproducibility story
- Some statistical aspects
- What should we solve – and how?

Sociological components

1. Research is a competitive space
2. Success is often defined by Sum(papers*IF*1/author-positions)
3. Articles and “narratives” are the clear winning currency for hiring / promotion
4. Other research objects have less impact
    1. Software, library, packages
    2. Educational material, courses
    3. Datasets
    4. Standards

Sociological components

1. Research is a competitive space
2. Success is often defined by Sum(papers*IF*author-positions)
3. Articles and “narratives” are the clear winning currency for hiring / promotion
4. Other research objects have less impact
    1. Software, library, packages
    2. Educational material, courses
    3. Datasets
    4. Standards

Time, fear and culture

1. Time
    1. I need to have 3 papers for my PhD, I don't have time to document the data or develop a good package.
    2. The conference deadline is in 3 weeks, no way I can study the robustness of the results
2. Fear
    1. Data: Scooping
    2. My code is not good enough to show
    3. Preregistration
3. Culture
    1. Medical school vs engineering (reusability)
    2. Competition vs Collaboration
    3. Importance of reputation
    4. Department or Institute's favorite topic / journal / ...

Time, fear and culture

1. Time
    1. I need to have 3 papers for my PhD, I don't have time to document the data or develop a good package.
    2. The conference deadline is in 3 weeks, no way I can study the robustness of the results
2. Fear
    1. Data: Scooping
    2. My code is not good enough to show
    3. Preregistration: can be used to protect yourself
3. Culture
    1. Medical school vs engineering (reusability)
    2. Competition vs Collaboration
    3. Importance of reputation
    4. Department or Institute's favorite topic / journal / ...

Time, fear and culture

1. Time
    1. I need to have 3 papers for my PhD, I don't have time to document the data or develop a good package.
    2. The conference deadline is in 3 weeks, no way I can study the robustness of the results
2. Fear
    1. Data: Scooping
    2. My code is not good enough to show
    3. Preregistration

3. Culture
    1. Biology/medicine vs engineering
    2. Competition vs Collaboration
    3. Reputation / the fusion between your work and your name
    4. Department or Institute's favorite topic / journal / ...

Fight the misplaced incentives

ROYAL SOCIETY OPEN SCIENCE

rsos.royalsocietypublishing.org

Research

The natural selection of bad science

Paul E. Smaldino[1] and Richard McElreath[2]

[1]Cognitive and Information Sciences, University of California, Merced, CA 95343, USA
[2]Department of Human Behavior, Ecology, and Culture, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

A business like none other

illustration for science publishing long read Illustration: Dom McKenzie

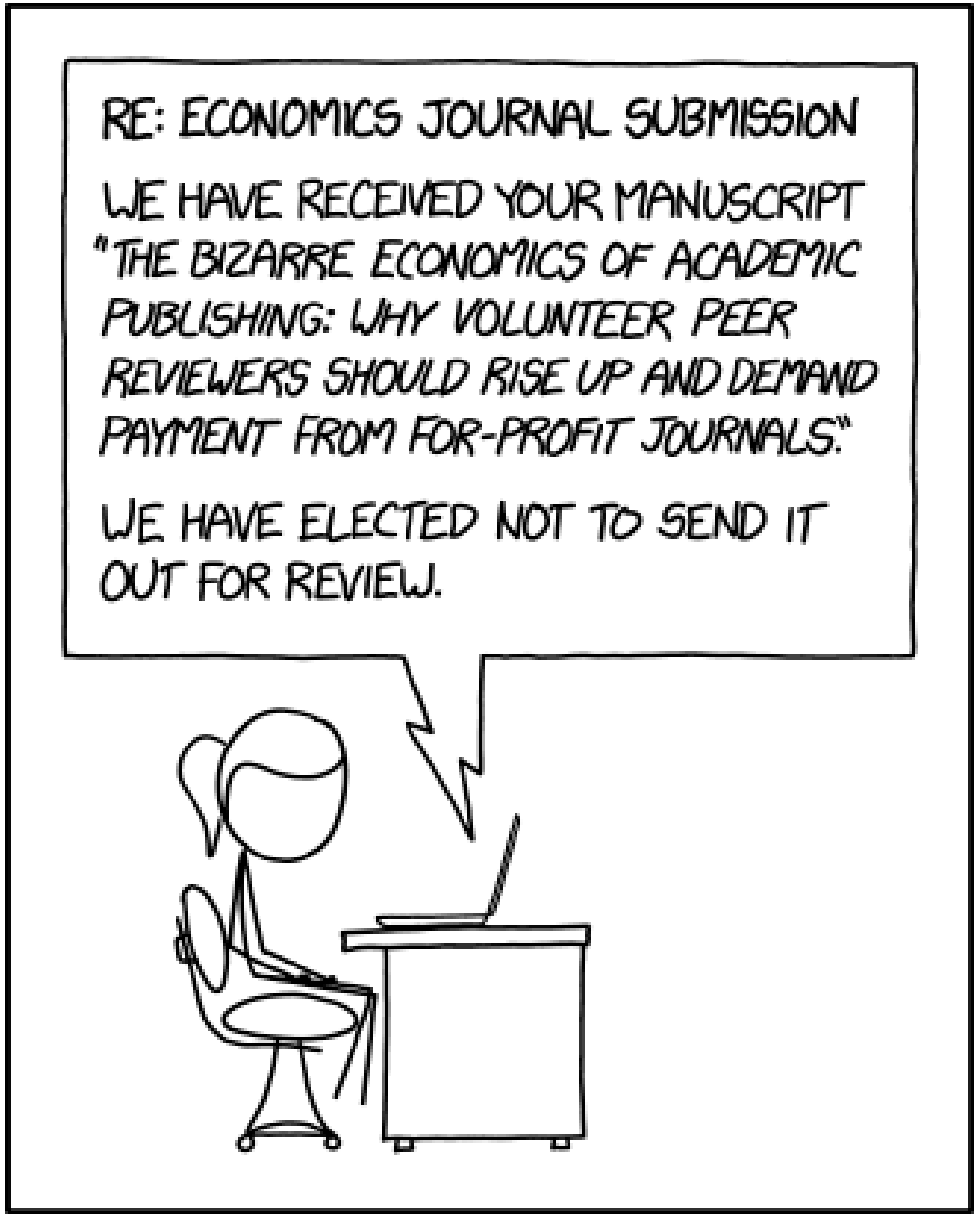
The long read

Is the staggeringly profitable business of scientific publishing bad for science?

It is an industry like no other, with profit margins to rival Google – and it was created by one of Britain's most notorious tycoons: Robert Maxwell

by Stephen Buranyi

Business and culture interaction



Reactions: the cost of knowledge

The Cost of Knowledge

20435 Researchers Taking a Stand. *See the list*

Academics have protested against Elsevier's business practices for years with little effect. These are some of their objections:

1. They charge exorbitantly high prices for subscriptions to individual journals.
2. In the light of these high prices, the only realistic option for many libraries is to agree to buy very large "bundles", which will include many journals that those libraries do not actually want. Elsevier thus makes huge profits by exploiting the fact that some of their journals are essential.
3. They support measures such as SOPA, PIPA and the ~~Research Works Act~~, that aim to restrict the free exchange of information.

The key to all these issues is the right of authors to achieve easily-accessible distribution of their work. If you would like to declare publicly that you will not support any Elsevier journal unless they radically change how they operate, then you can do so by filling in your details on this page.

More information:

Add your name to the list.

First and Last Name

Affiliation

Email

only used once to verify your identity; never displayed, never shared

Subject: Mathematics

Comments *(optional)*

Link *(optional)*

such as a link to a blog post of yours explaining your position

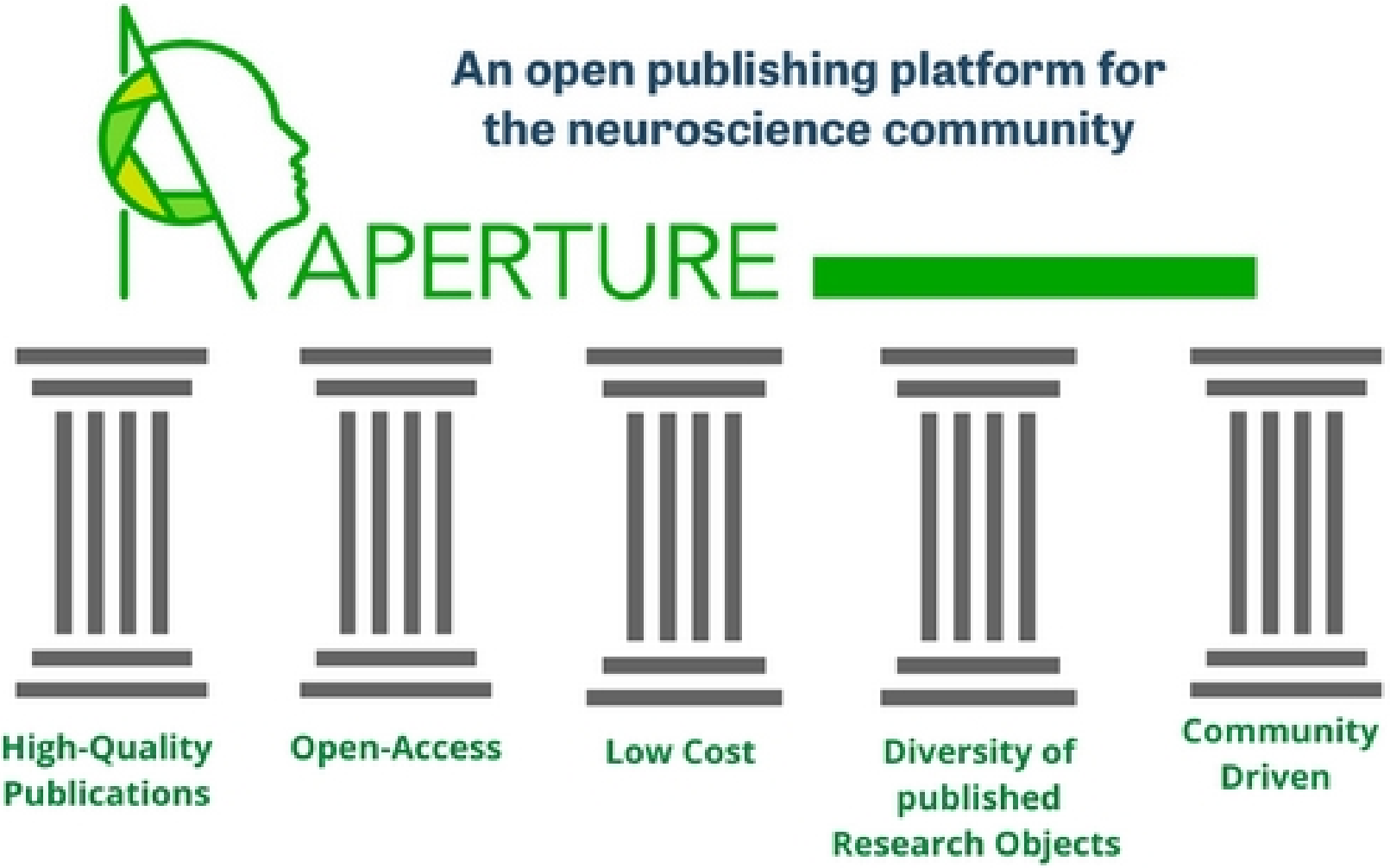
I plan to refrain from:

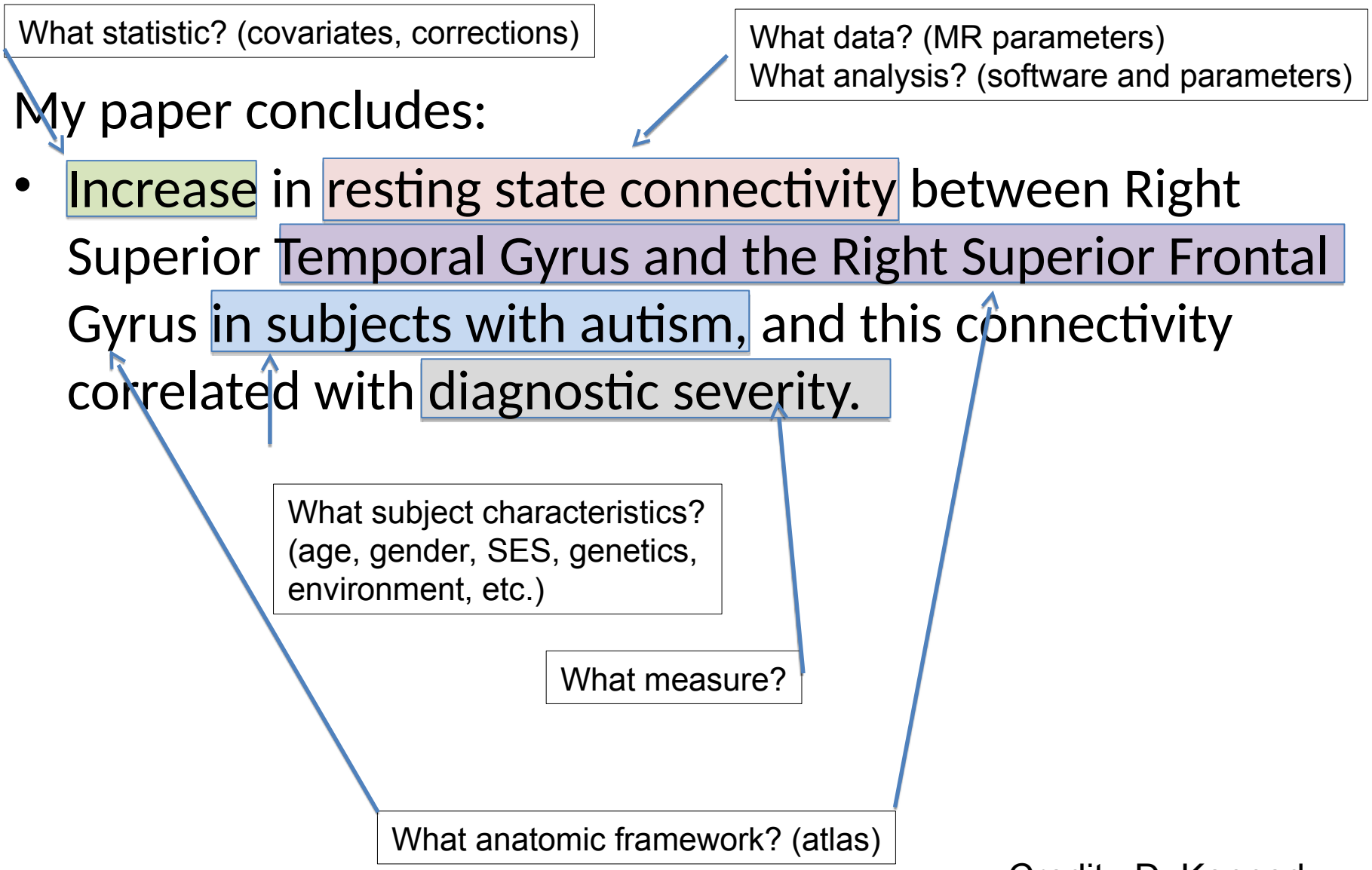
☐ publishing ☐ refereeing ☐ editorial work

Add My Name

Cost of knowledge started by Timothy Gowers. In 2019, the UC announced that it was cancelling its Elsevier subscriptions, followed by MIT (2020) SUNY, etc. … but deals were then struck

What is needed ?





Credit : D. Kennedy

Thank you

- Lab@McGill**: https://neurodatascience.github.io/**
- **McGill** colleagues: S. Brown, T. Glatard, G. Kiar, A. Evans, C. Greenwood, A. DeGuise and others
- **ReproNim** colleagues: D. Kennedy, D. Keator, S. Ghosh, M. Martone, J. Grethe, M. Hanke, Y. Halchenko
- **Berkeley** colleagues: S. Van der Walt, M. Brett, J. Millman, Dan Lurie, M. D'Esposito, et al
- **Pasteur** colleagues: G. Dumas, R. Toro, T. Bourgeron, and others
- **Paris** colleagues: B. Thirion, G. Varoquaux, V. Frouin, et al
- **Funders**: McGill HBHL, HBHL NeuroHub, NIH, NIMH

Proposal for a project

This would help understand the consequences of effects such as

- File drawer
- P-hacking
- Low / high study power
- Pre-registration

- Simulate the research world with N laboratories with different culture and practices

- **Science on a wire** : **https://github.com/neurodatascience/QLS-course-materials/tree/main/project**