

The issues with statistics that will make your work irreproducible and what we should do about it

Jean-Baptiste Poline

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

“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

So, if we want to learn about reality, we need to
FALSIFY

What are our tools for falsification ? Is rejecting the
null falsification ? Which null should be rejected
then ?

If I want to learn how TBI affect brain connectivity,
what do I need to reject ?

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](https://doi.org/10.1016/S0140-6736(15)60696-1)

- Most common causes :
- Small sample size
 - Analytical flexibility
 - p-hacking
 - Poor statistical training

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

		Data	
		Same	Different
Code & Analysis	Same	Reproducible	Replicable
	Different	Robust	Generalisable

- 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-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

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.

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.

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.

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

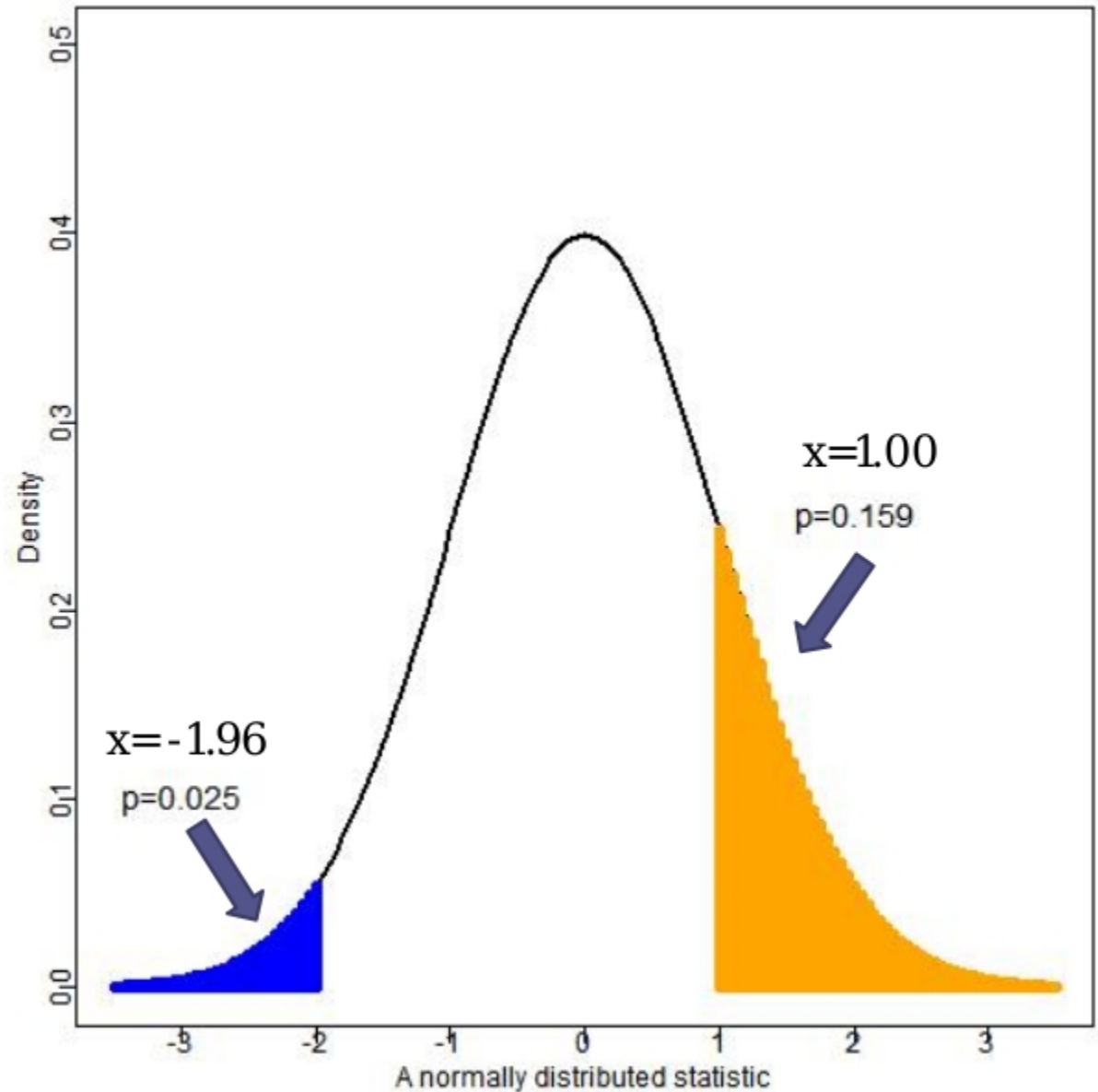
- What is a “statistic” ?
 - Any function of the data
 - the mean,
 - the SD,
 - the mean/SD,
 - the t statistics,
 - the z statistics,
 - etc

- 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

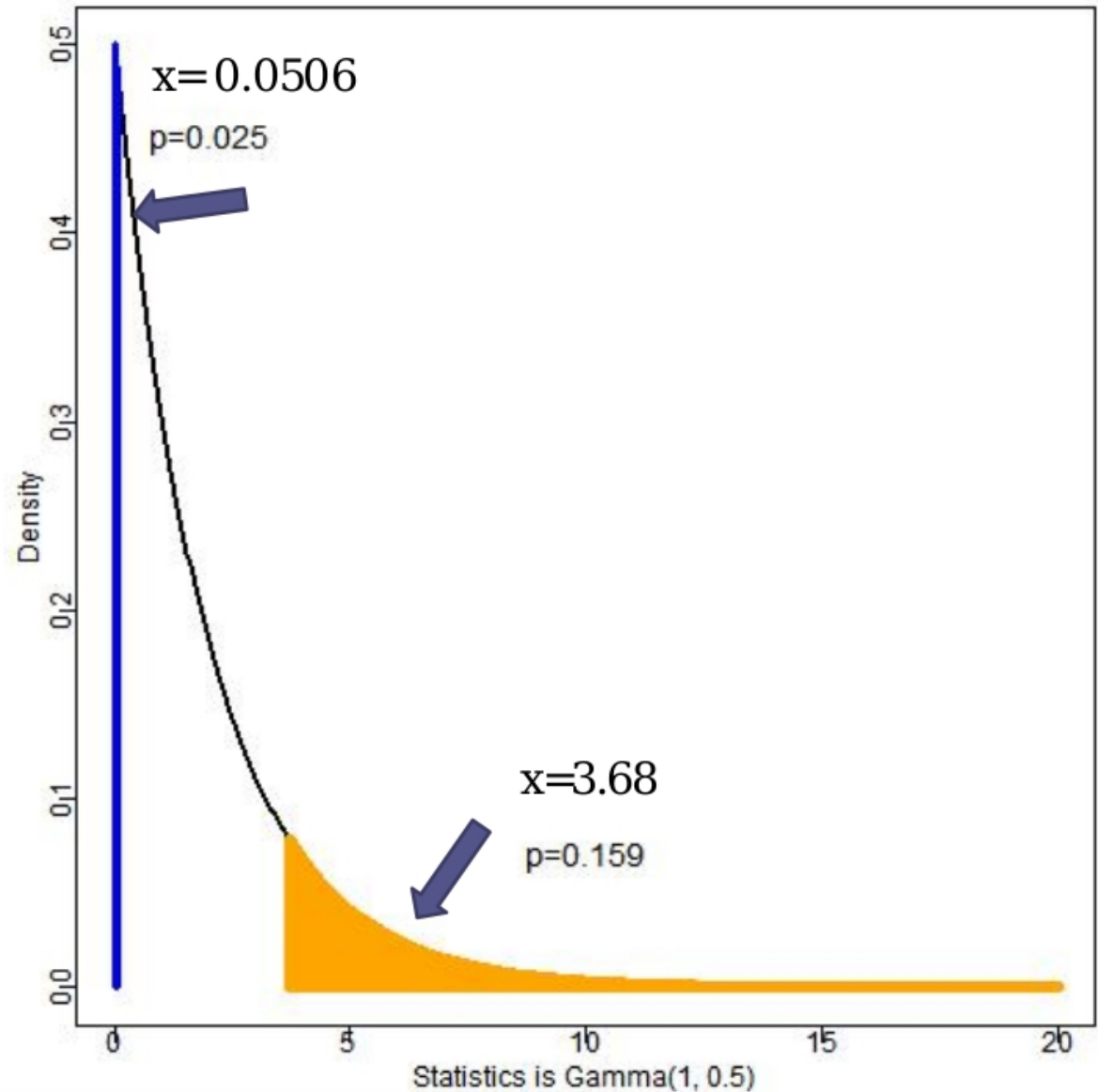
- *“We are inclined to think that as far as a particular hypothesis is concerned, no test based upon a theory of probability can by itself provide any valuable evidence of the truth or falsehood of a hypothesis”*

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.

A normally
distributed
statistic

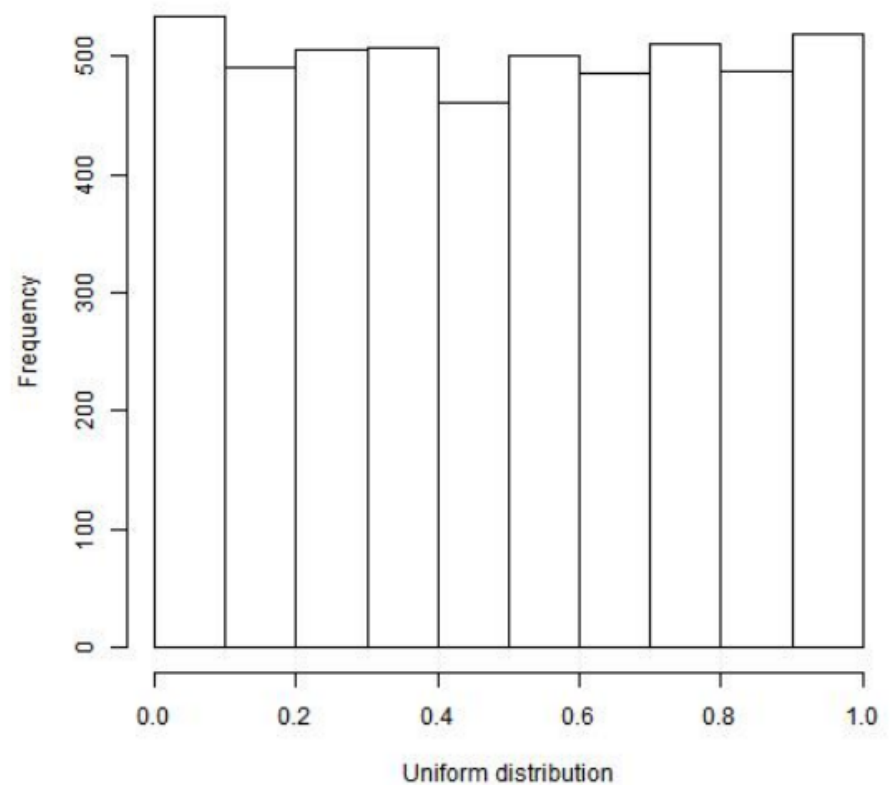


A gamma-
distributed
statistic
 $\text{Gamma}(1, 0.5)$



Uniform distribution

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

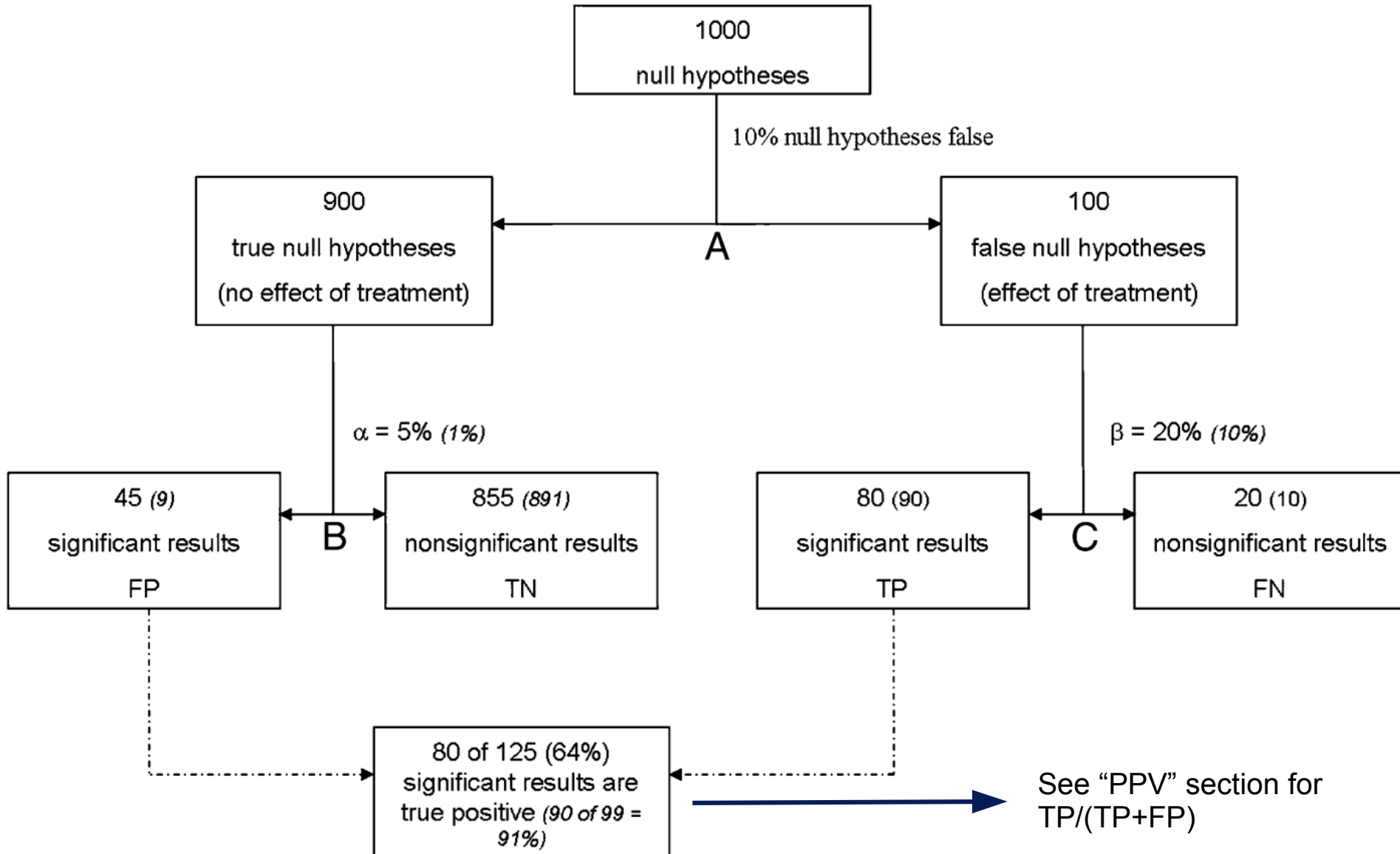


- We want the probability of rejecting the null to be alpha : $P(\text{observed-statistics} > \text{quantile-}\alpha) = \alpha$
- The probability that the p-value p is smaller than 5% is the probability that the observed statistic is above the 95th percentile of the null, so is 5%
- Hence: $P(p \leq x) = x$
 - *Fact used in the “p-hacking test”*
- A p-value is a statistic : it is a function of the data - therefore random

- “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.

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?



- 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 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 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

False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant

Joseph P. Simmons¹, Leif D. Nelson², and Uri Simonsohn¹

¹The Wharton School, University of Pennsylvania, and ²Haas School of Business, University of California, Berkeley

Table 1. Likelihood of Obtaining a False-Positive Result

Researcher degrees of freedom	Significance level		
	$p < .1$	$p < .05$	$p < .01$
Situation A: two dependent variables ($r = .50$)	17.8%	9.5%	2.2%
Situation B: addition of 10 more observations per cell	14.5%	7.7%	1.6%
Situation C: controlling for gender or interaction of gender with treatment	21.6%	11.7%	2.7%
Situation D: dropping (or not dropping) one of three conditions	23.2%	12.6%	2.8%
Combine Situations A and B	26.0%	14.4%	3.3%
Combine Situations A, B, and C	50.9%	30.9%	8.4%
Combine Situations A, B, C, and D	81.5%	60.7%	21.5%

- 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

- Pre-registration
- Ban p-values
- Change α
- 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

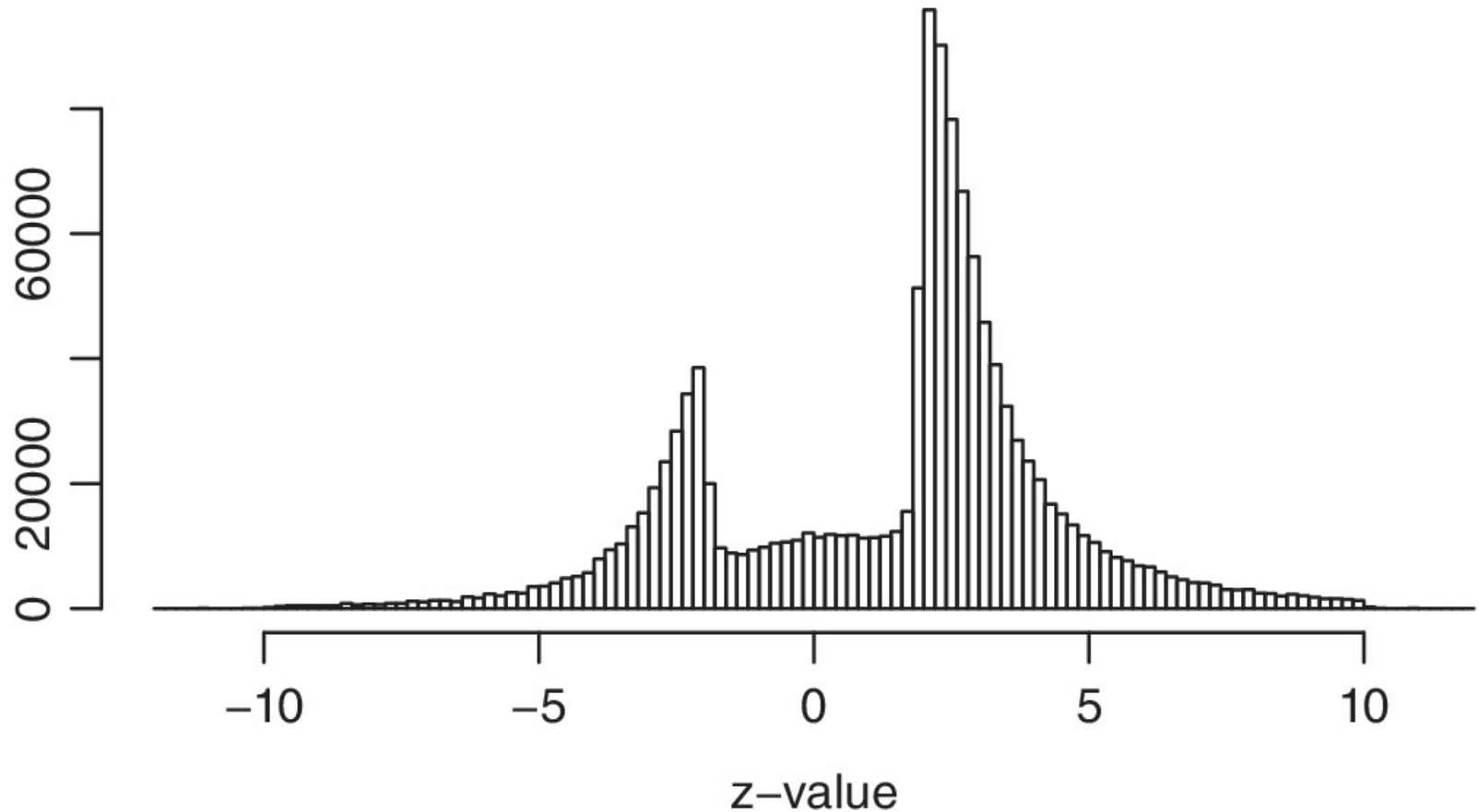
The “File Drawer Problem” and Tolerance for Null Results

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.

- 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.



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

- Pre-registration !
 - Cf “reproducibility lesson”
- Convince journals that null results need to be published
- Can the importance of a null result be stated? ie.
Can we assess the 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

- 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

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

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

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

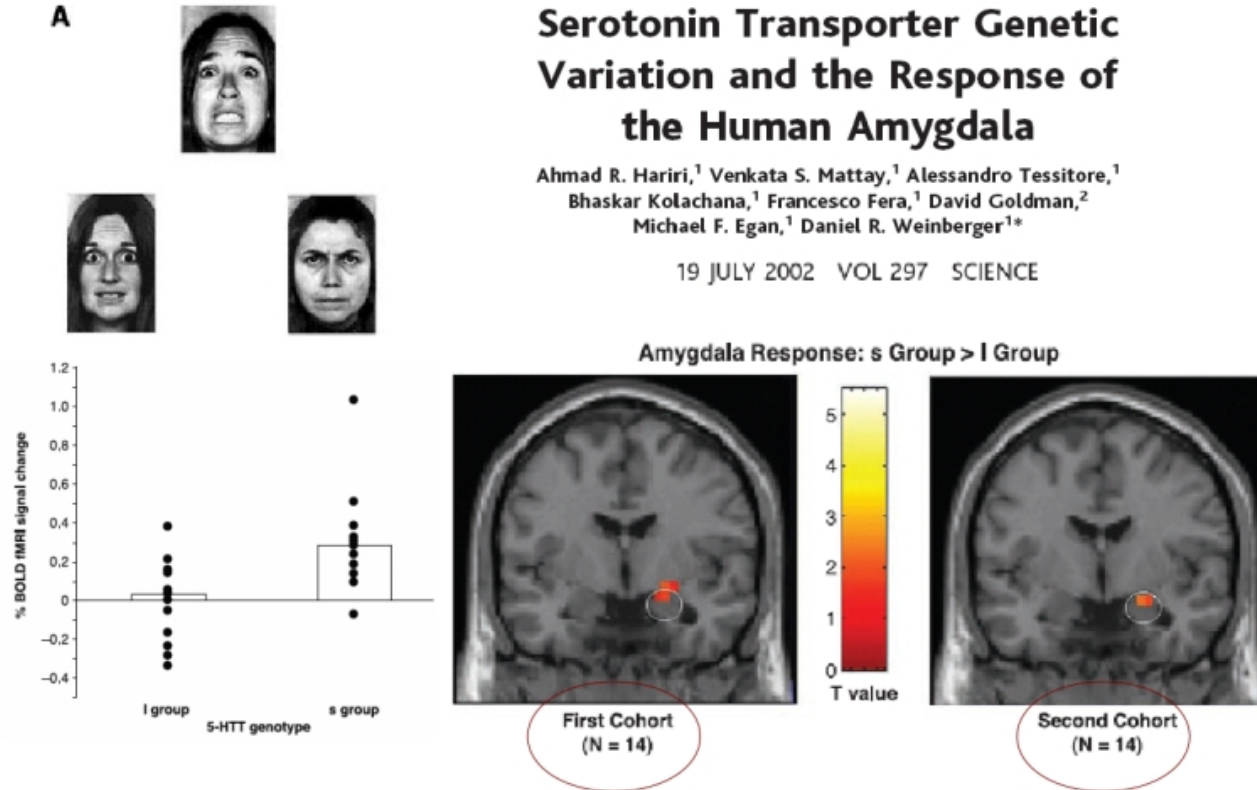
- A vague notion because any “measure of interest” can be an effect size
- Good quality for an “effect size” :
 - Simple
 - Interpretable - with units !
 - Standard in your field of science
- Examples:
 - Percentage of variance explained by a model
 - Correlation
 - Difference between means of two groups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

- 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 unnormalized
 - Standard deviations and standard deviation of the means
 - Confidence / Credible intervals
 - (not generally in guidelines) : Some bootstrapping of the data if possible to give an idea of the “results distribution”

Serotonin Transporter Genetic Variation and the Response of the Human Amygdala

Ahmad R. Hariri,¹ Venkata S. Mattay,¹ Alessandro Tessitore,¹
Bhaskar Kolachana,¹ Francesco Fera,¹ David Goldman,²
Michael F. Egan,¹ Daniel R. Weinberger^{1*}

19 JULY 2002 VOL 297 SCIENCE



- Authors report
 $m_1 = .28$, $m_2 = .03$, $SDM_1 = 0.08$, $SDM_2 = 0.05$, $N_1 = N_2 = 14$
- How do we compute the effect size ?

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

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

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

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

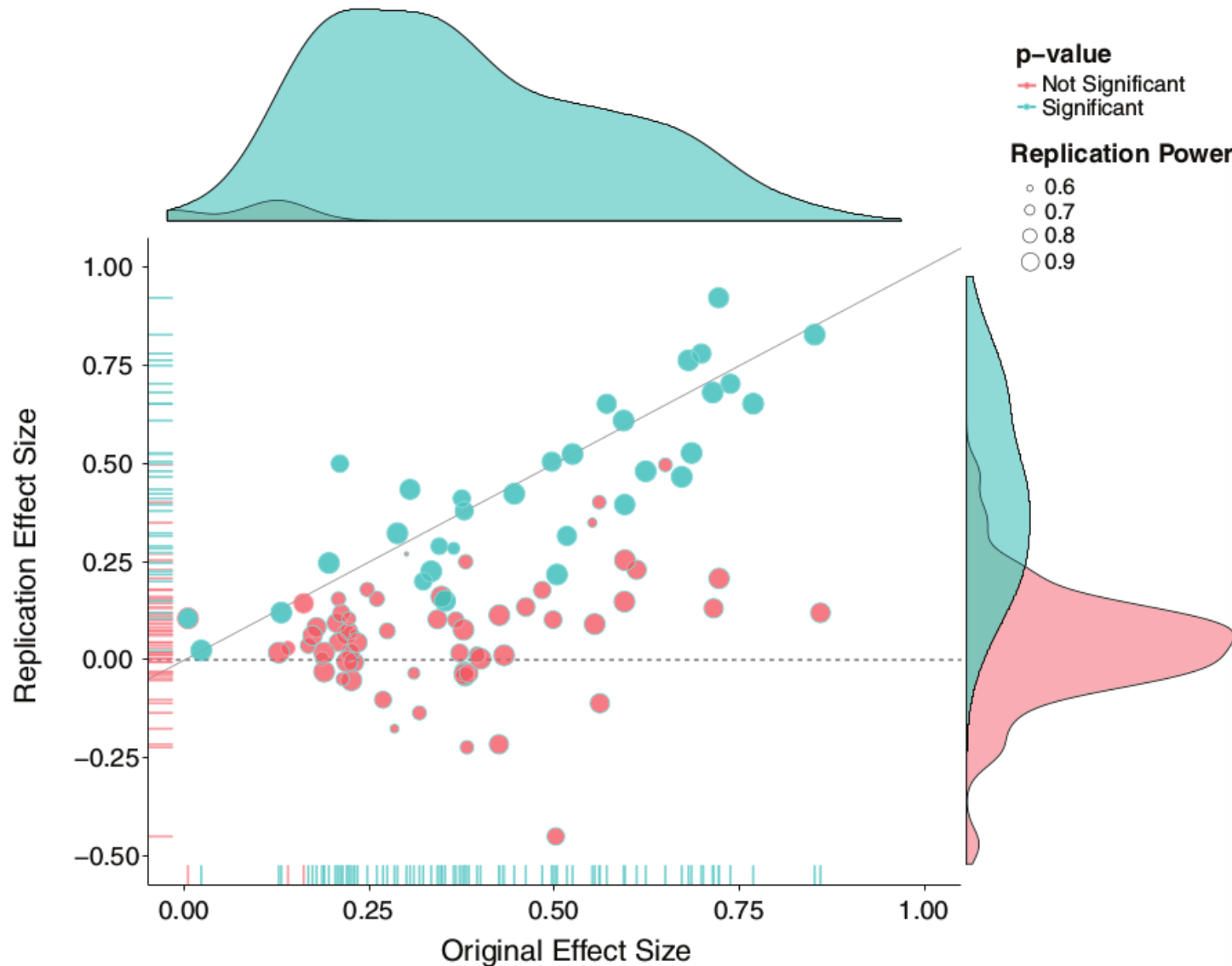
$$m_1=.28, m_2=.03 \quad d = \frac{m_1 - m_2}{\sigma} = 1.05$$

- What is the percentage of variance explained ?
- Write the estimated model: $Y = [1 \dots 1]^t [m_1 - m_2] + \text{residual}$
- Compute the total sum of square $Y^t Y$, then the proportion:

$$V_e = \frac{(n_1 + n_2)(m_1 - m_2)^2}{n_1 s_1^2 + n_2 s_2^2 + (n_1 + n_2)(m_1 - m_2)^2} > 40\%$$

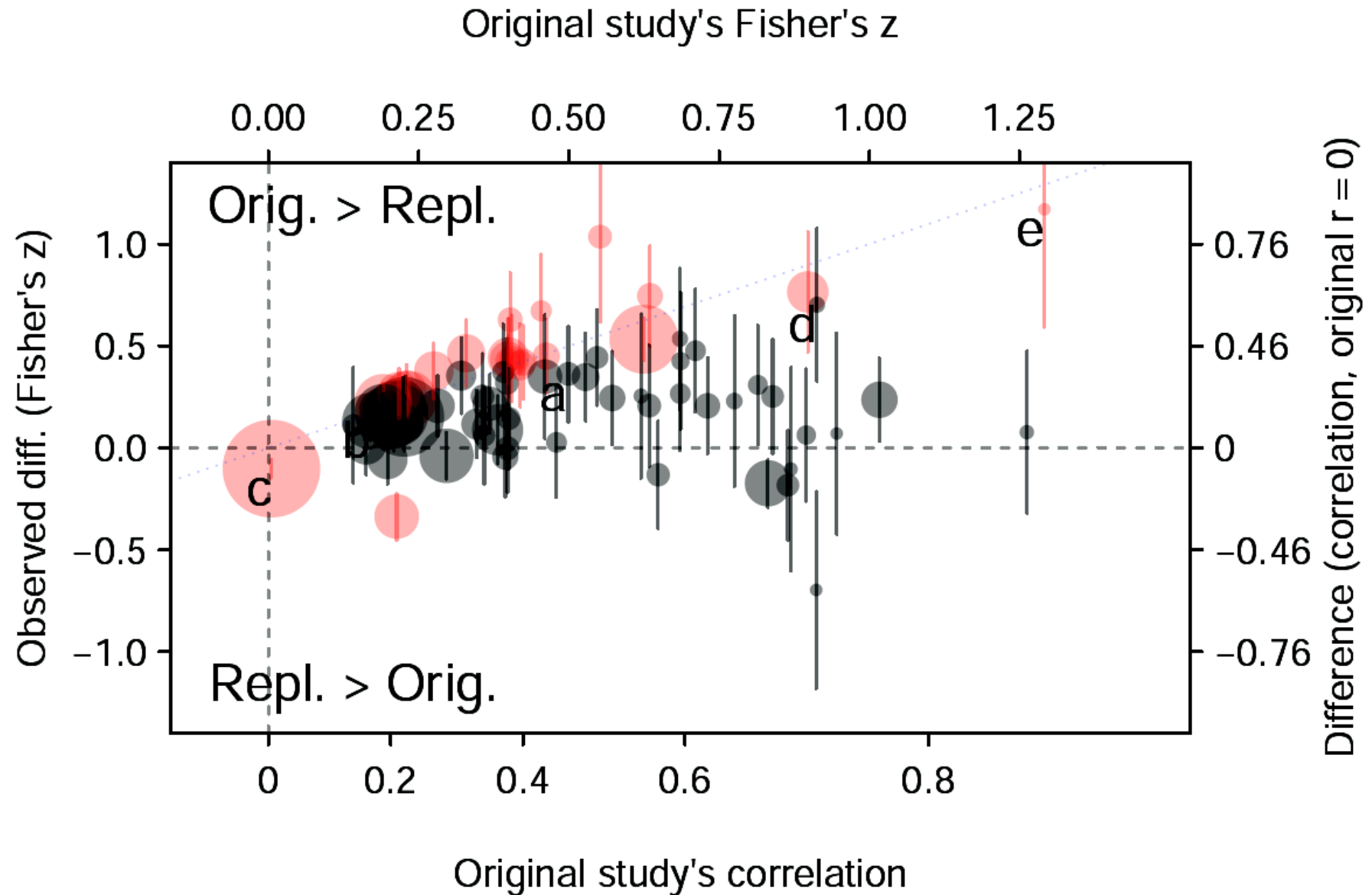
First, compute the standard deviation of the data from the SDM

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



* 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 ($M r = 0.403$, $SD = 0.188$)

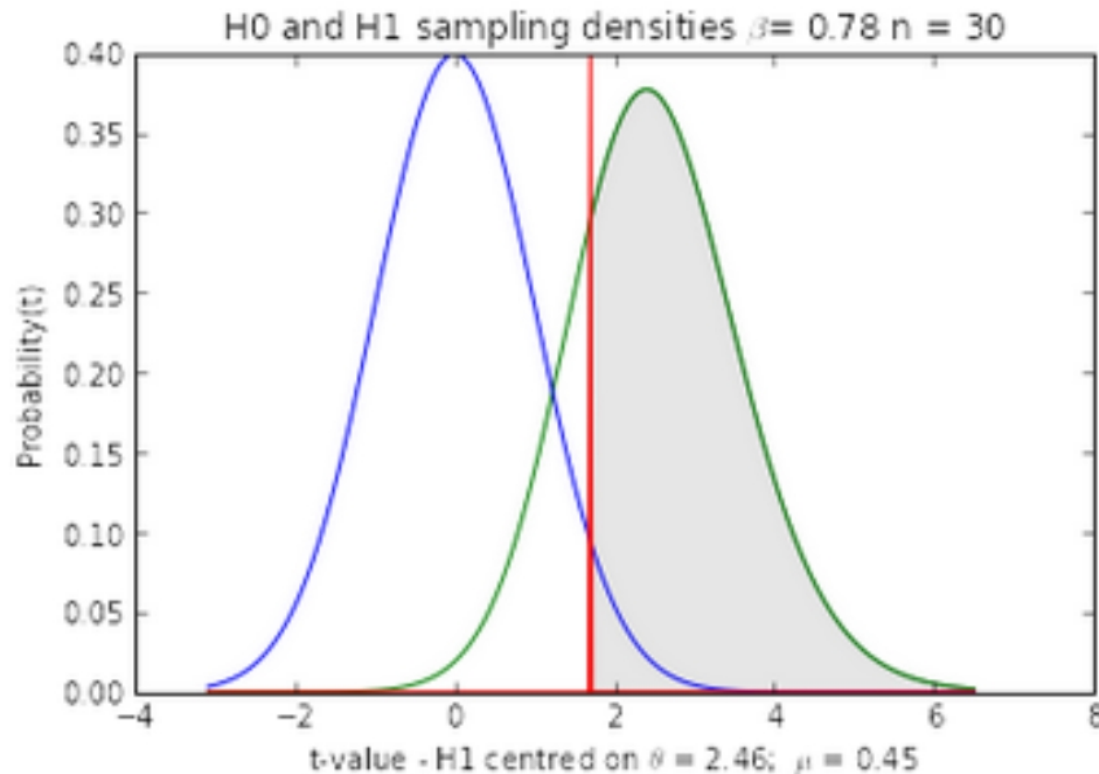
* **39%** of effects were rated to have replicated the original effect



- 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

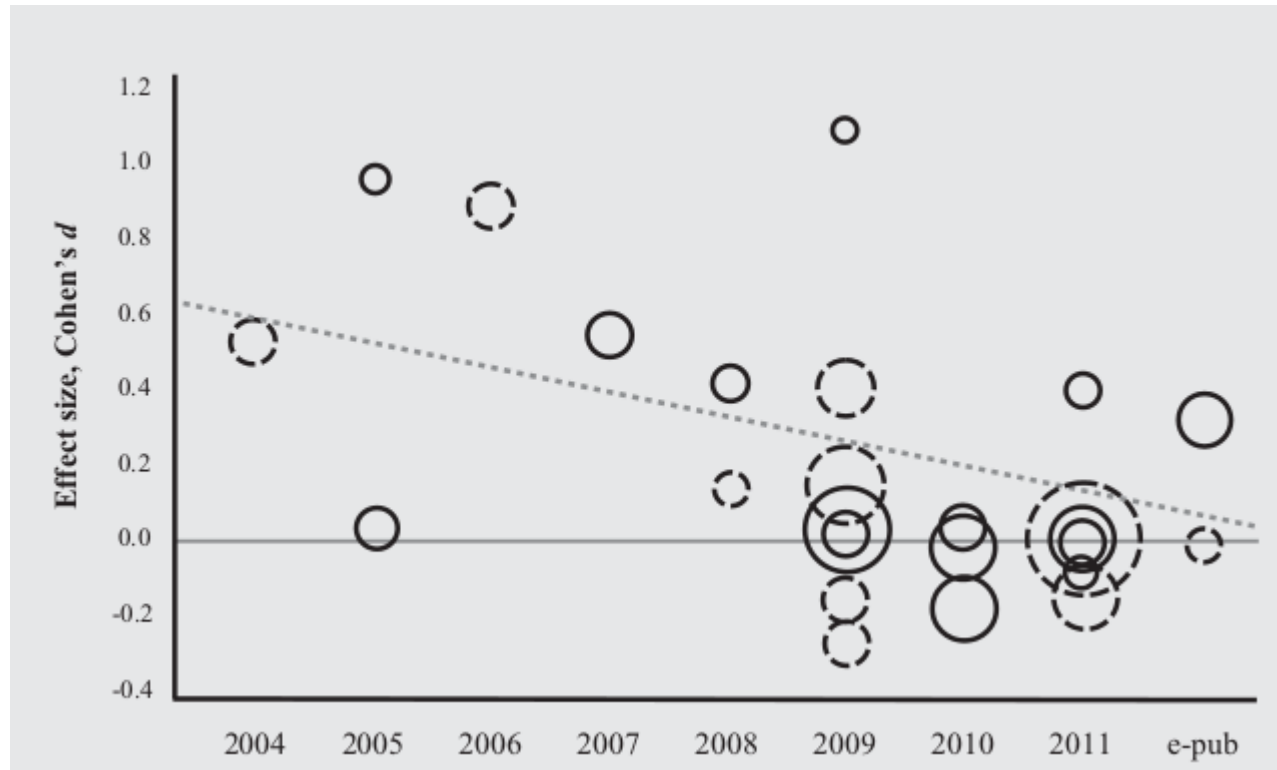
1. P-value and the null hypothesis statistical testing (NHST)
2. P-hacking
3. File drawer
3. Winner's curse
4. Effect sizes
- 5. Power**
6. PPV
7. Statistical generalizability

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



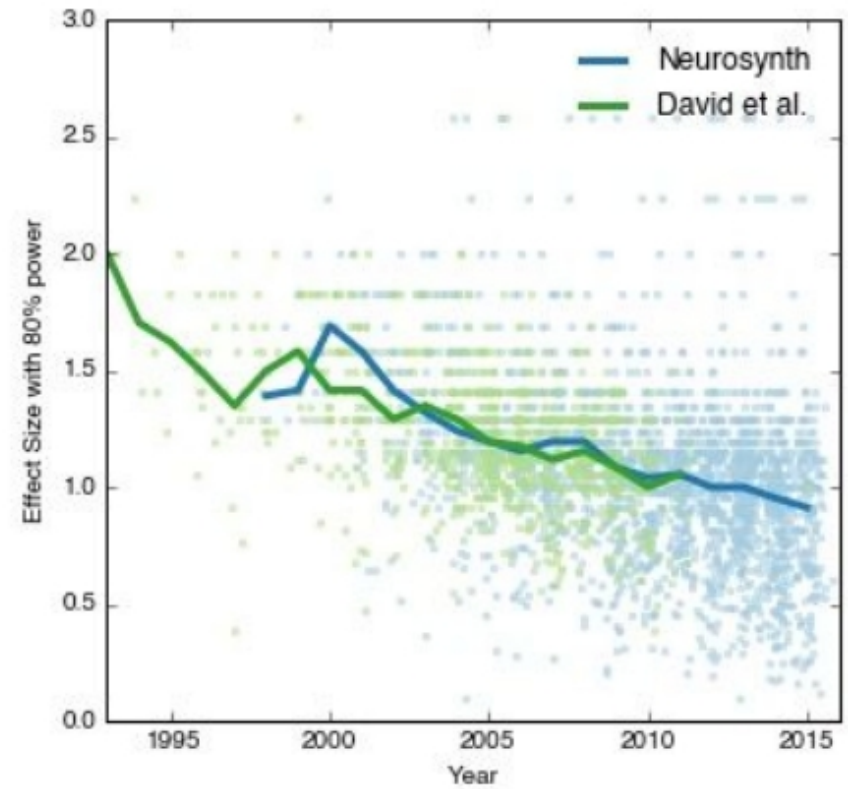
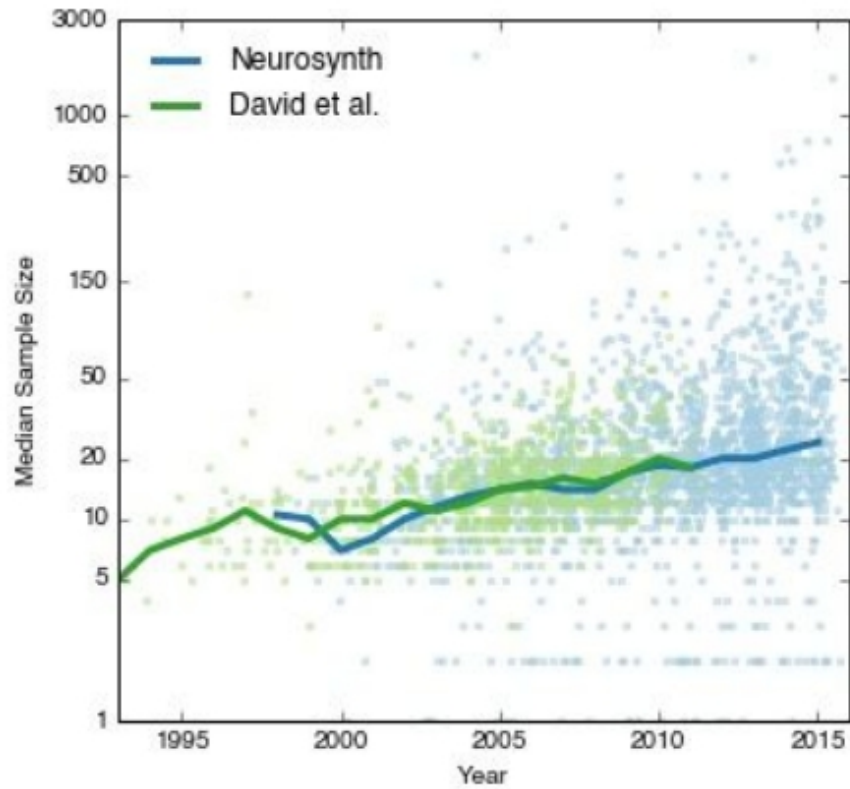
1. Power is a key measure because
 - Without good power, the study is not worth doing
 - Without good power, the study results are doubtful (see PPV)
2. Power is hard - or very hard to measure
 - You have to know H_0 , H_1 , effect size
3. Despite (2.), attempting to estimate power is important !

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



Molendijk, 2012: BDNF and hippocampal volume

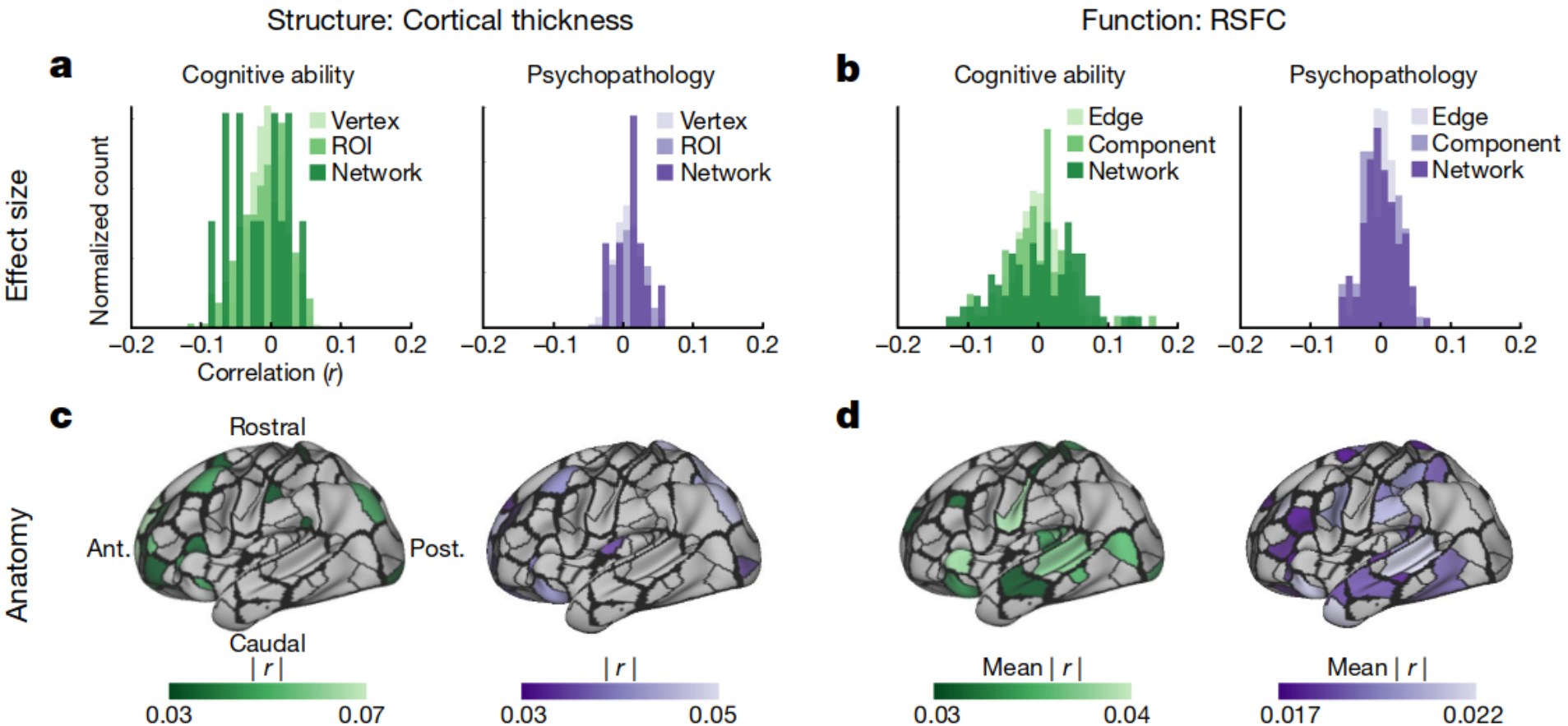
See also : Mier, 2009: COMT and DLPFC



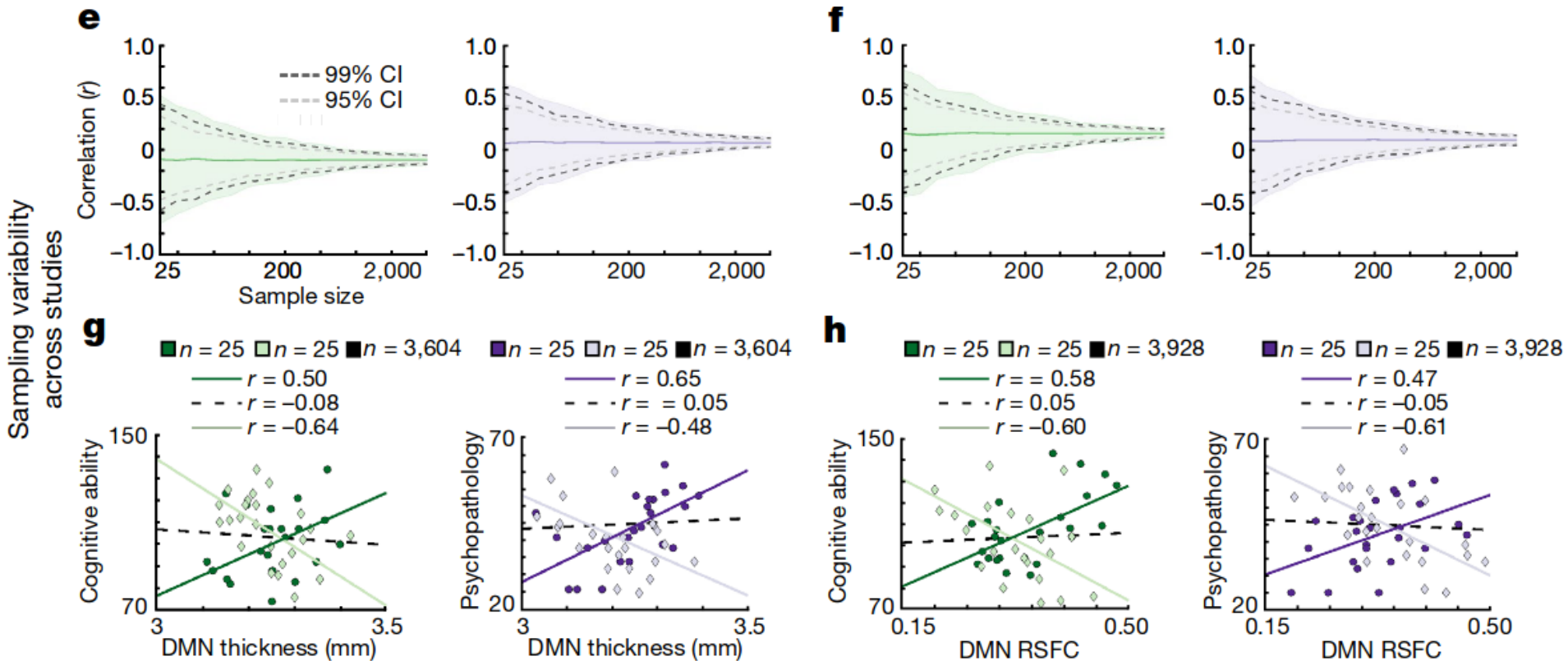
Poldrack et al., PNAS, 2016

Paradigm	Intersection mask	mask size (vox)	Cohen D			BOLD		
			P10	median	P90	P10	median	P90
MOTOR	Bilateral Precentral Gyrus	12894	0.158	0.628	1.070	0.505	2.707	8.582
	Bilateral Supplementary motor cortex	3418	0.211	0.716	1.197	0.911	4.033	12.510
	Left putamen	1532	0.114	0.513	0.864	0.586	2.388	4.318
	Right putamen	1437	-0.008	0.369	0.749	-0.045	1.696	3.609
WM	Bilateral Middle frontal gyrus	7116	0.101	0.474	0.837	0.130	0.986	2.504
EMOTION	Left amygdala	1133	0.265	0.534	1.065	0.516	1.198	3.379
	Right amygdala	1082	0.308	0.645	1.140	0.581	1.350	3.557
GAMBLING	Left accumbens	455	0.138	0.310	0.461	0.369	0.849	1.440
	Right accumbens	417	0.141	0.332	0.488	0.373	0.981	1.618

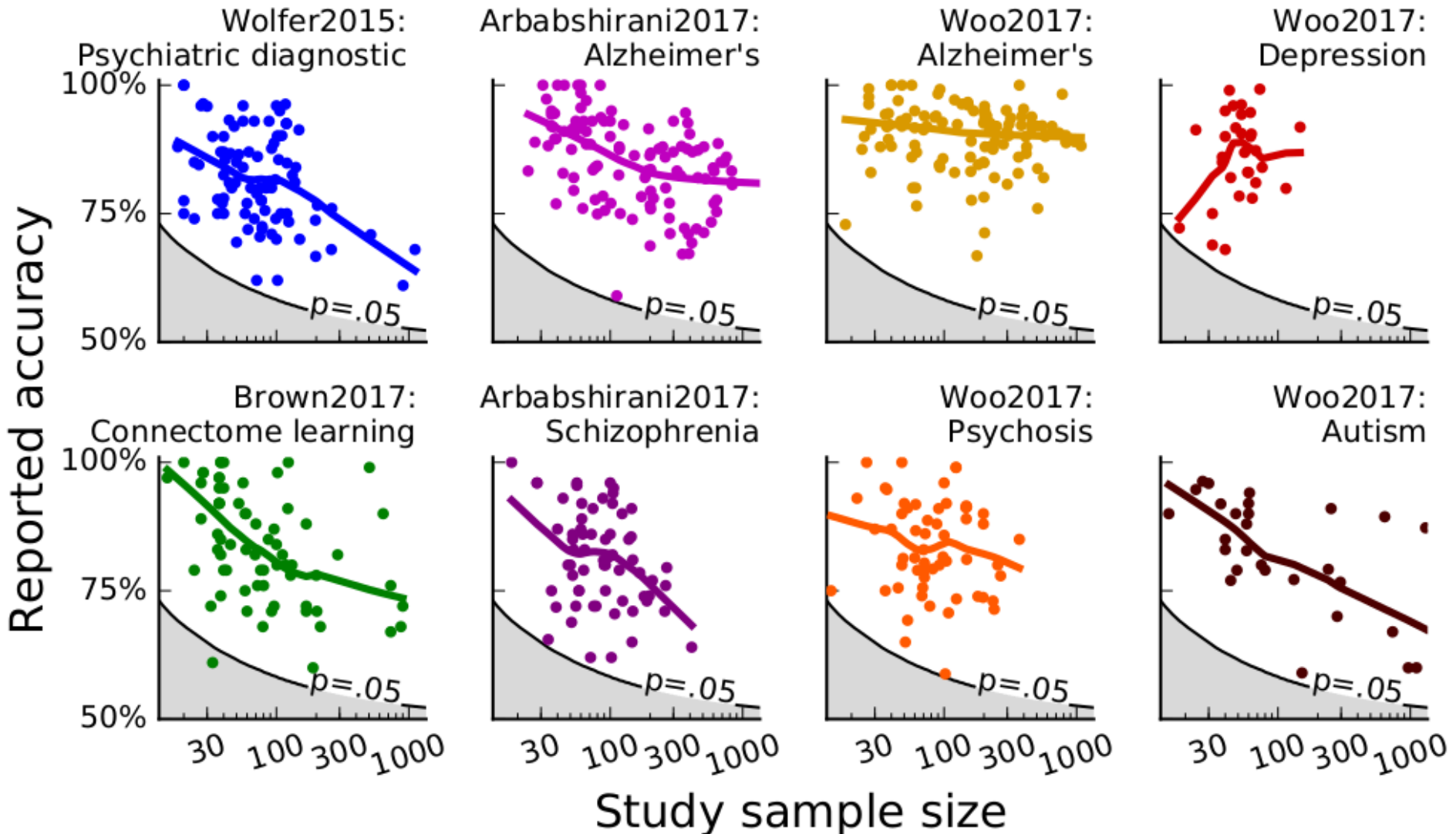
With effect size = 0.5 => Power ~ 30%



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



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



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

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

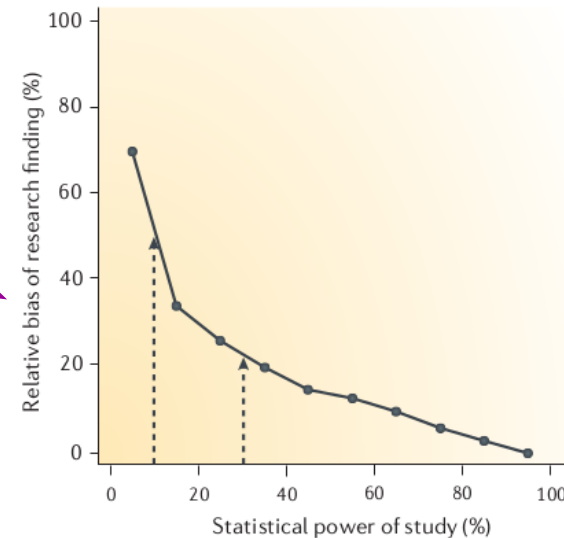
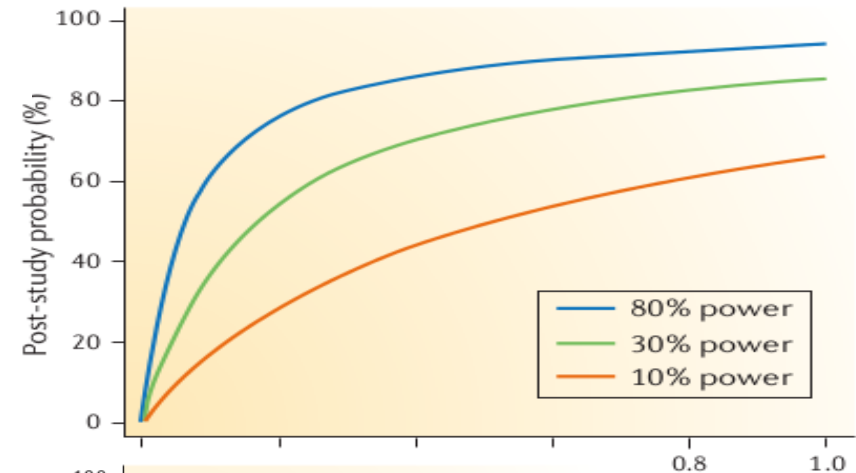
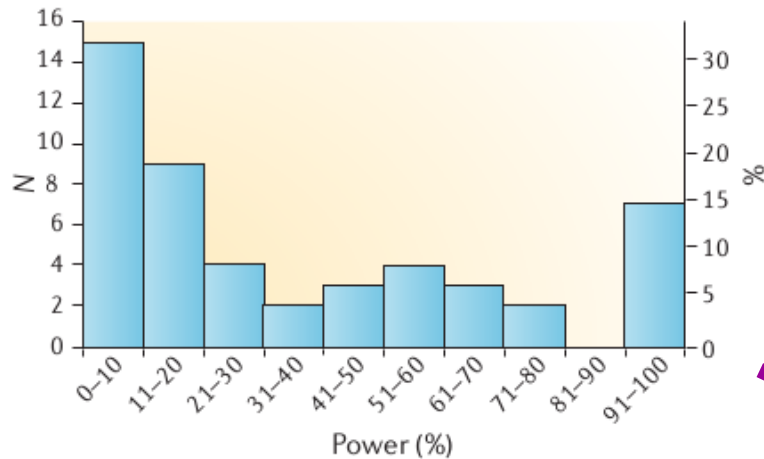
[Note :Requires that “hypothesis” has a probability]

Open access, freely available online

Essay

Why Most Published Research Findings Are False

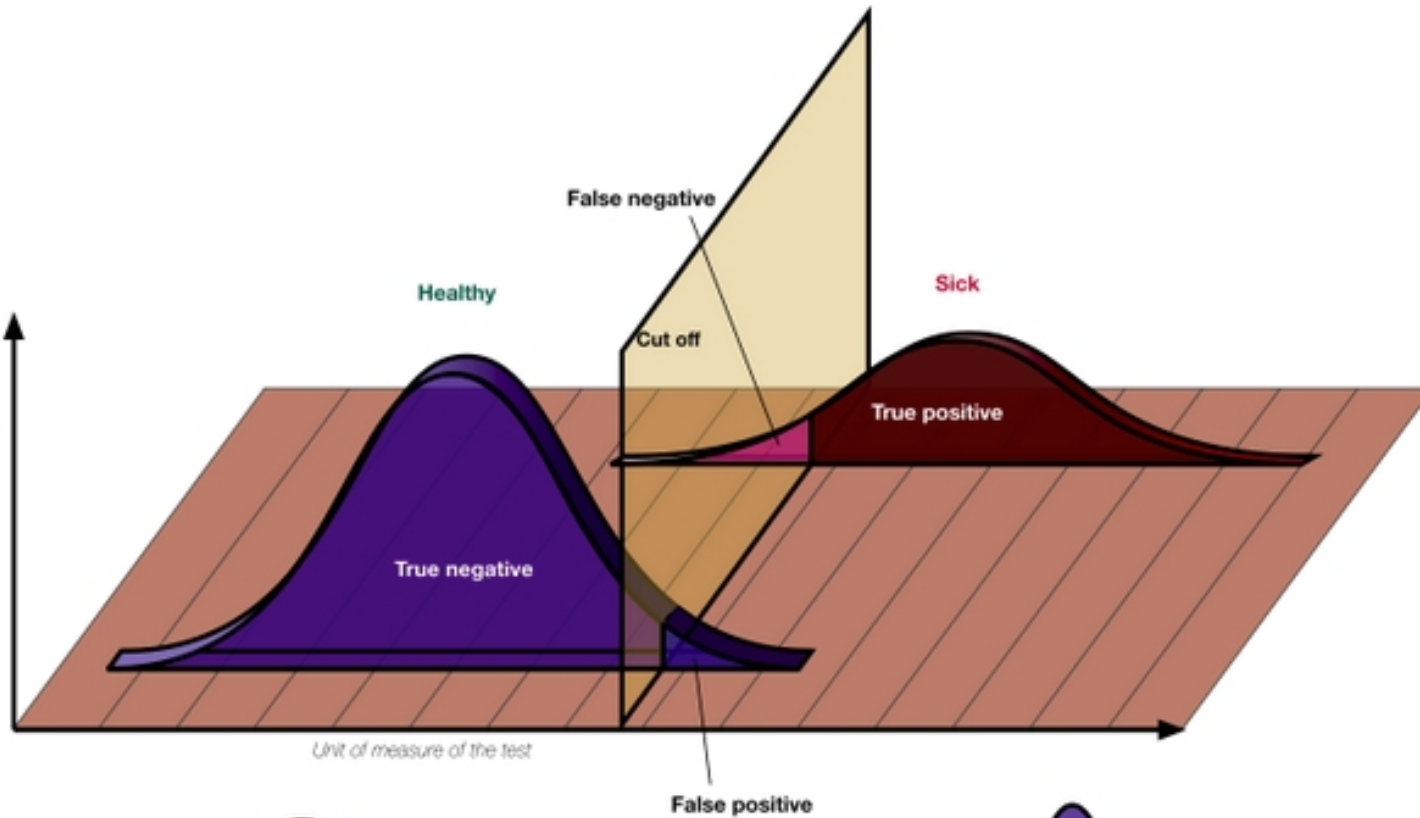
John P.A. Ioannidis



Button et al., NNR, 2013

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

$$\text{PPV} = \frac{WR}{WR + \alpha} = \frac{WP_1}{WP_1 + \alpha P_0} = \frac{TP}{TP + FP}$$



$$PPV = \frac{\text{True positive}}{\text{True positive} + \text{False positive}}$$

$$NPV = \frac{\text{True negative}}{\text{True negative} + \text{False negative}}$$

$$\text{Sensitivity} = \frac{\text{True positive}}{\text{True positive} + \text{False negative}}$$

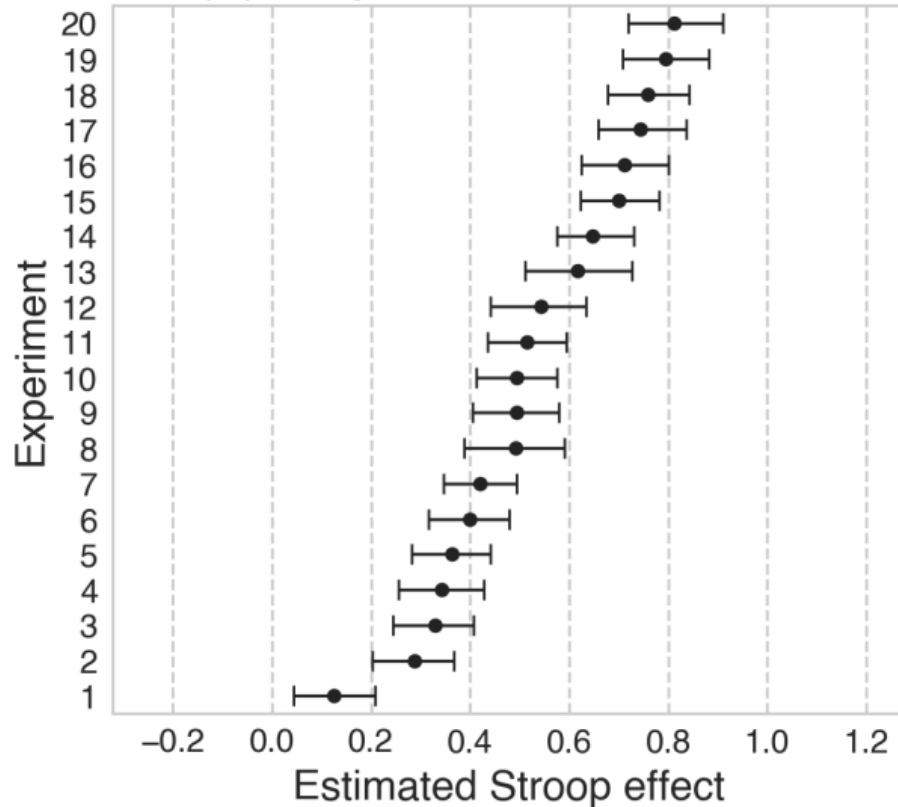
$$\text{Specificity} = \frac{\text{True negative}}{\text{True negative} + \text{False positive}}$$

- 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

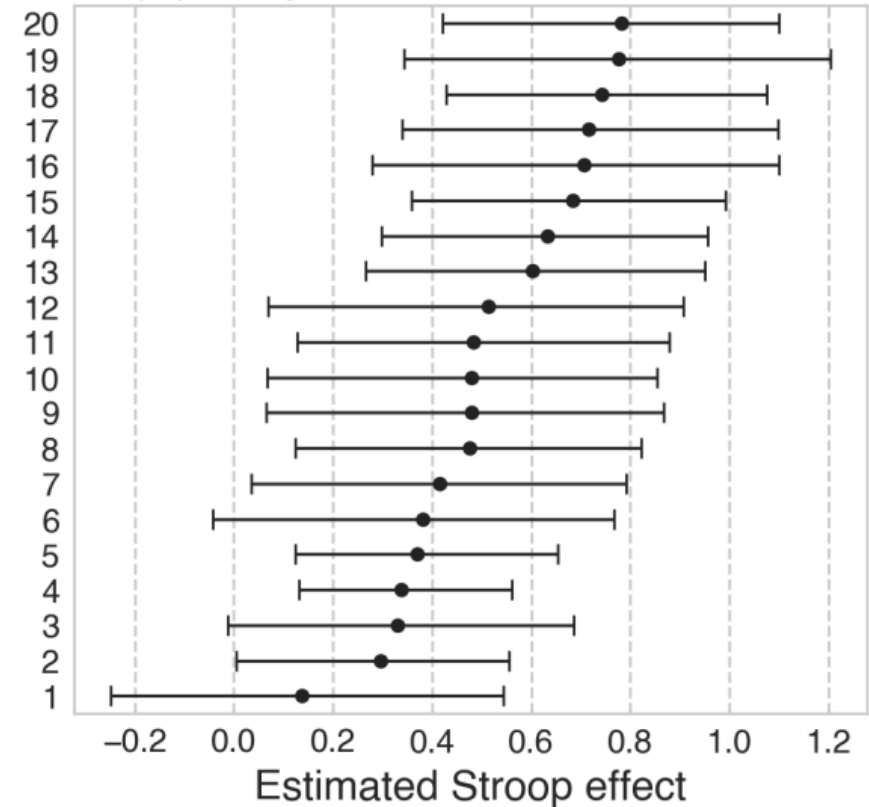
1. P-value and the null hypothesis statistical testing (NHST)
2. P-hacking
3. File drawer
3. Winner's curse
4. Effect sizes
5. Power
6. Positive Predictive Value
- 7. Statistical generalizability**

- 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
 - $Y = X\mathbf{\underline{b}} + Z\mathbf{\underline{g}} + \mathbf{e}$
 - Random: consider another source of variance
 - Fixed: one source of randomness
 - Example: several observation per groups, many groups, linear regression where slope and intercepts vary with groups

(A) Subjects modeled as fixed



(B) Subjects modeled as random



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.

- It is critical to understand the statistical framework that we are working with - often the NHST framework
- Results in the literature may be biased due to a number of effects :
 - File drawer
 - Winner's curse
 - P-hacking
 - Poor statistical modelling

- Some literature is showing that TBI is inducing decrease in connectivity compared to a normal control group
- To “advance science” I will test / reject:
 - $\text{TBI-connectivity} == \text{NC-control connectivity}$
- But the theory is **not** that TBI and NC group have the same connectivity
- I should try to reject that:
 - $\text{TBI-connectivity} \neq \text{NC-control connectivity}$
 - In what situation the connectivity is unchanged ?
 - What is the effect size and SD of the change ?

- Are general hypotheses better than more precise ones ?
 - If you want to learn a lot, you should reject large hypothesis (ie, the “null” should be “large”)
- In general, our nulls are very narrow, hence we learn “little”
- Better hypothesis : the association is in the $[-.3, .3]$ interval - let's reject that :)

- **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