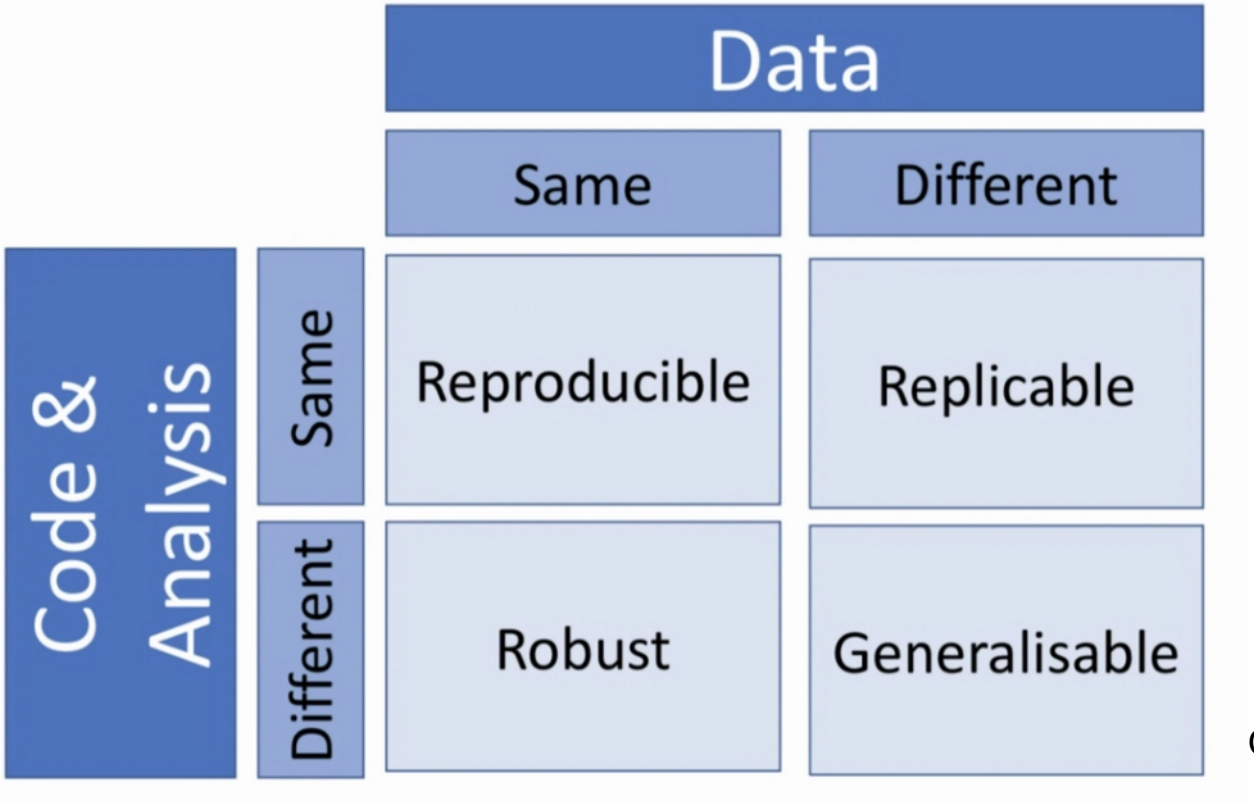
Classical statistics pitfalls and remedies

Jean-Baptiste Poline

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

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



Credit: J.Pineau

- 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
- P-value 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 "statistics" ?
  - Any function of the data
    - the mean,
    - the SD,
    - the mean/SD,
    - the t statistics,
    - the z statistics,
    - etc

Under the null hypothesis

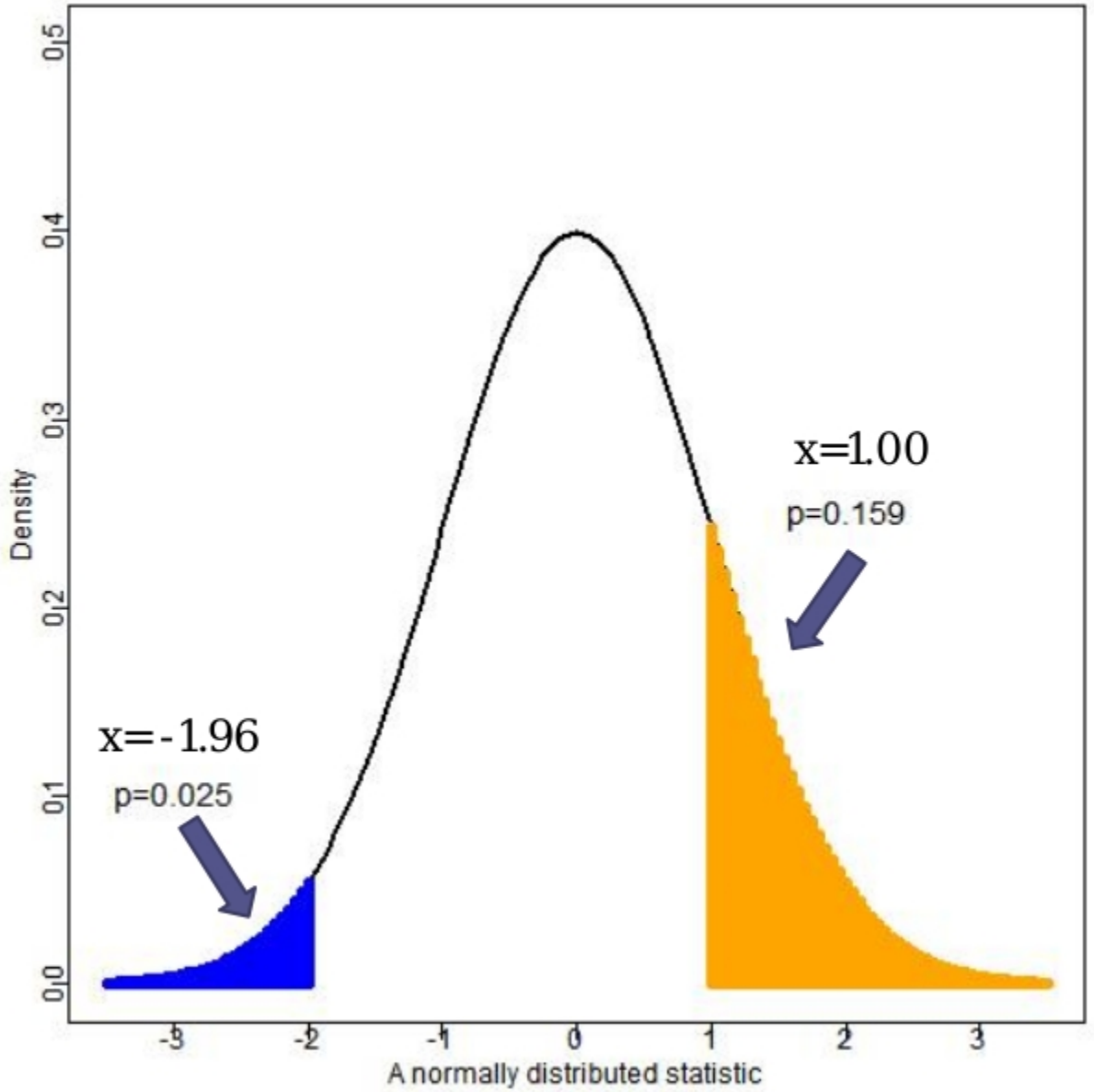
- Knowledge of the null hypothesis
- Choice of a statistic
- Concept of repeating the whole study in the same way
    - Same study design
    - Same sampling scheme
    - Same definition of the statistics
    - Same population sampled

- "*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.

Examples of null distributions

A normally distributed statistic

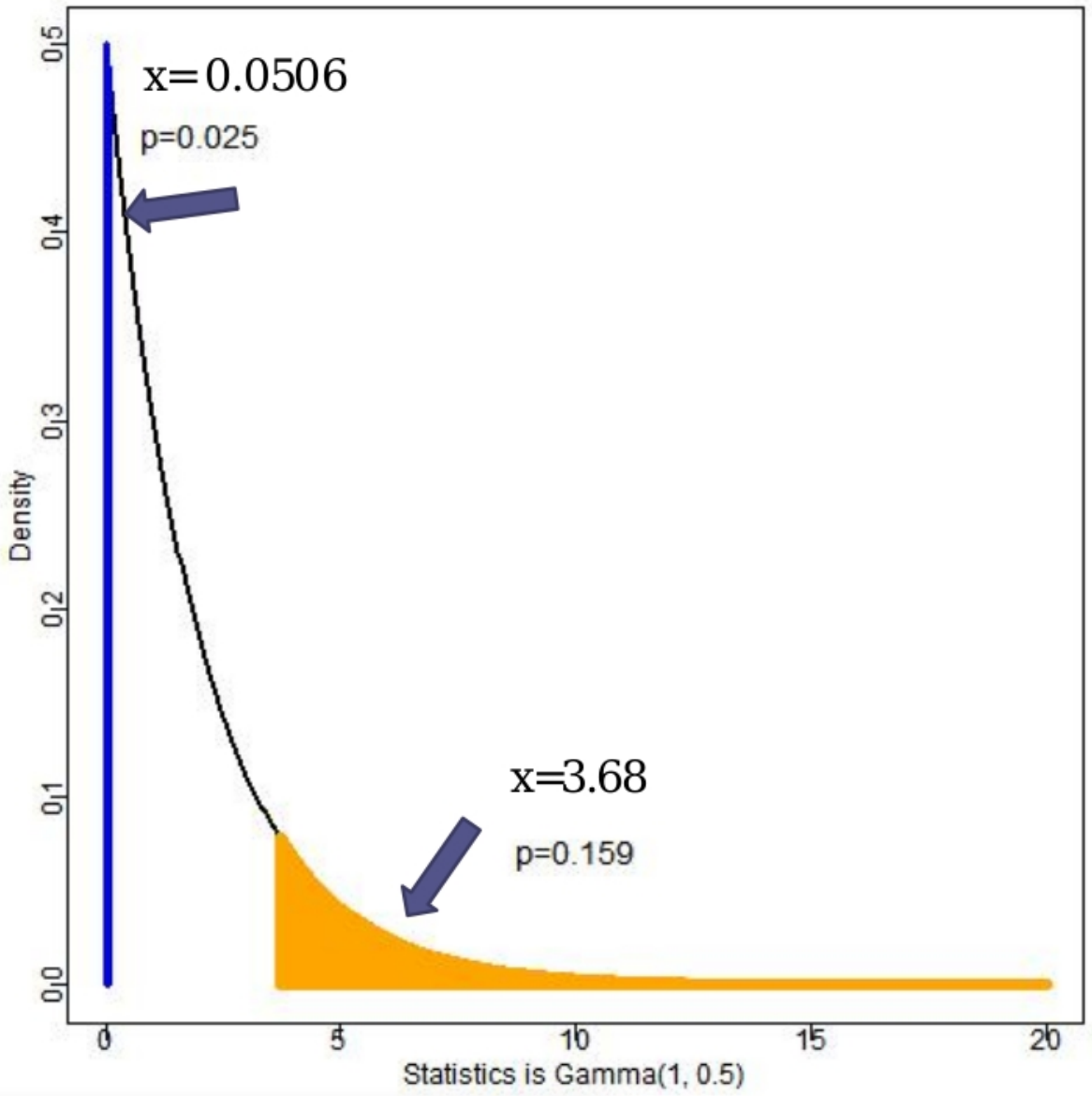


Credit: Celia Greenwood McGill

Gamma

A gamma-distributed statistic

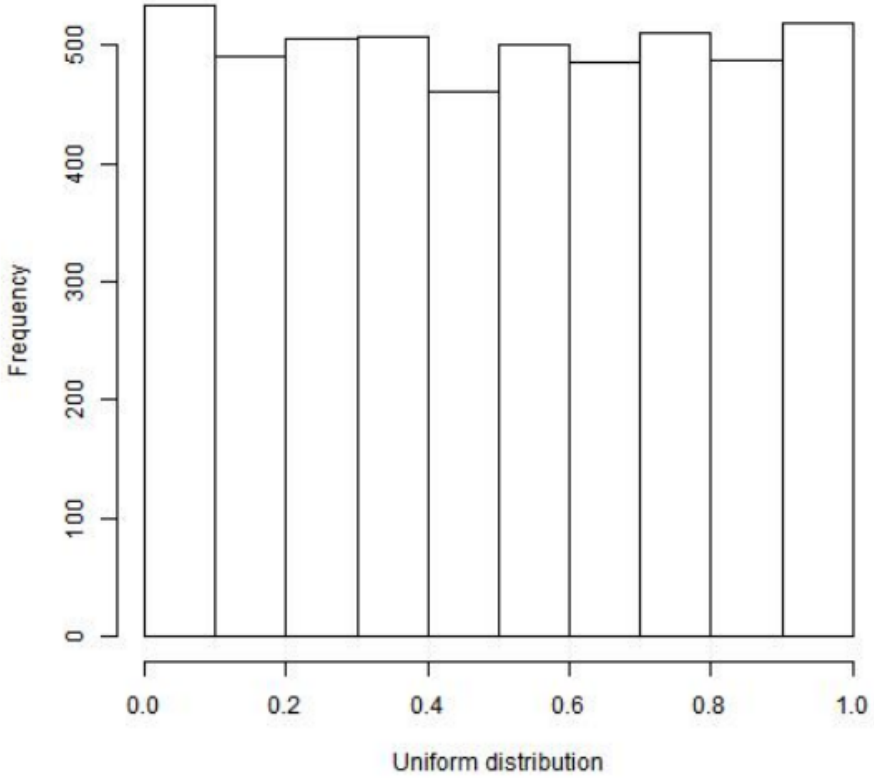
Gamma(1, 0.5)



Credit: Celia Greenwood McGill

Uniform distribution

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



Why p-value are uniform under the null ?

- We want the probability of rejecting the null to be alpha : P(*observed-statistics* > *quantile-*α) = α
- 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* <= *x*) = *x*

Credit: Jerôme Dockes

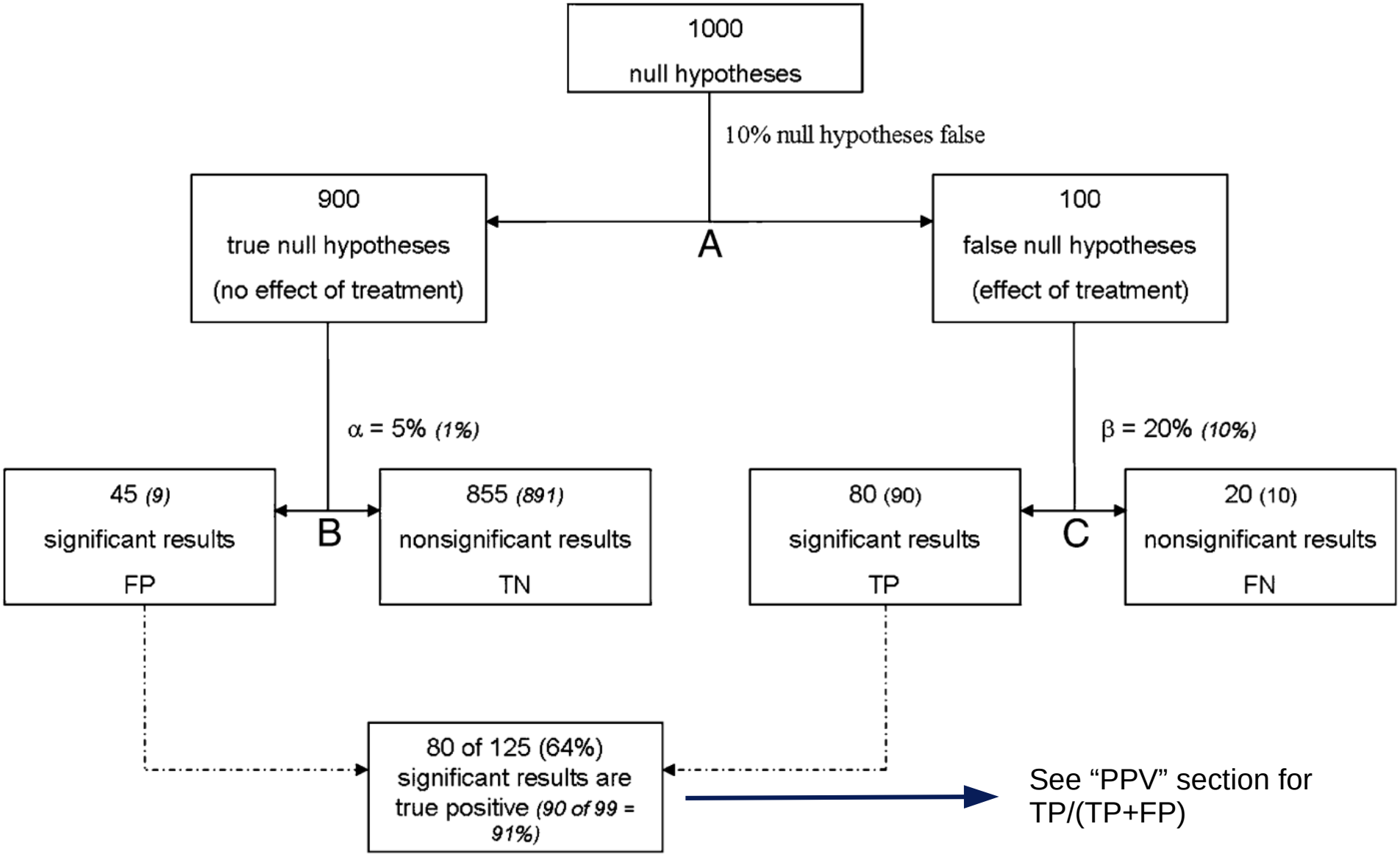
- A p-value is a statistic : it is a function of the data - therefore random !
  - Fact used in the “p-hacking test”

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.

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

- A long history - 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 !
  - 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

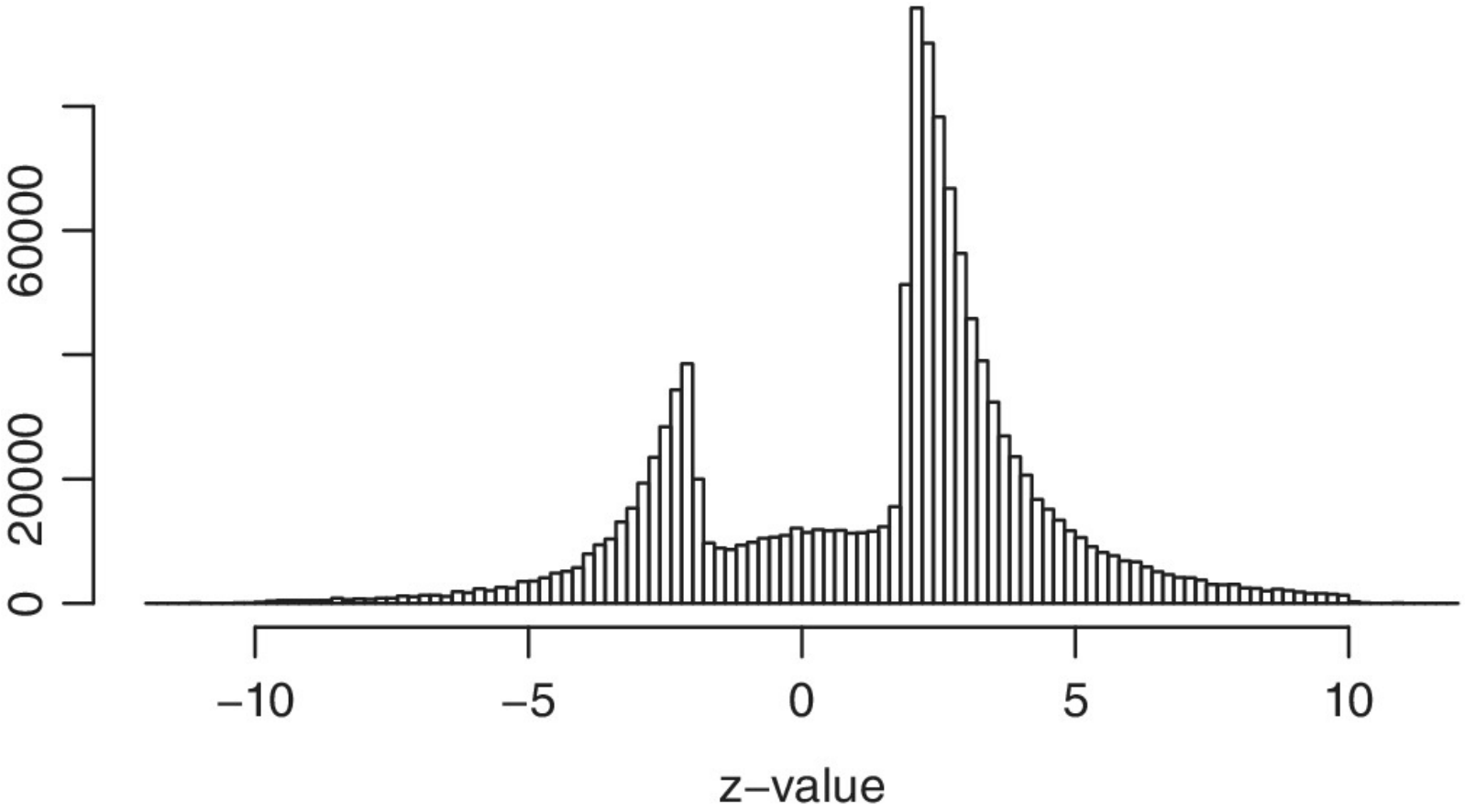
For any given research area, one cannot tell how many studies have been conducted but never reported. The extreme view of the "file drawer problem" is that journals are filled with the 5% of the studies that show Type I errors, while the file drawers are filled with the 95% of the studies that show non-significant 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 !
  - 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

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

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

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

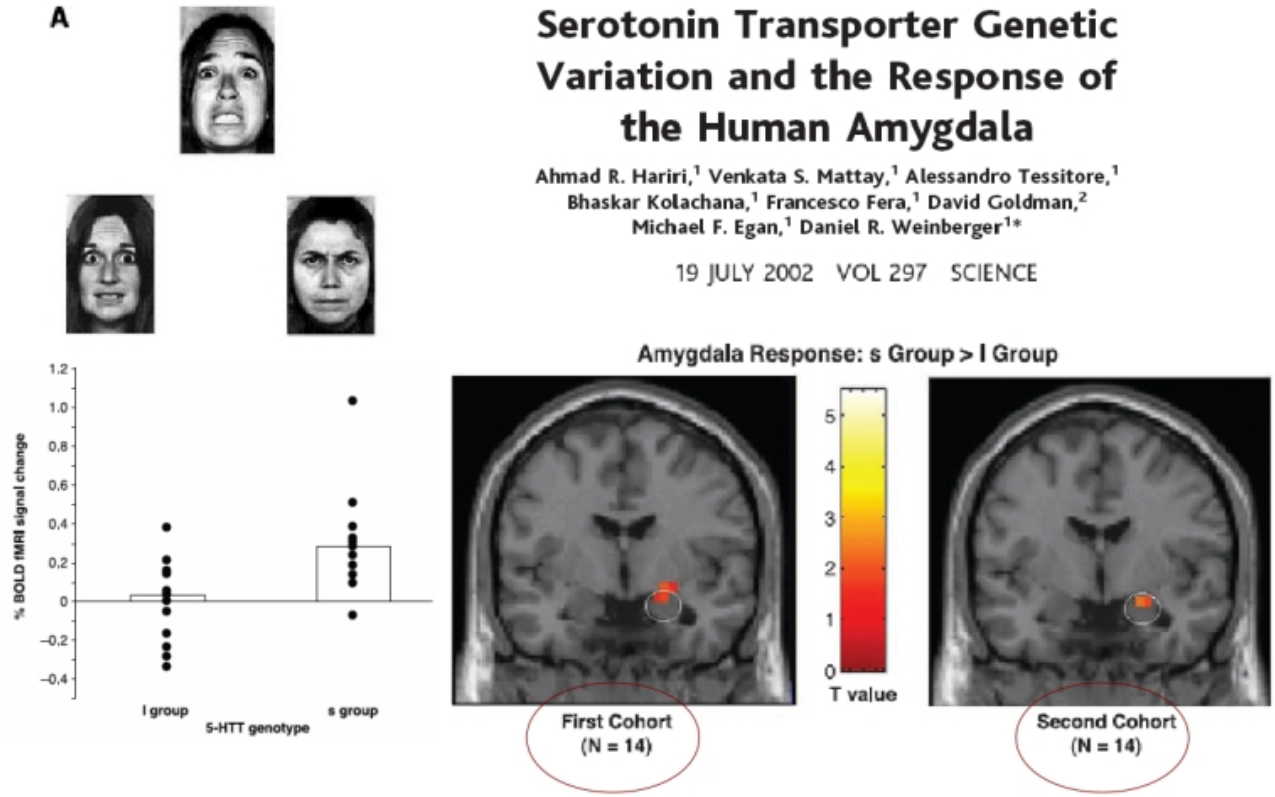
Effect size ?

- 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 ?

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”

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: $Y = [1 \ldots 1]^t [m_1 - m_2] + \mathrm{residual}$
- 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

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 ?

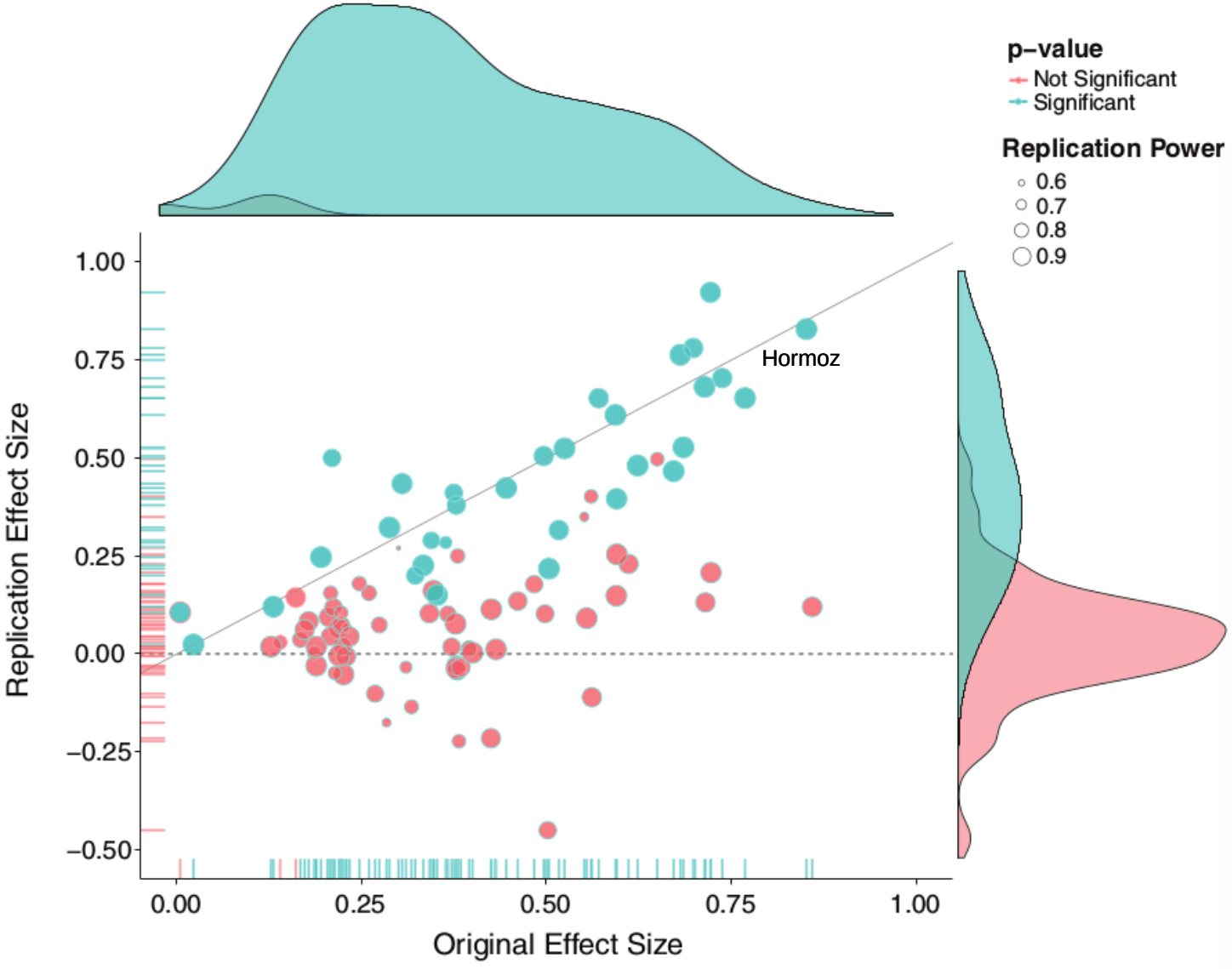
First, compute the standard deviation of the data from the SDM

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

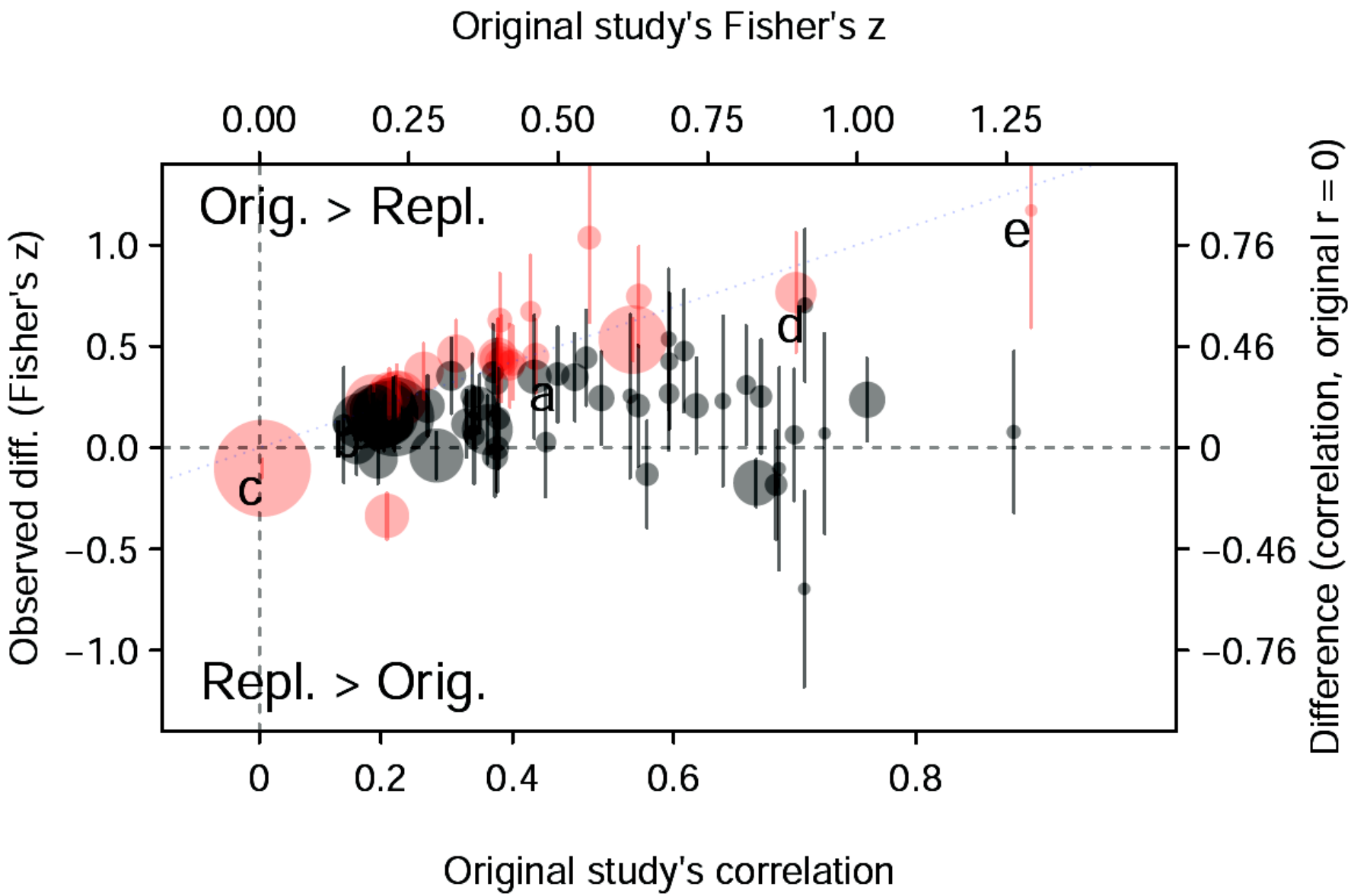
B. Nosek, Estimating the reproducibility of psychological science, Science 2015



* 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”

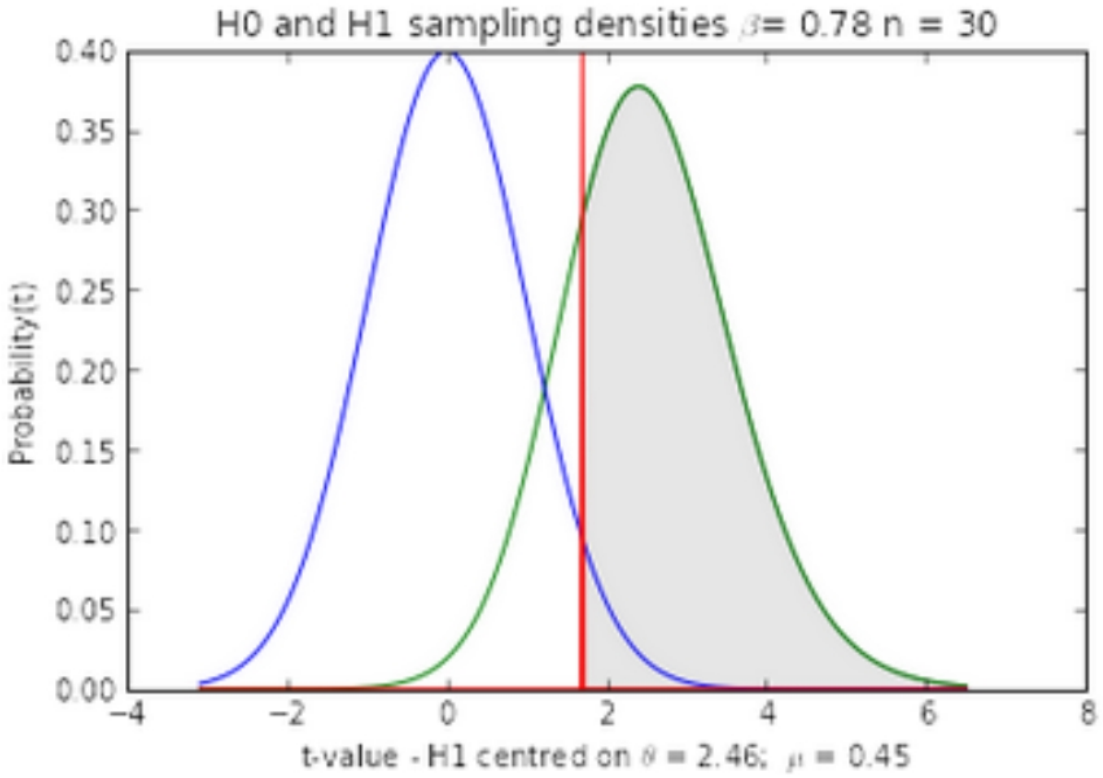


Morey and Lakens, 2016

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

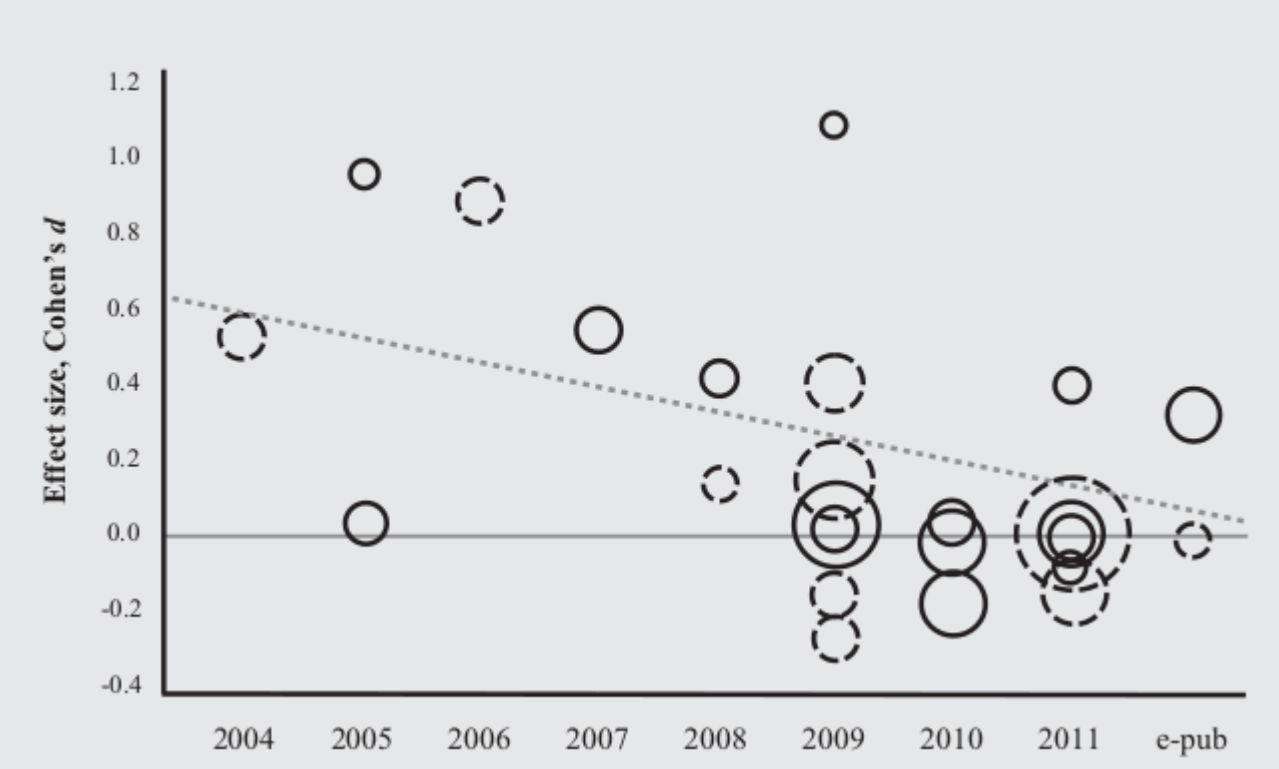
Decision/H	H0 True	H1 True
reject	α (type I)	$1 - \beta$ (Power)
not reject	1 - α	β (type II)



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 !

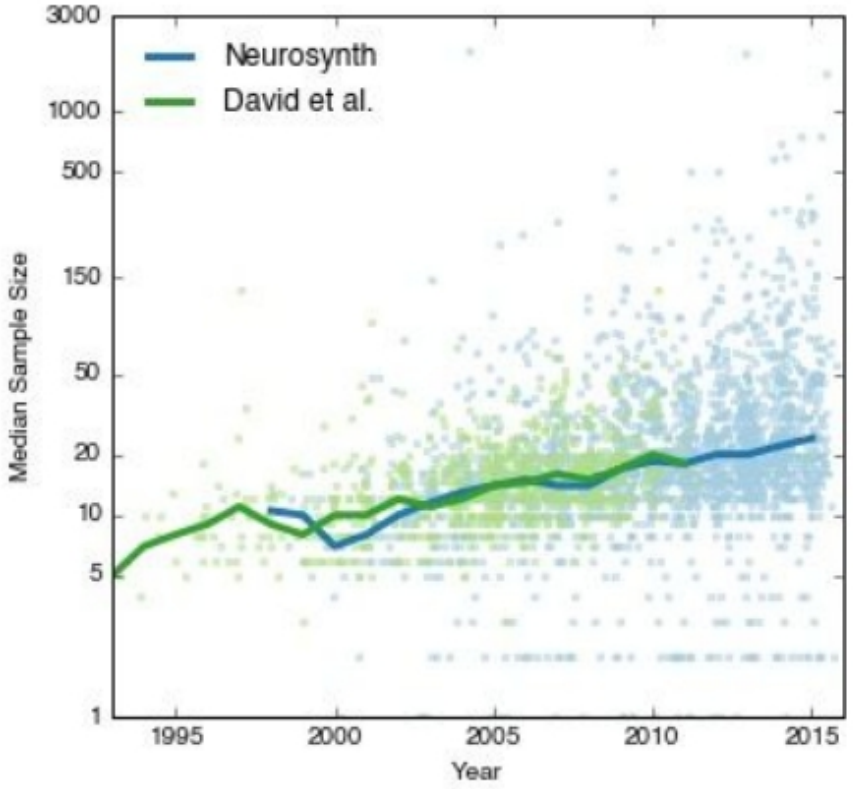
Low power - high effect size

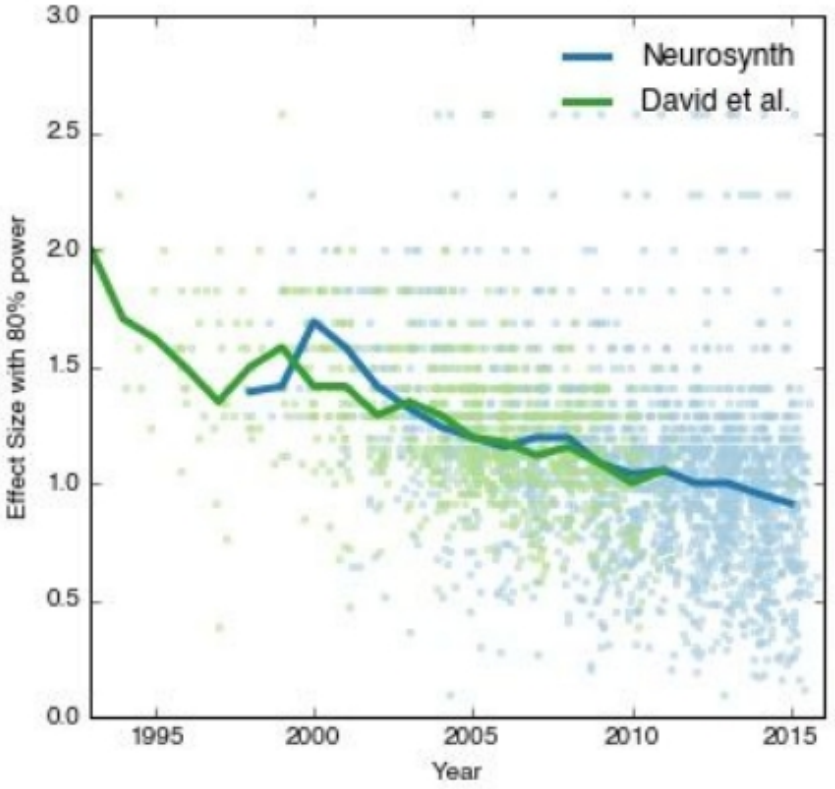


Molendijk, 2012: BDNF and hippocampal volume

See also : Mier, 2009: COMT and DLPFC

Feeling the Future





Poldrack et al., PNAS, 2016

Feeling the Future

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%

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 !

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)

Recall Bayes

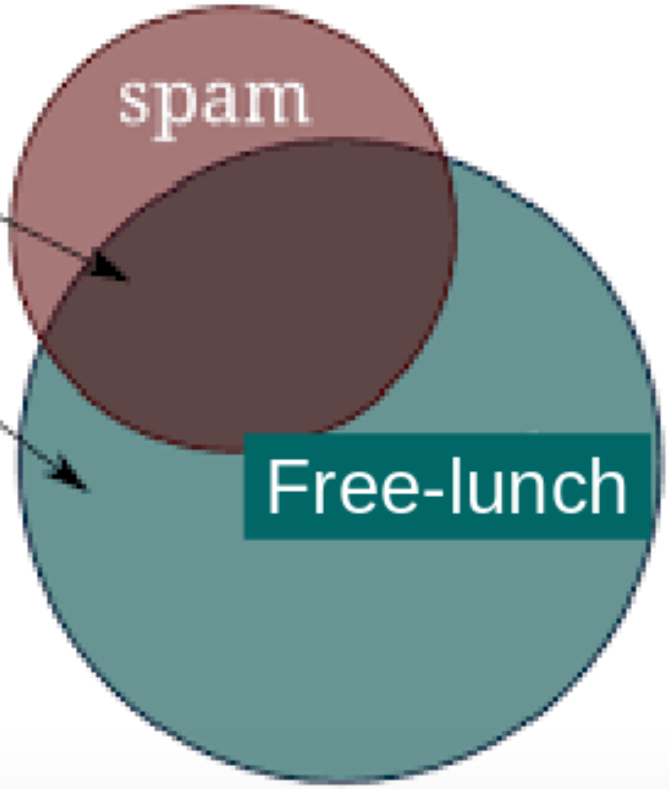
$$P(H, D) = P(H|D)P(D)$$

$$P(H, D) = P(D|H)P(H)$$

$$P(H|D) = \frac{P(D, H)}{P(D)} = \frac{P(D|H)P(H)}{P(D)}$$

You've received an email promising a free lunch

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$



A is the set of **spam emails**,
B is the set of "**free lunch**" emails

What is the probability that I have a spam knowing that it is labeled as "free lunch" ?

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/

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

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

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

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

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

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

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

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

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

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

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

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

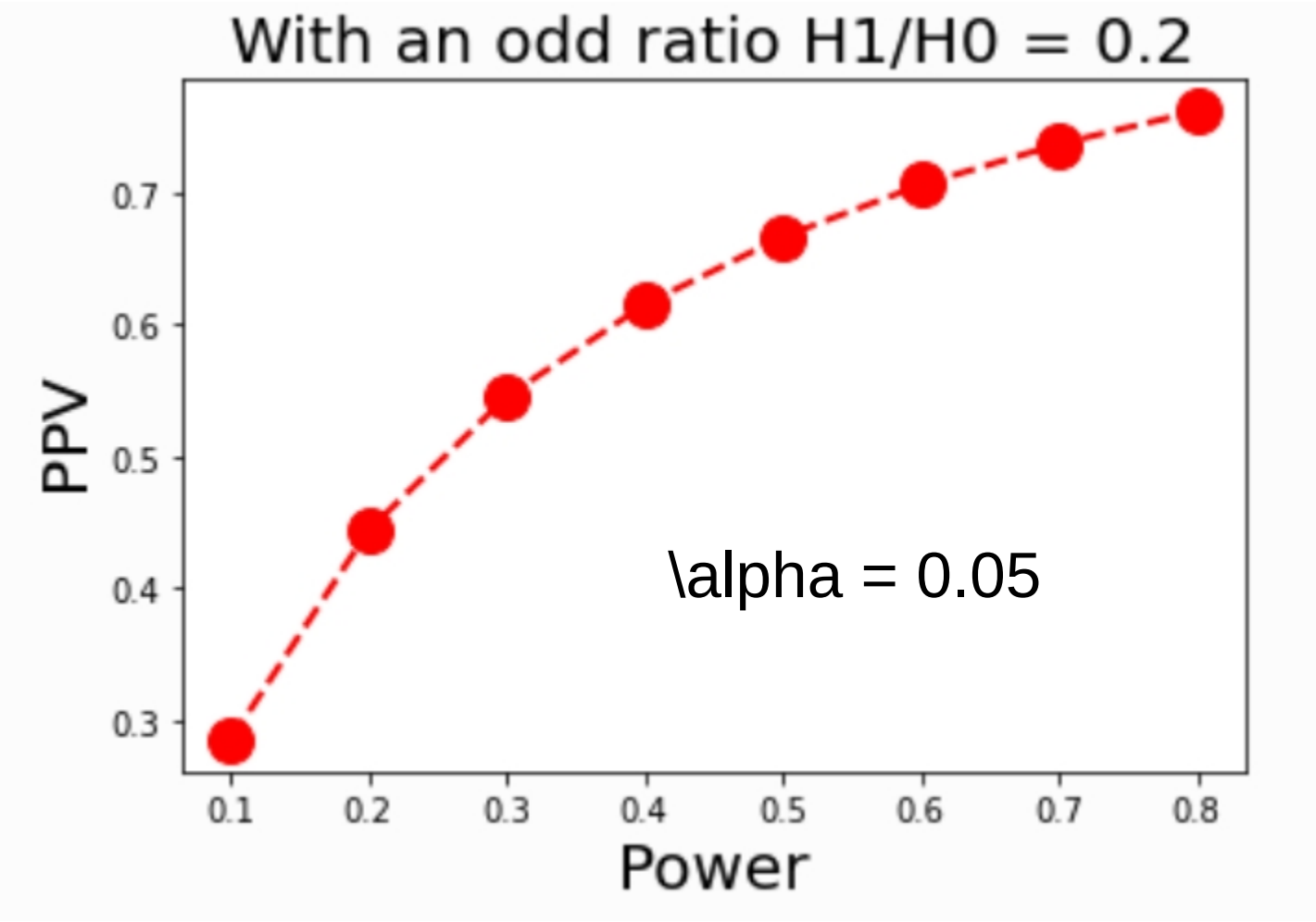
$$= \frac{P(T_S \mid H_A)P(H_A)}{P(T_S \mid H_A)Pr(H_A) + Pr(T_S \mid H_0)Pr(H_0)}$$

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

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

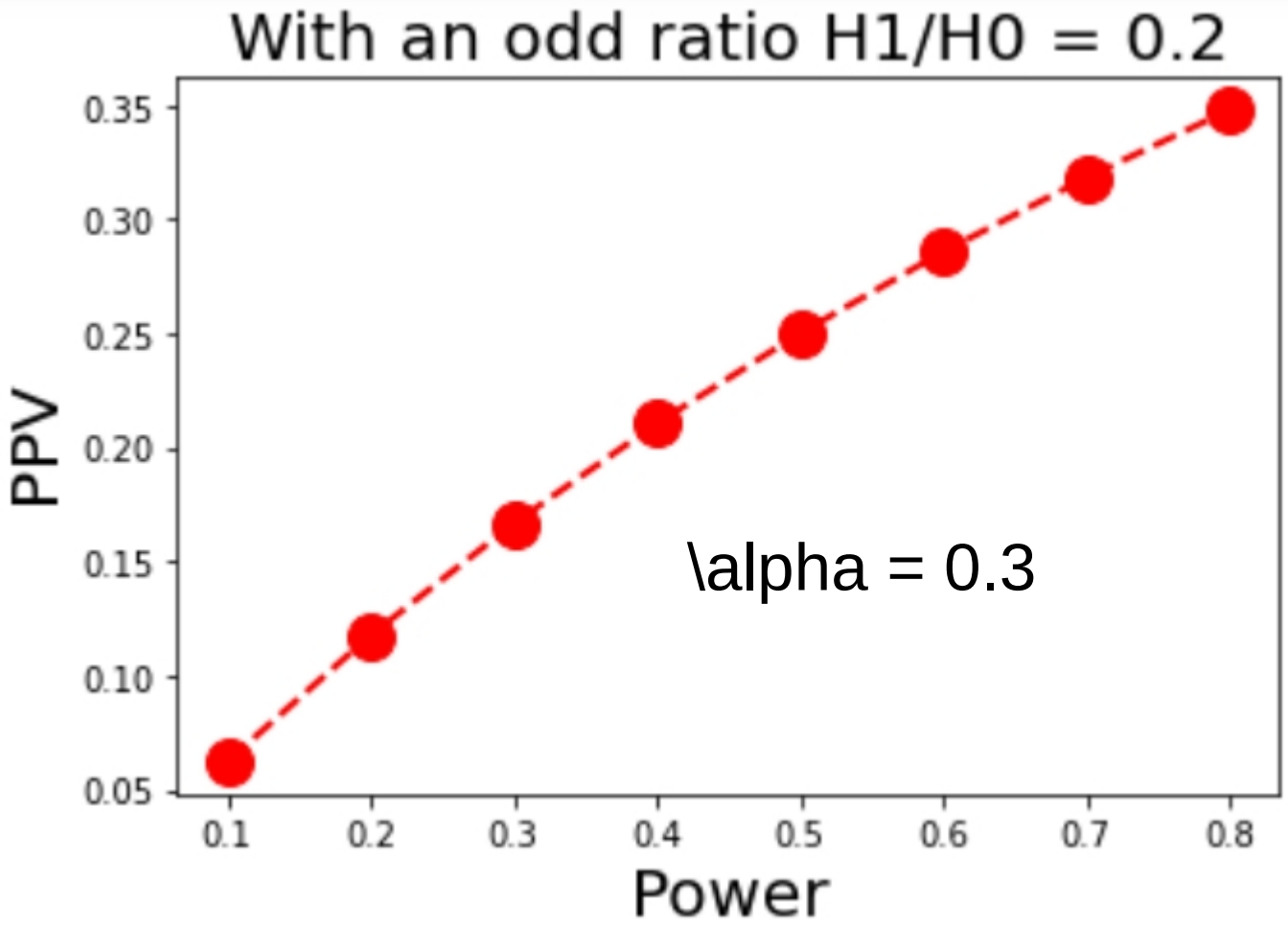
Deriving PPV

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

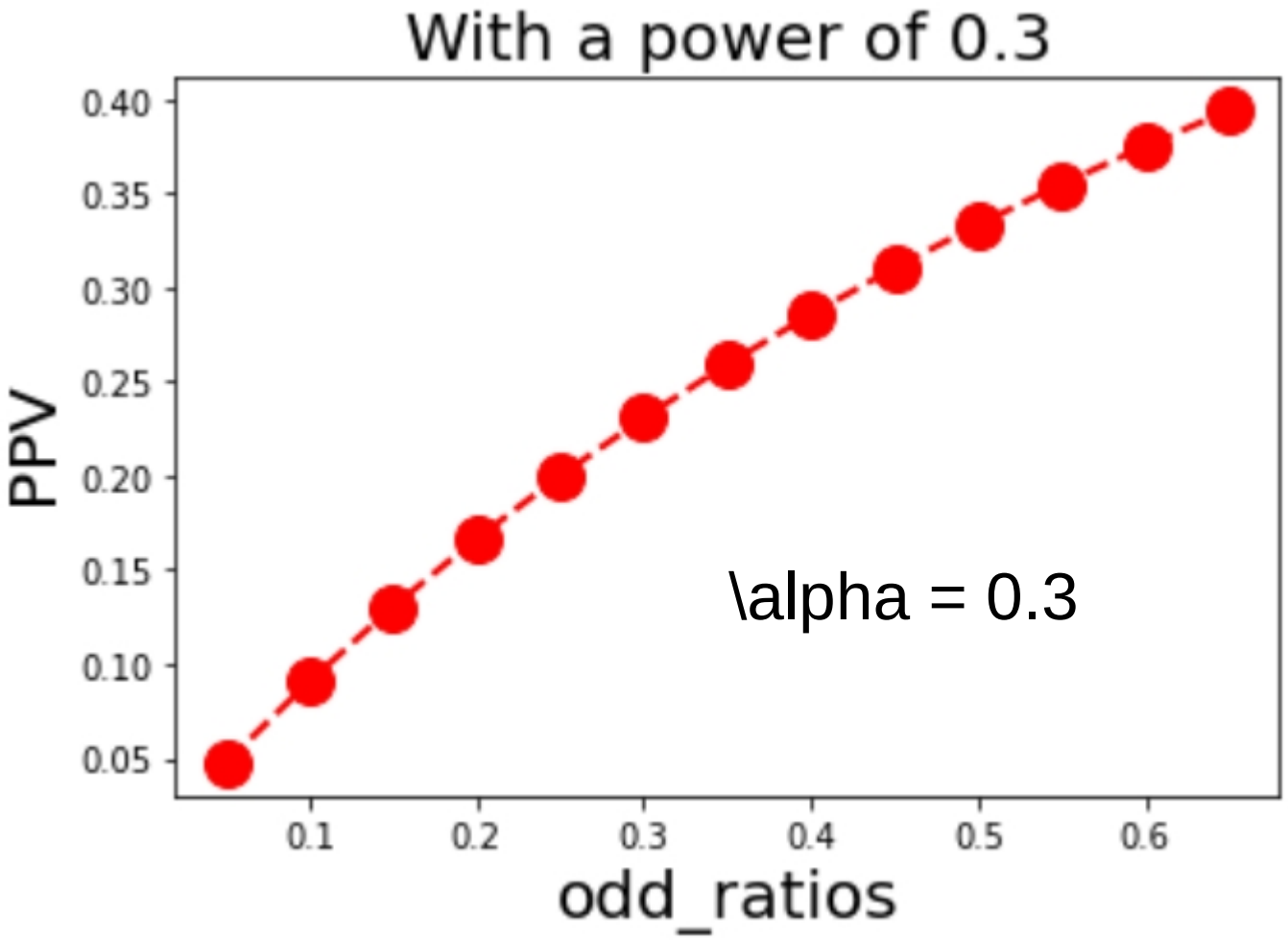


Deriving PPV

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



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



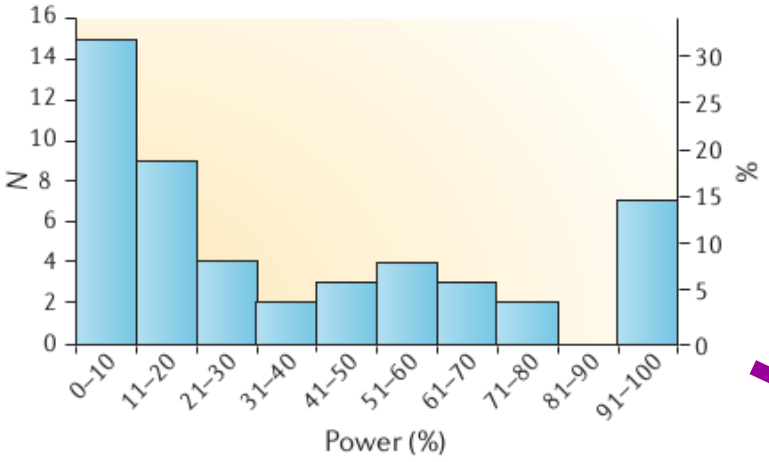
Power issues lead to PPV issues

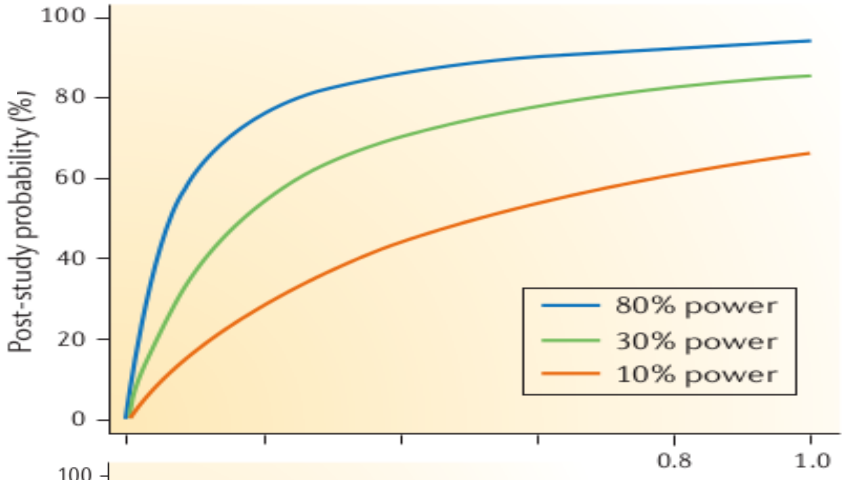
Open access, freely available online

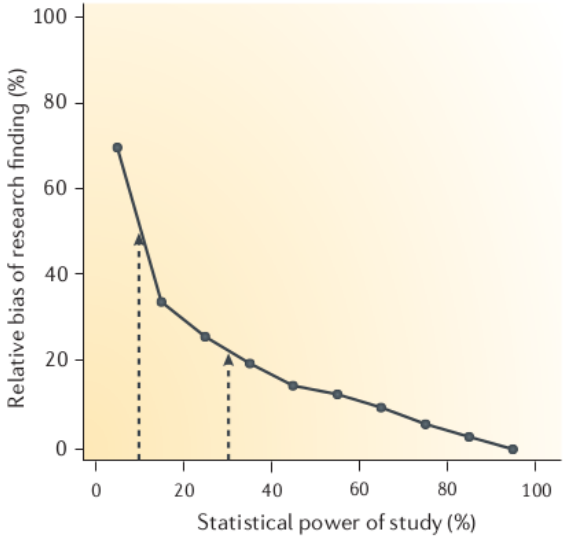
Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis







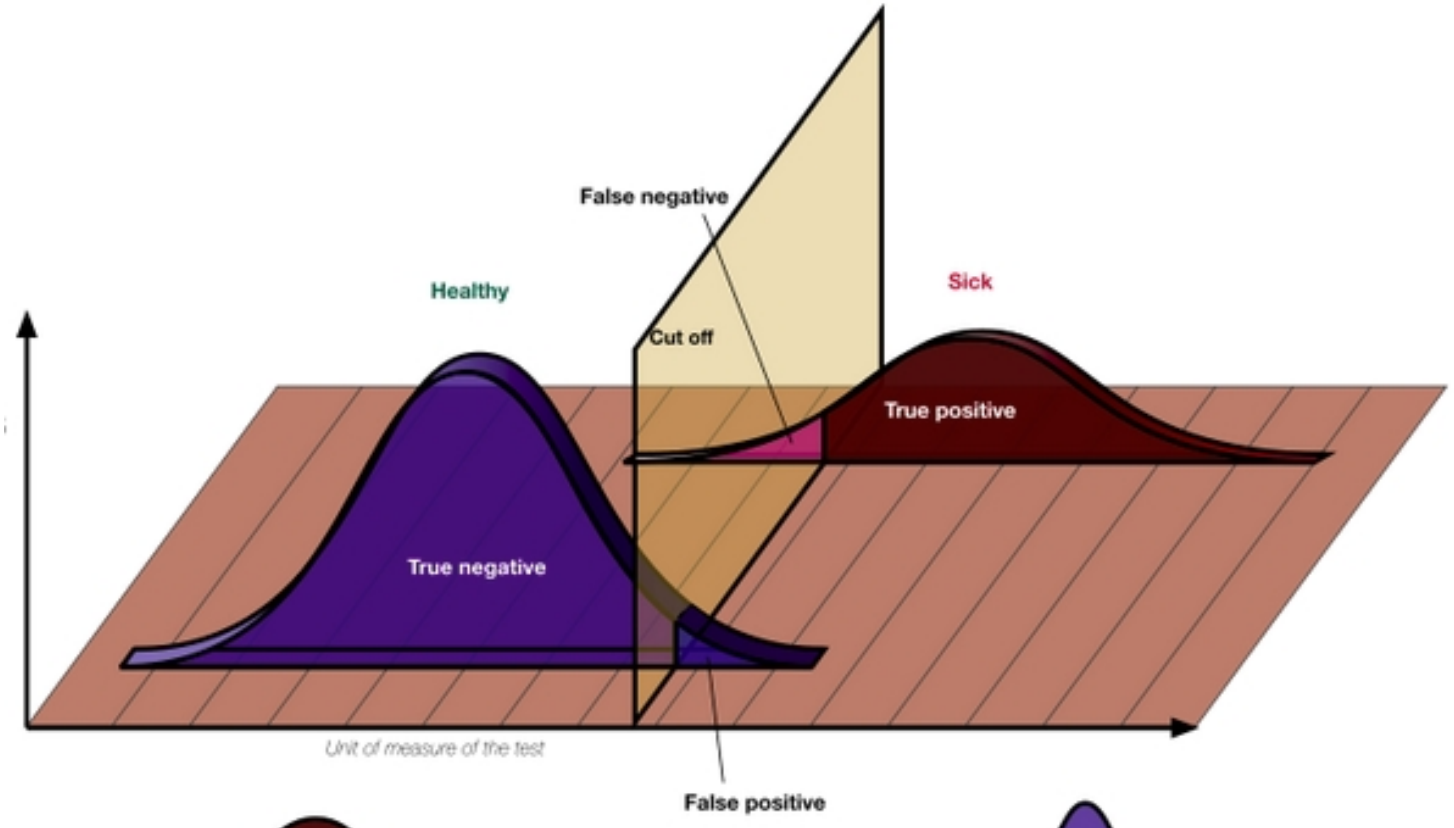
Button et al., NNR, 2013

PPV is also “precision” in ML

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

$$\text{PPV} = \frac{WR}{WR + \alpha} = \frac{W\text{P}_1}{W\text{P}_1 + \alpha\text{P}_0} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

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

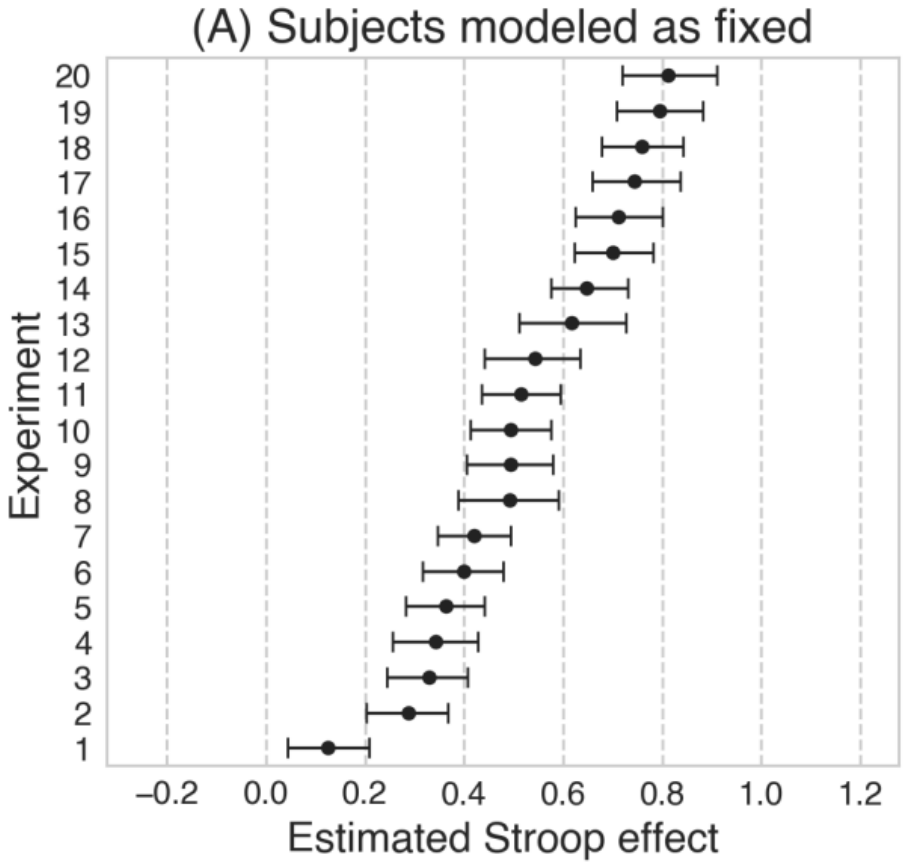
- These are usually unknown, but can be estimated
  - The process of estimating this quantities helps assess the solidity of the result
  - PPV may be in general quite small

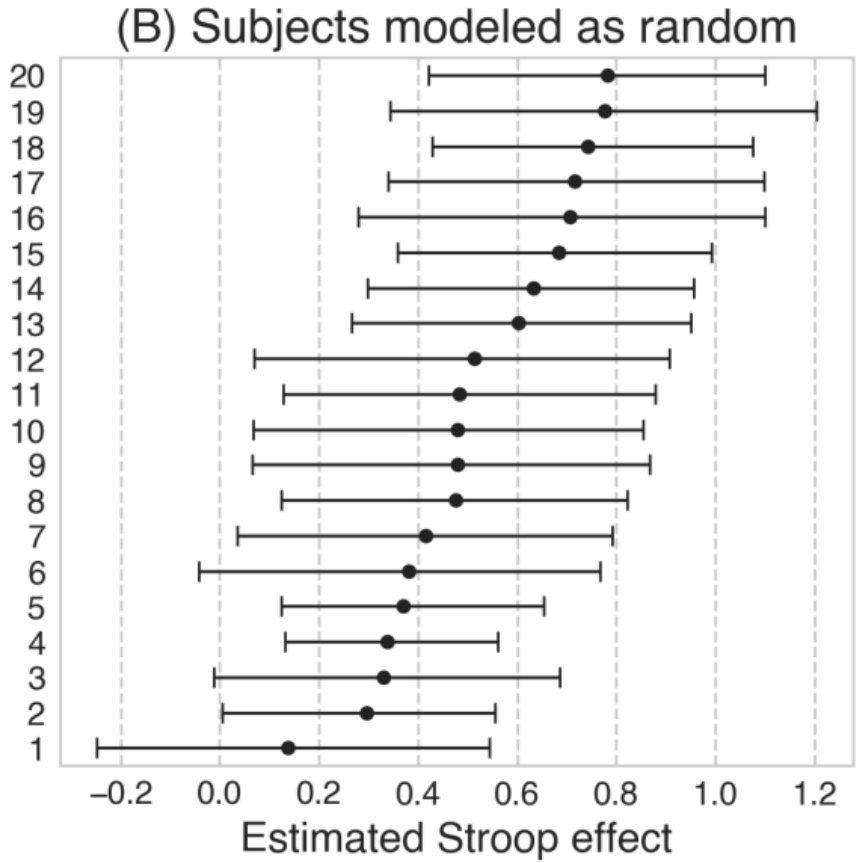
1. P-value and the null hypothesis statistical testing (NHST)
2. P-hacking
3. File drawer
3. Winner's curse
4. Effect sizes
5. Power
6. Positive Predictive Value
7. **Statistical generalizability**

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: one 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.

Conclusion

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

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