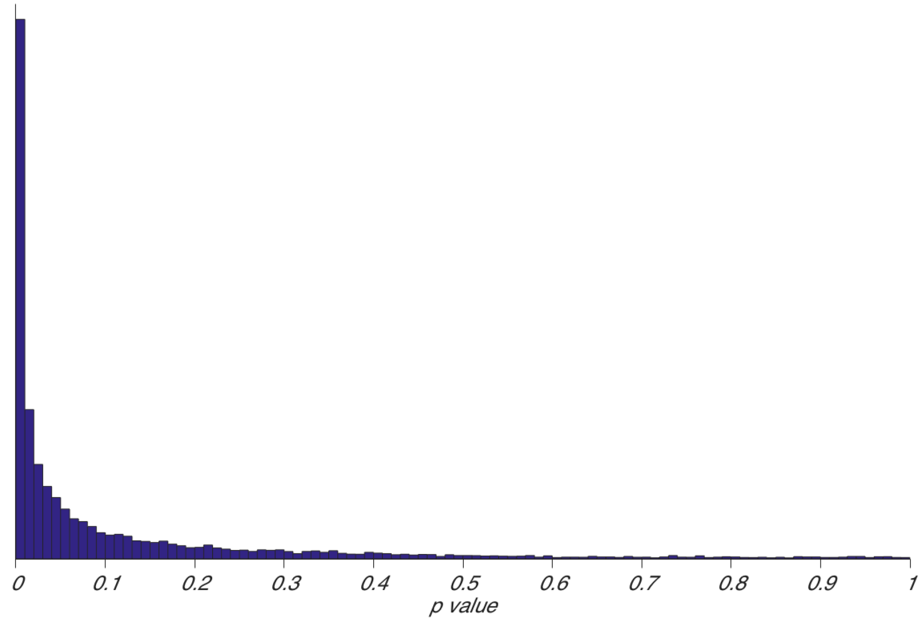
Smallest font

Calibration slide

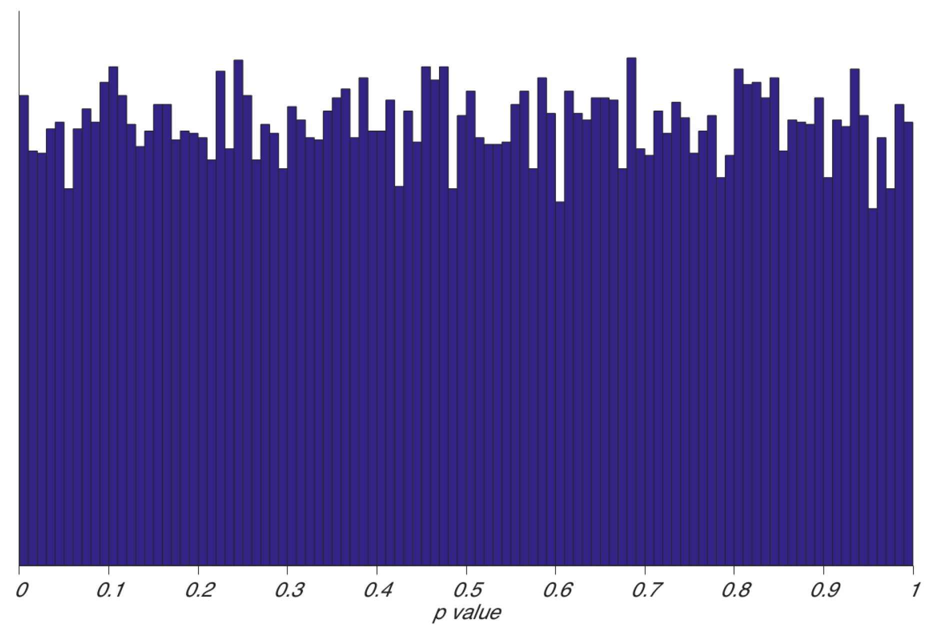
Smallest font

Distribution of p values

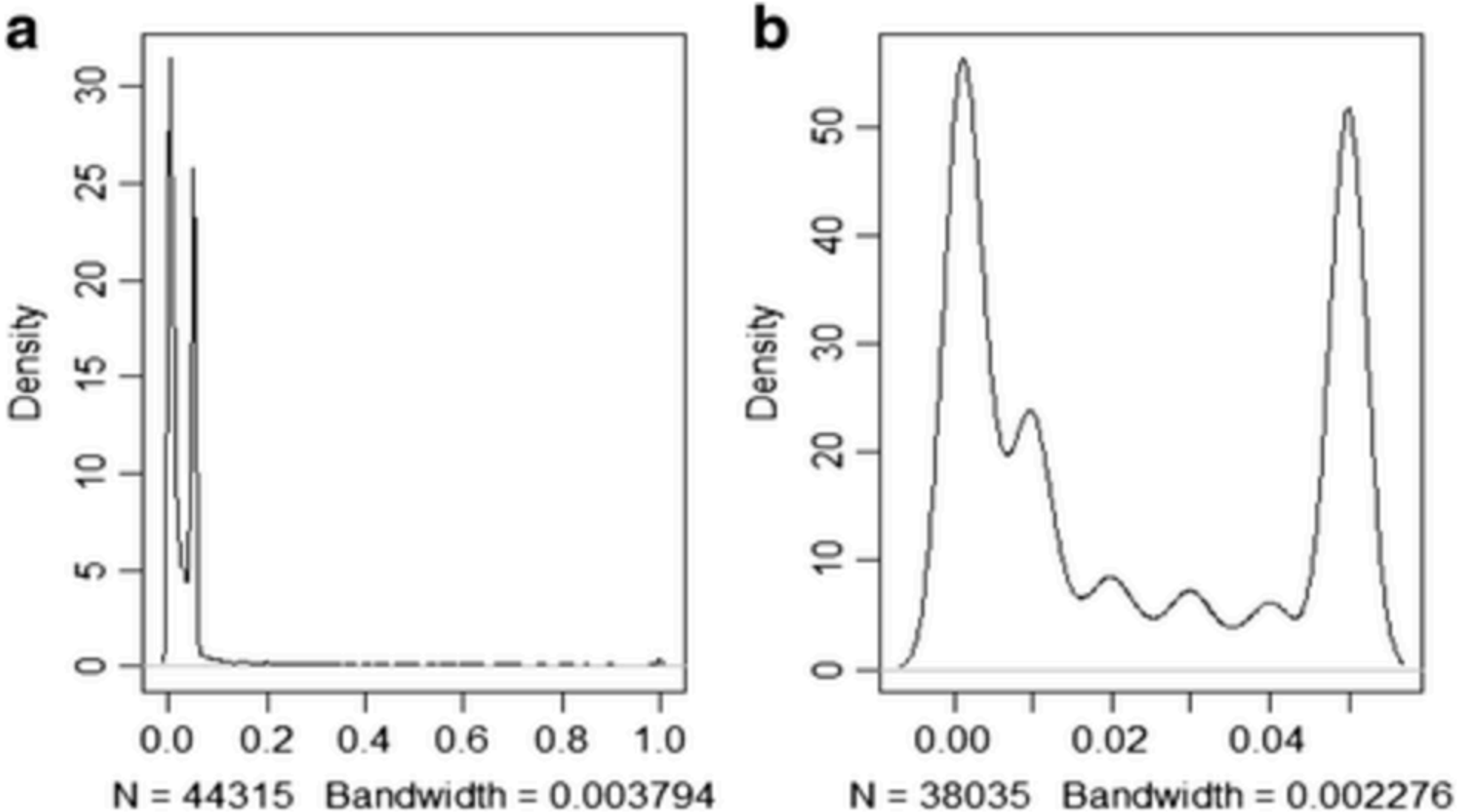
If H0 is false:



If H0 is true:



Not just neuroscience and psychology: Evidence of p-hacking in oral health



Kagereki et al., 2016

Power

&

Confidence

What is (statistical) power?

$1-\beta$

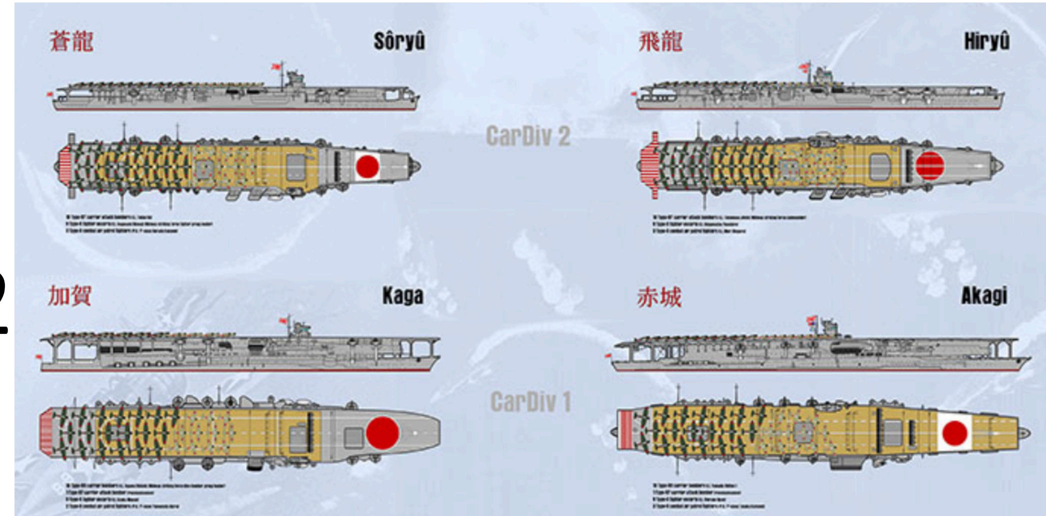
		Reality	
		Effect	No effect
Significant?	Yes	**Correct**	**Type I error** **False positive** **α error**
	No	**Type II error** **False negative** **β error**	**Correct**

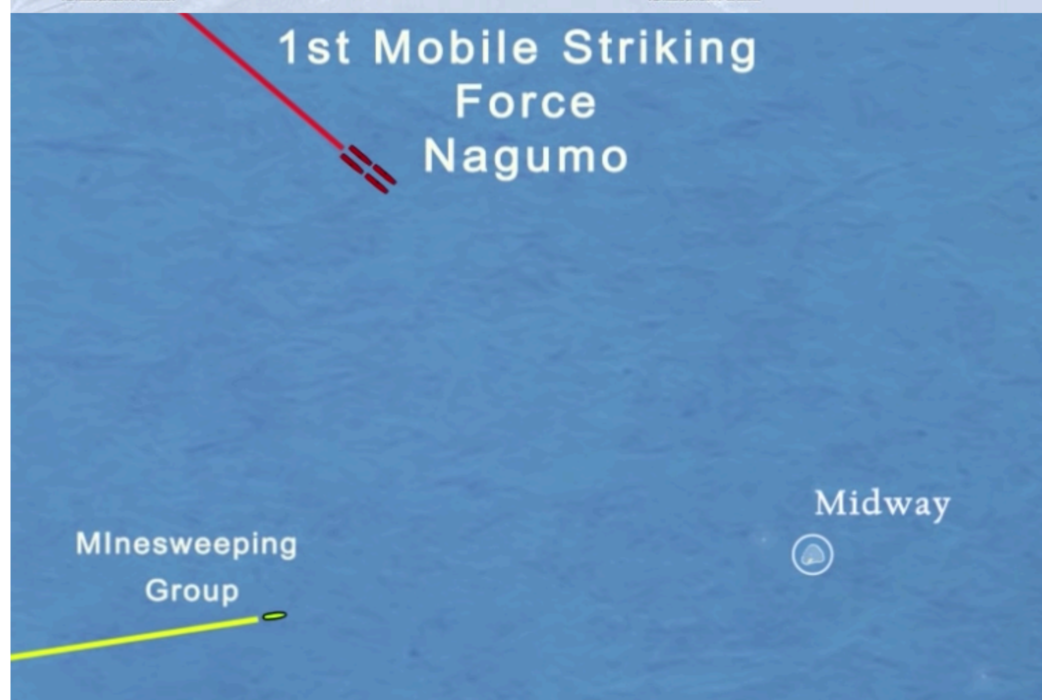
What is power really?

Why does it matter so much?

Kidō Butai
(Mobile Striking Force)

Pearl Harbor 12/1941
Wake Island 12/1941
Hong Kong 12/1941
Malaya 01/1942
Darwin Raid 02/1942
Singapore 02/1942
Borneo 03/1942
Java 03/1942
Sumatra 03/1942
Bataan 04/1942
Mandalay 05/1942
Philippines 05/1942

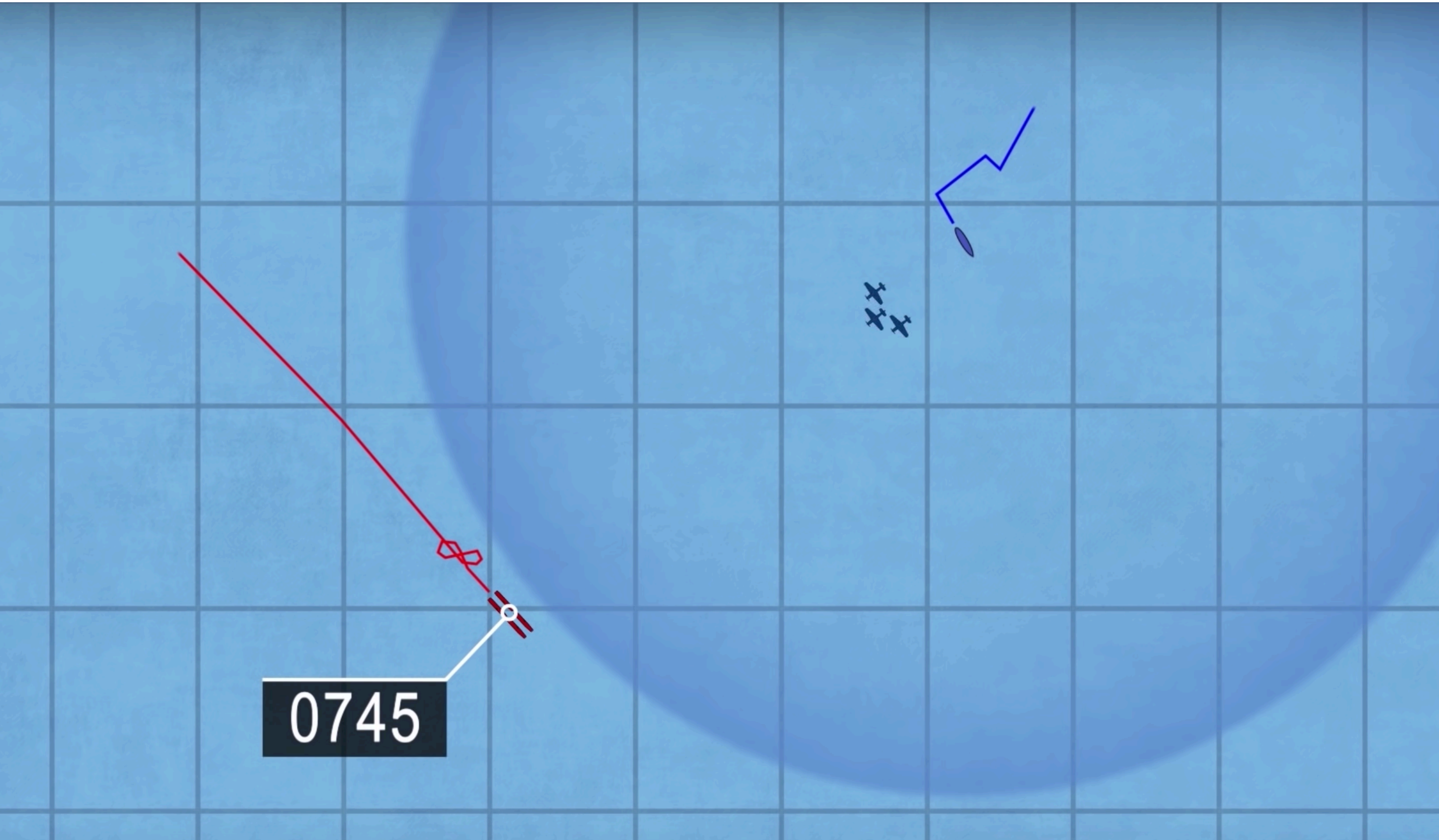




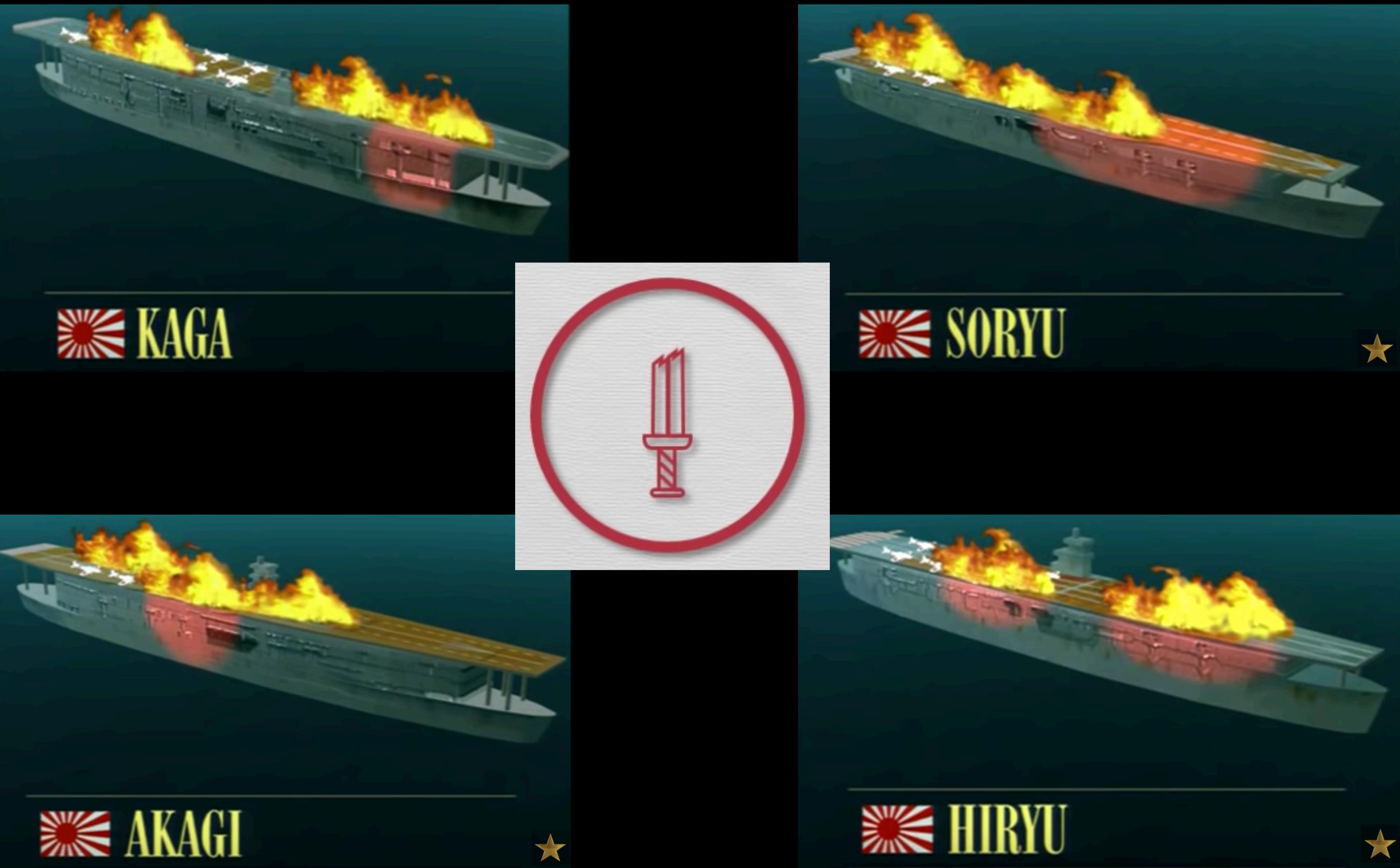
No US carriers spotted = evidence of absence

Nagumo

Reality: Absence of evidence!



Lack of power can be devastating



The link to confidence (intervals)

- Higher power shrinks confidence intervals
- CI = interval estimate of a population parameter
- If we repeated a study often, x% (usually 95 or 99%) of the CIs would contain the population parameter.

How to calculate a confidence interval

- The idea is to multiply a critical z value with the SEM.

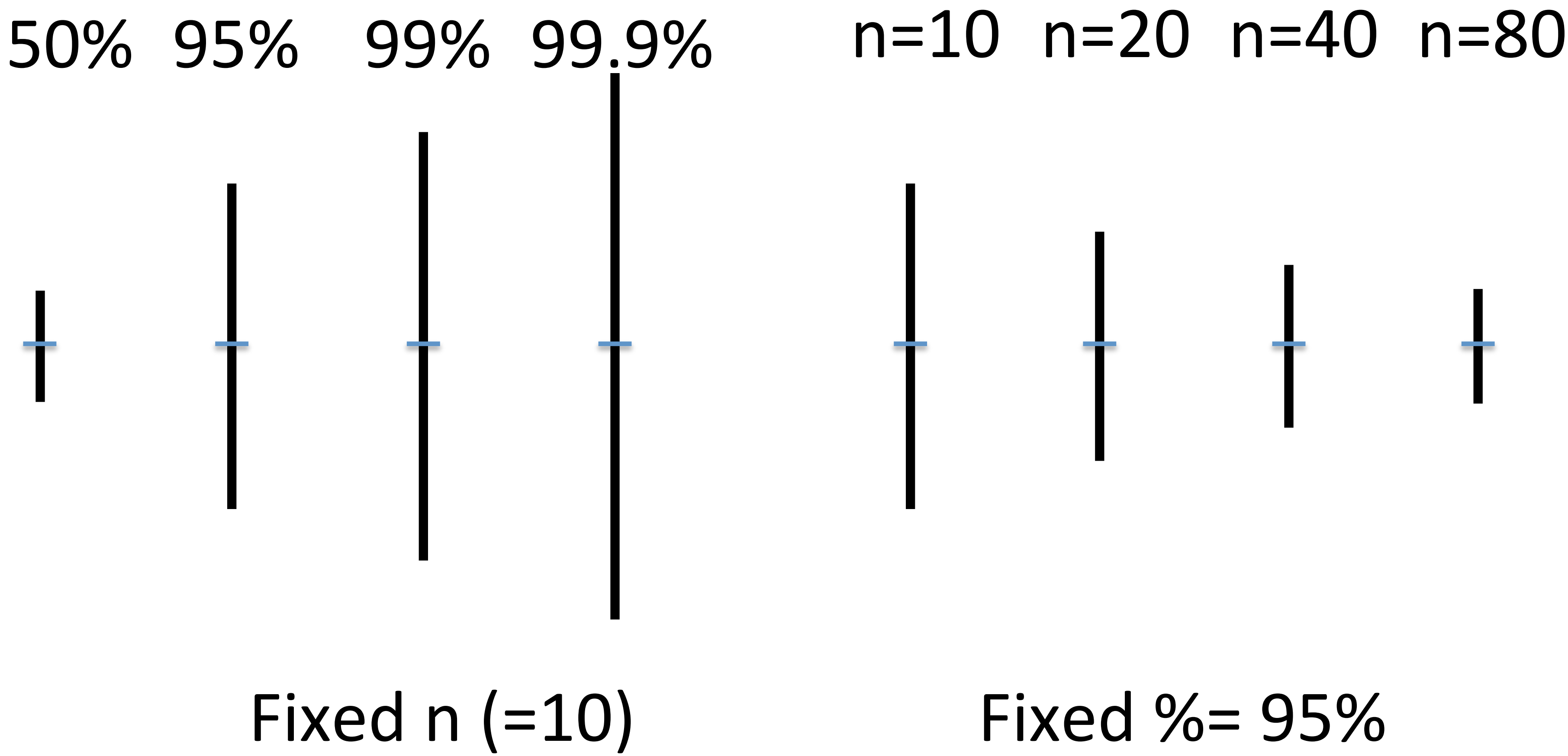
$$SEM = \frac{\sigma}{\sqrt{n}}$$

- **Critical z-value**: Found by taking the confidence level (e.g. 95%) and looking it up in a z-table: $\mathbf{z_{a/2}}$

$$CI = \overline{x} \pm z_{a/2} * \frac{\sigma}{\sqrt{n}}$$

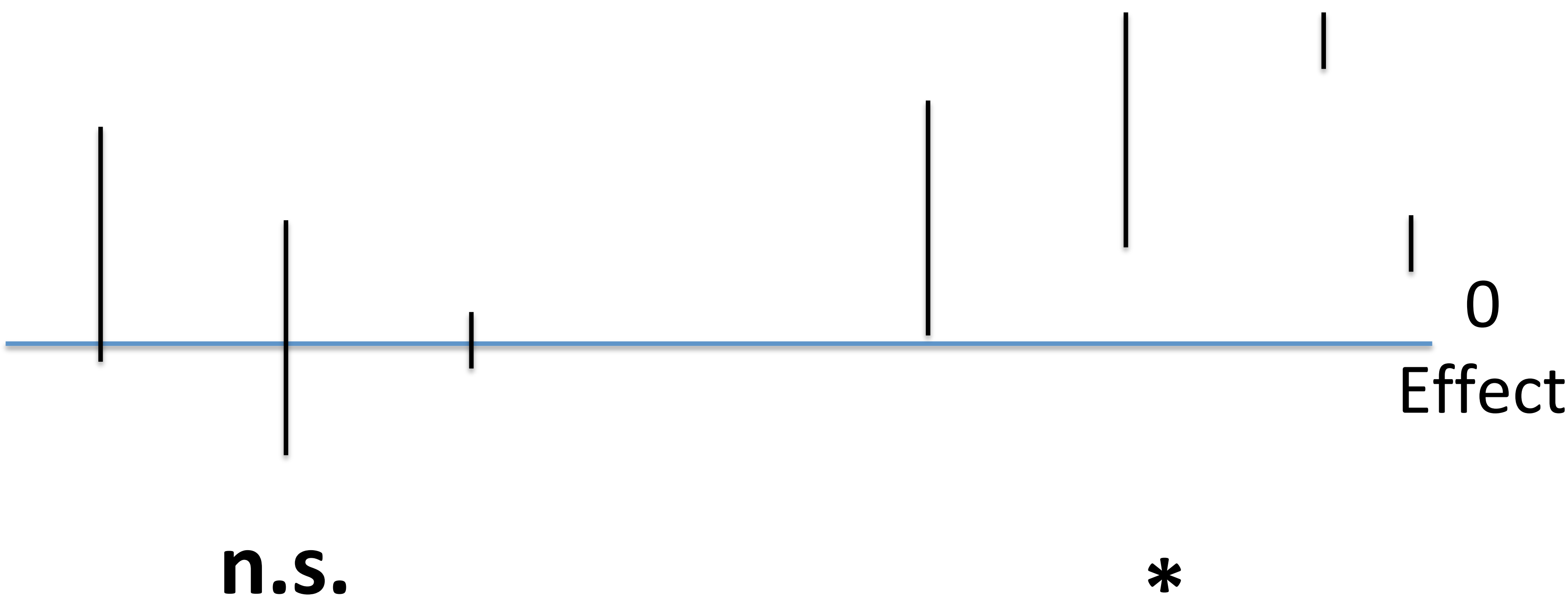
50%: 0.67
95%: 1.96
99%: 2.58
99.9%: 3.29

Clamped confidence intervals



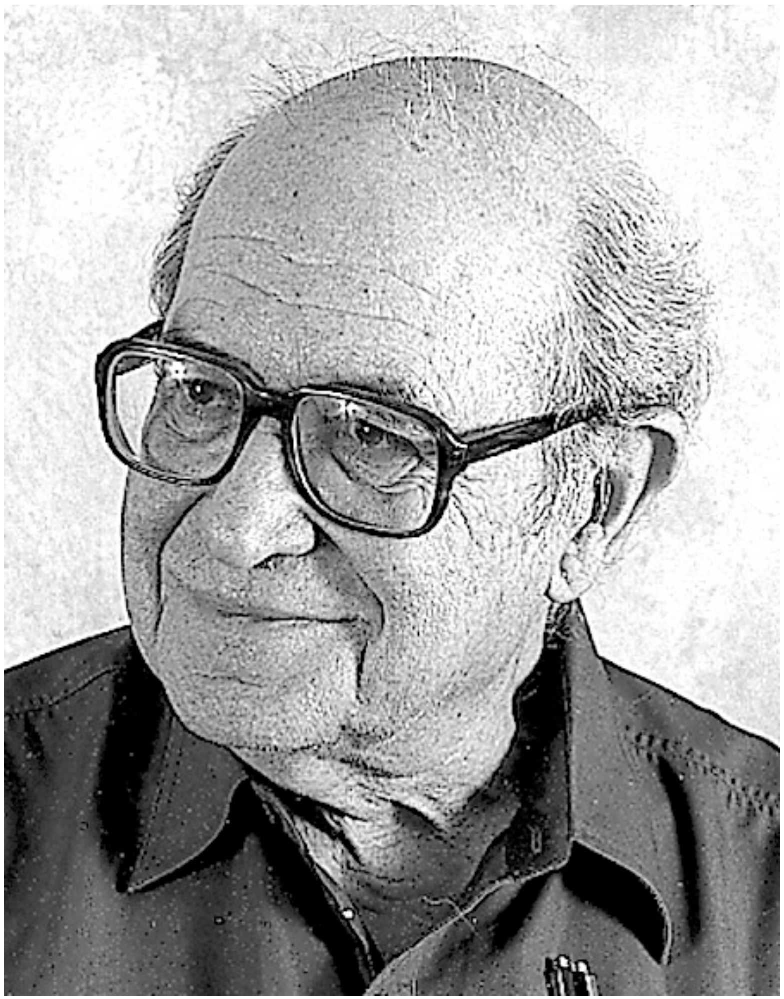
Why is CI a more informative measure than significance alone?

- It allows to distinguish absence of evidence from evidence of absence



The link to effect sizes

- We need to talk about effect sizes



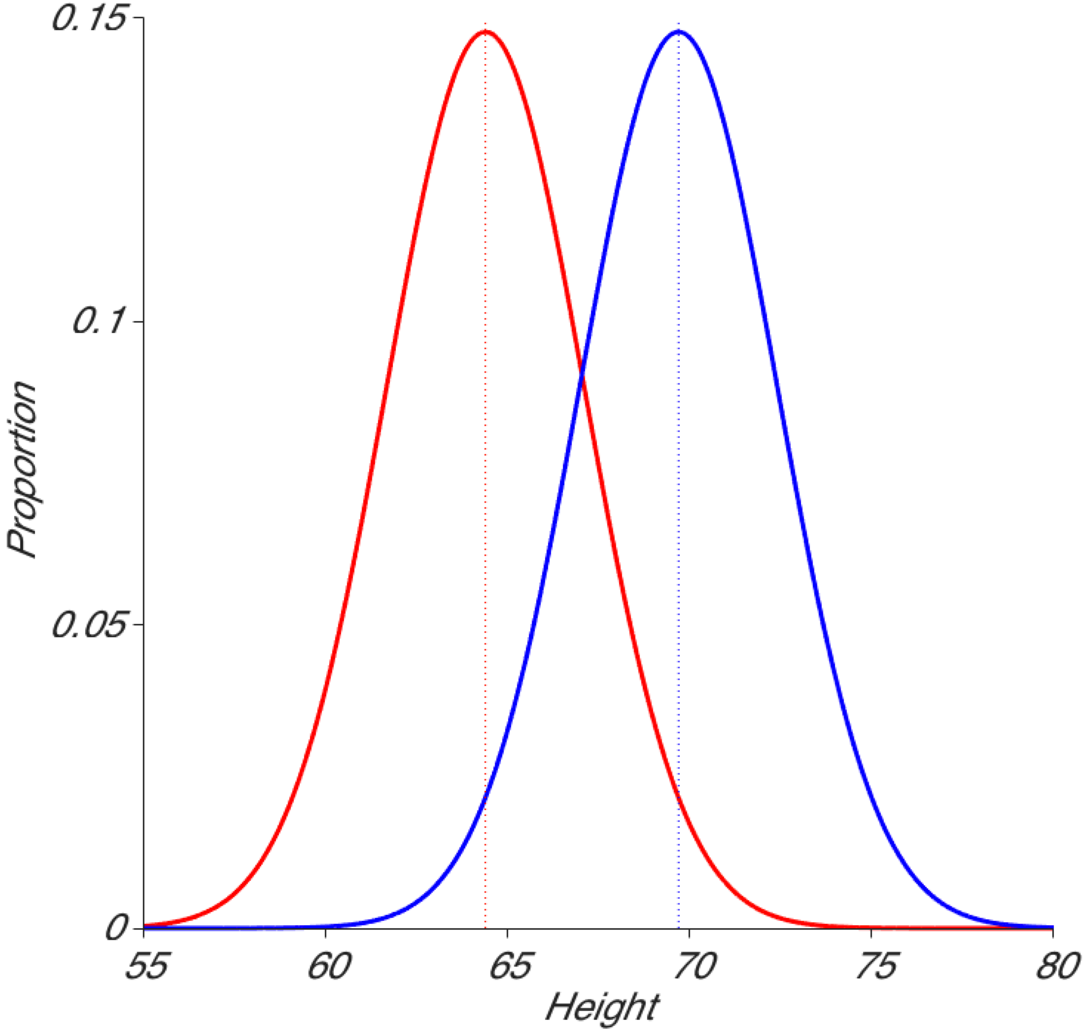
Jacob Cohen

Cohen's d:

$$d = \frac{\bar{X}_1 - \bar{X}_2}{\mathrm{SD}}$$

Effect sizes are counterintuitive

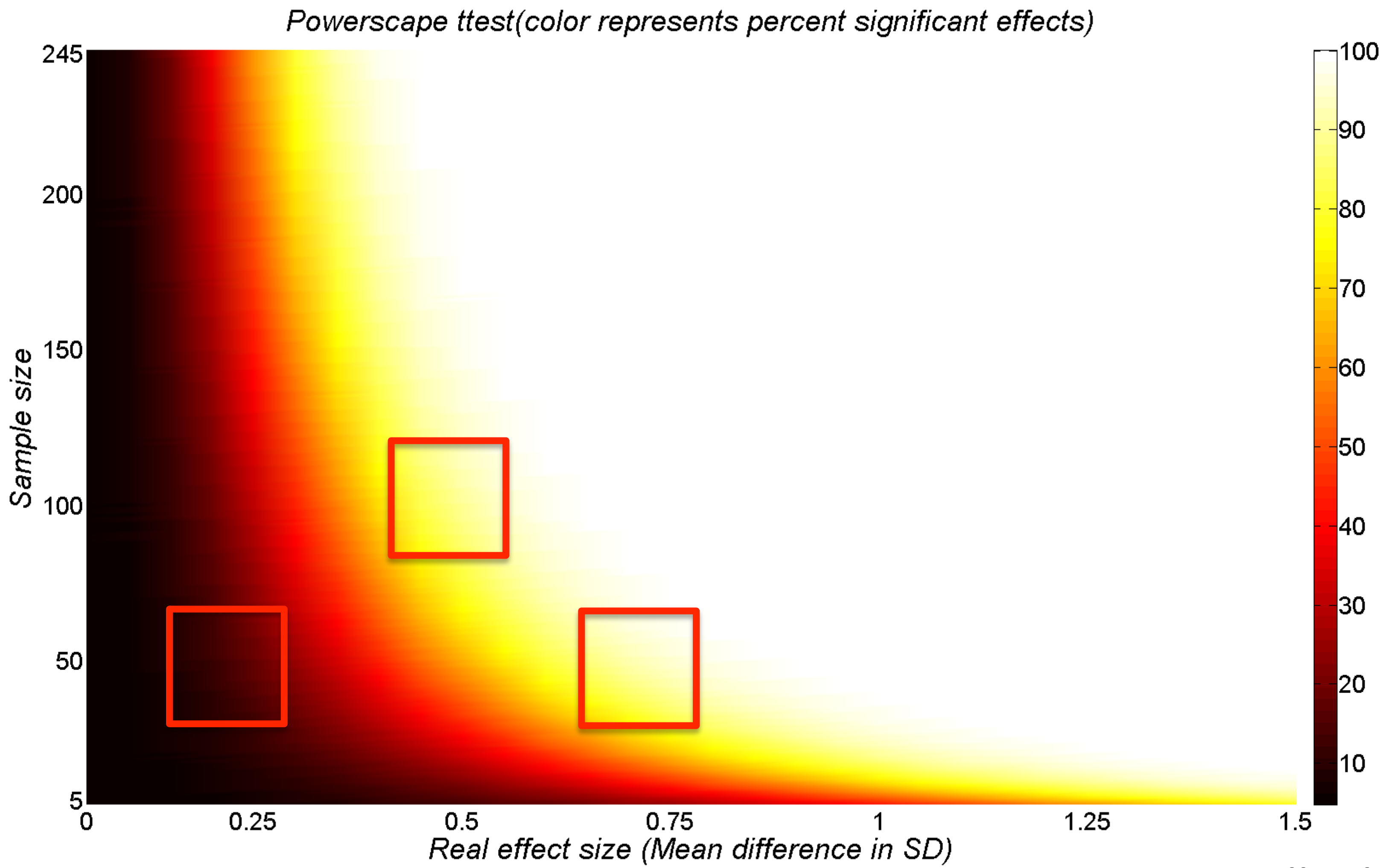
- Undiscovered big effects are rare.
- The average size of asteroids discovered annually shrinks by 2.5 percent each year (Arbesman, 2012)
- Similar things are true for effects – the probability of discovering them is not independent of their size.



Large effects are rare

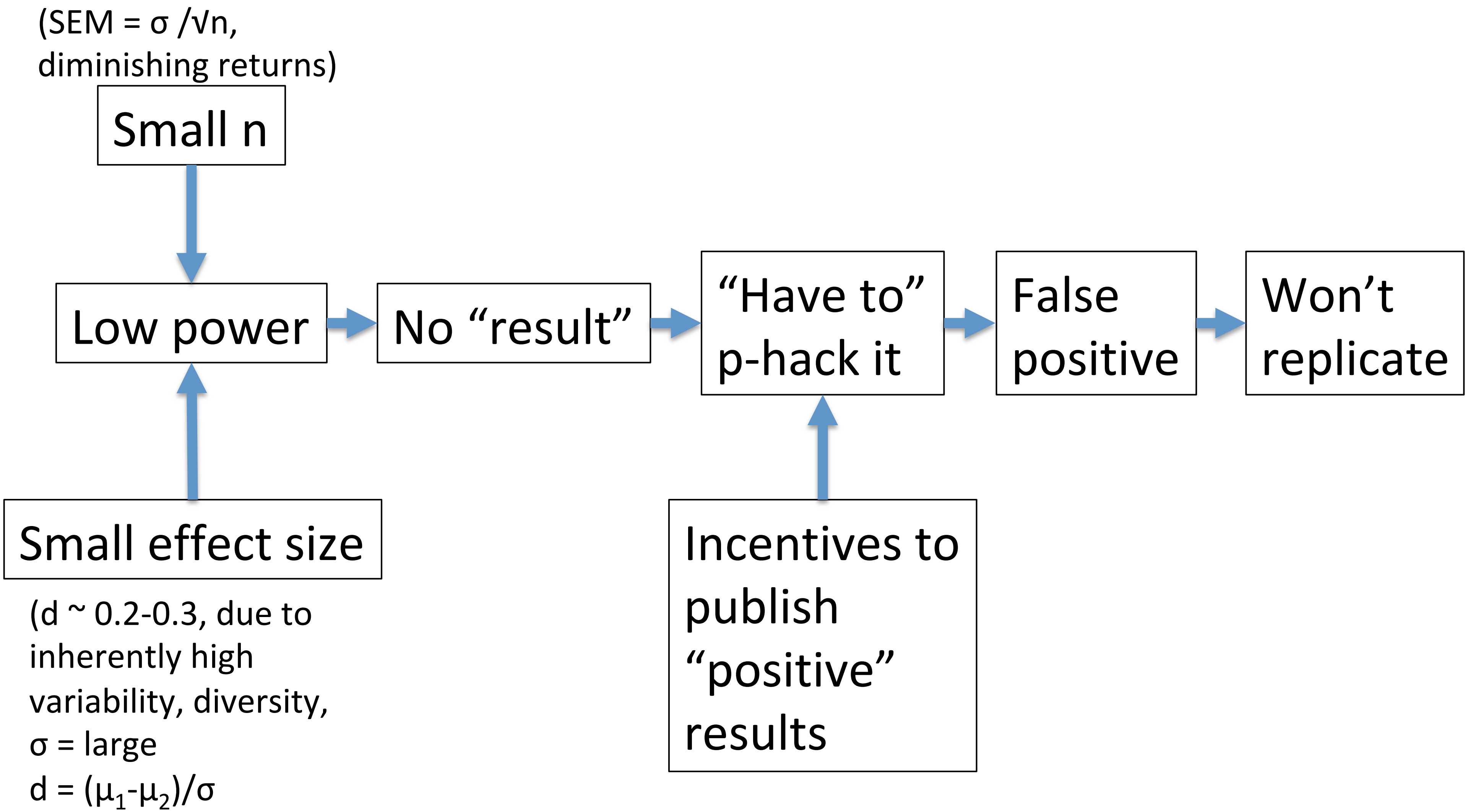
- Average pygmy height = 4 ft, 11”
- Average NBA player height = 6 ft, 7”
- Cohen’s d = 8
- The effect size of SSRI antidepressants is − on average − 0.30. (Cipriani et al., 2018).
- The effect size of psychotherapy is − on average − 0.30 (Hoyt et al., 2017).

Why does this matter? The powerscape



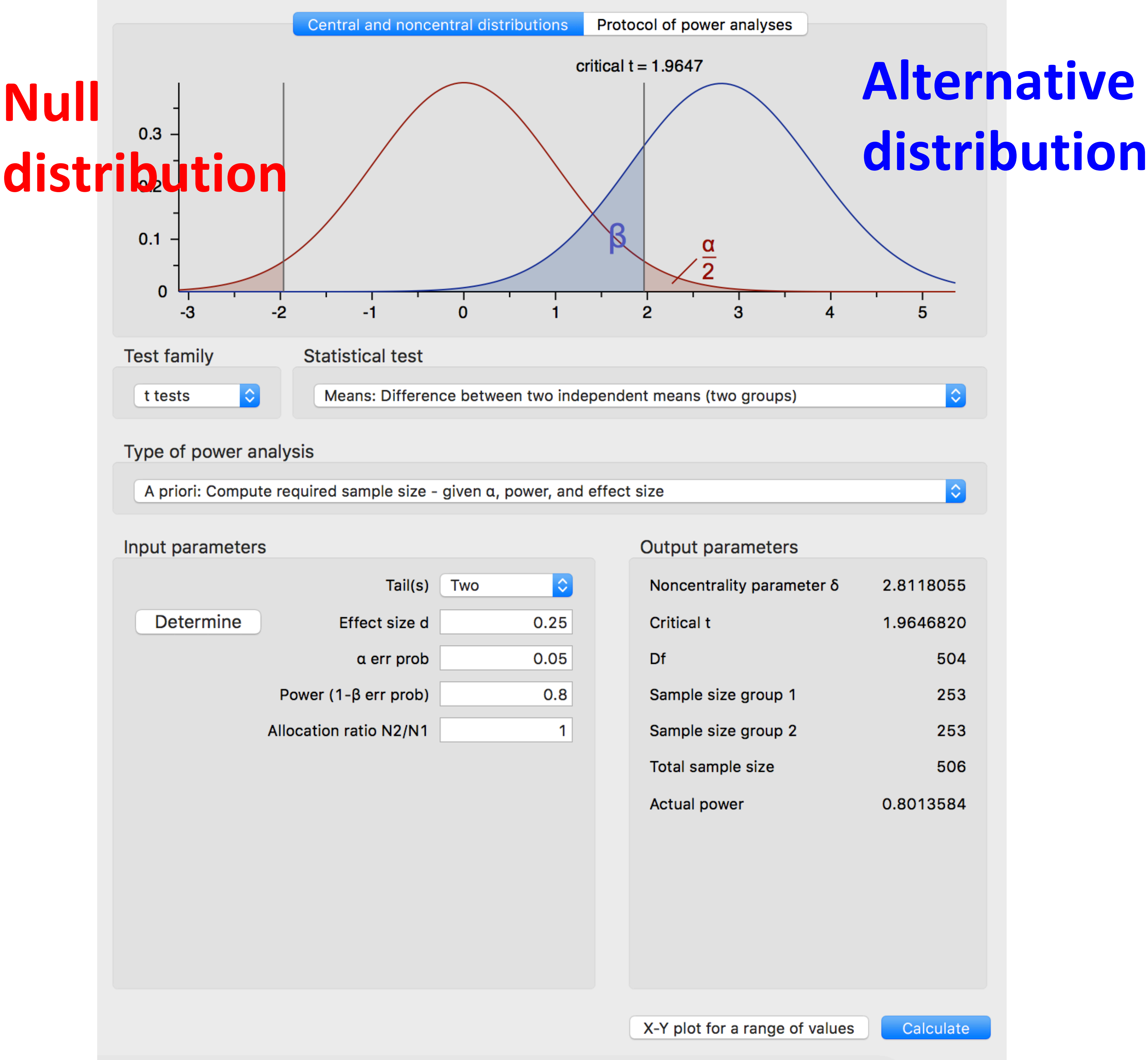
Wallisch, 2015

The replication crisis as a result of power failure

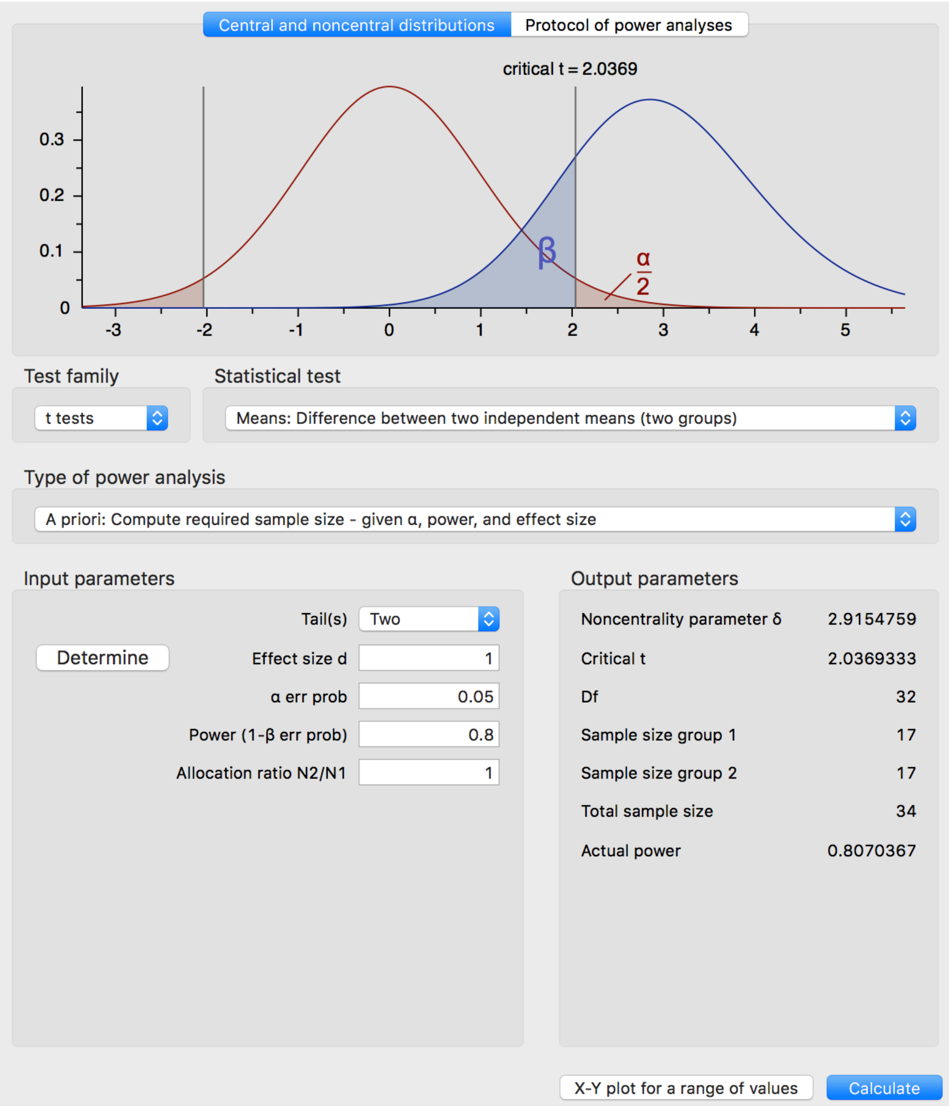


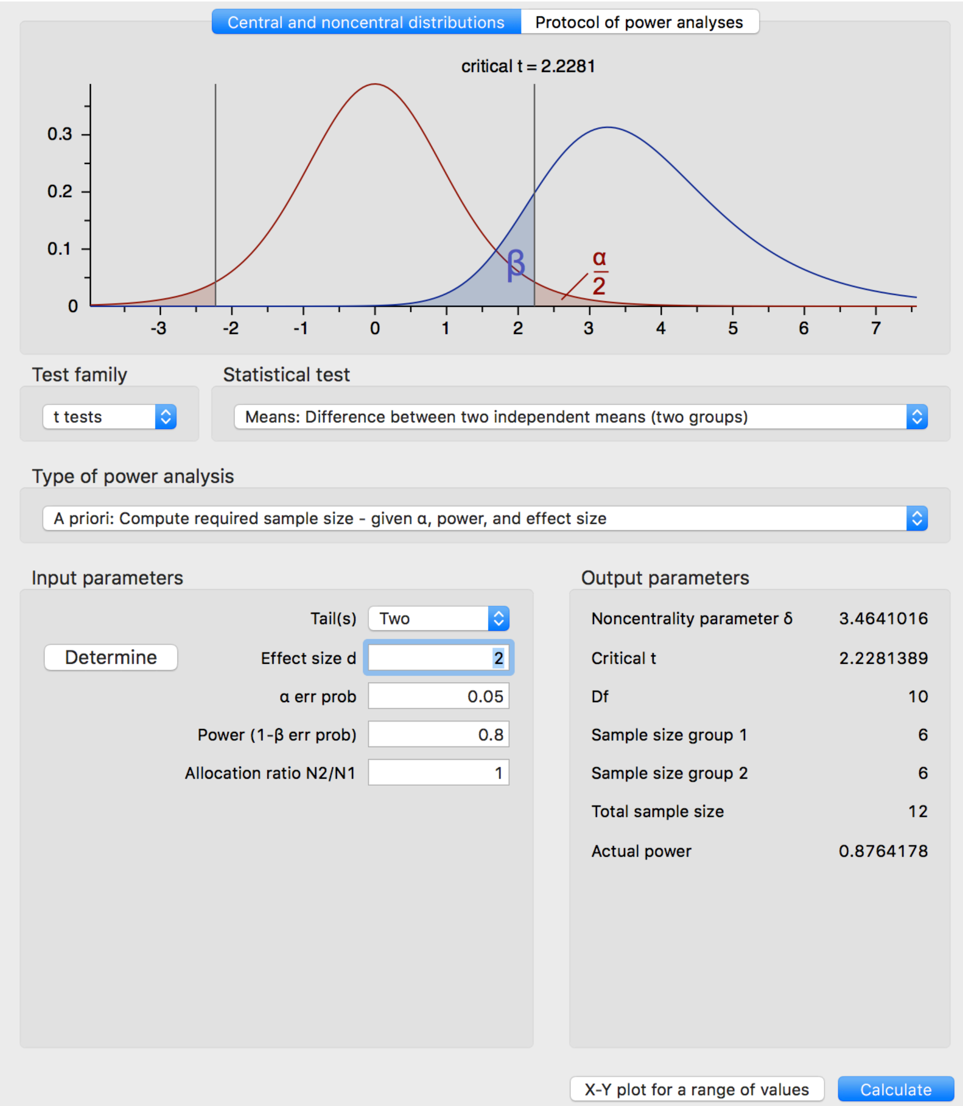
What to do?

- *A priori* (before the study) power analyses.
- Recommended tool: G*Power (free)

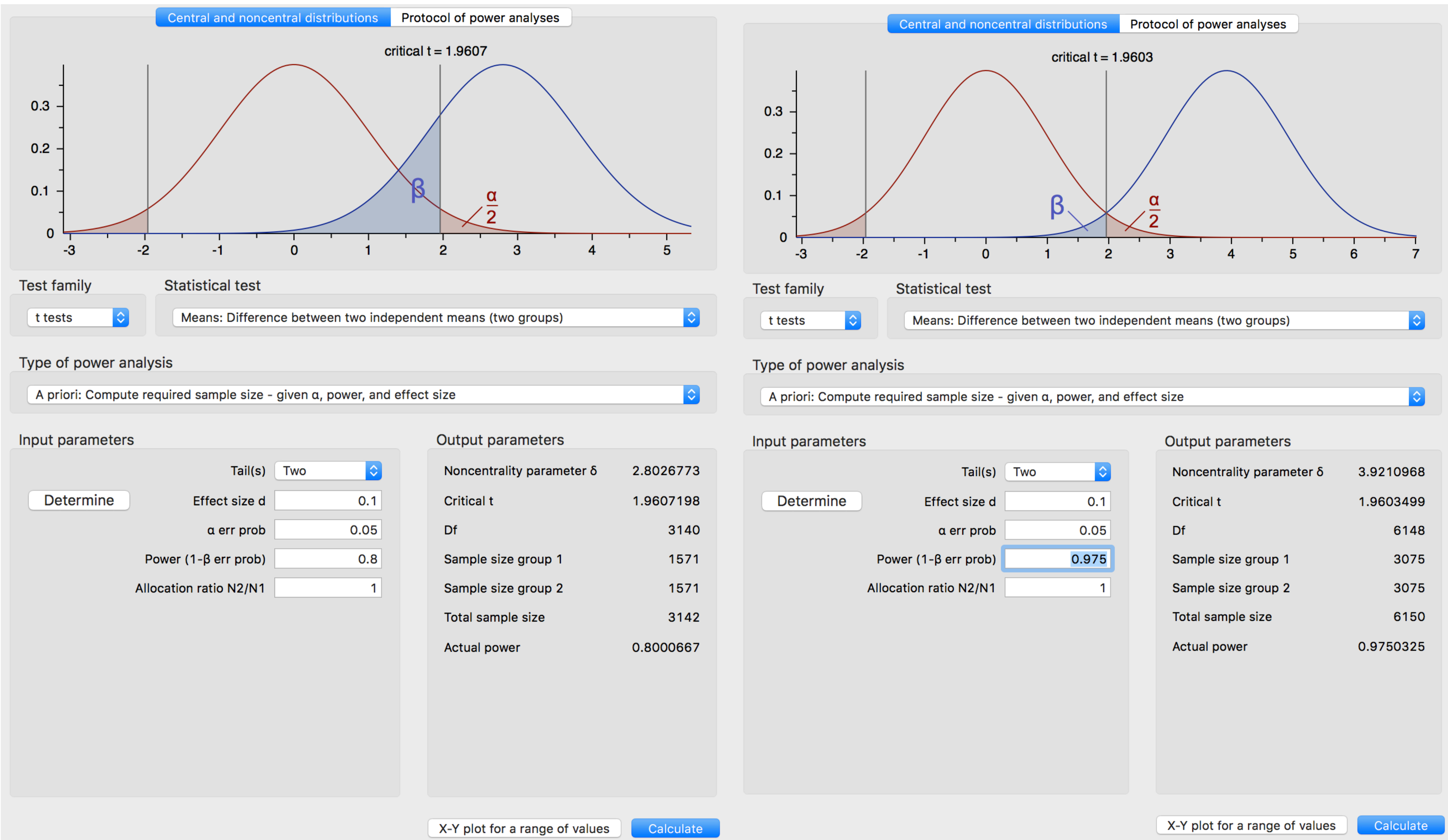

Null distribution
Alternative distribution
Central and noncentral distributions
Protocol of power analyses
critical t = 1.9647
β
α/2
Test family
t tests
Statistical test
Means: Difference between two independent means (two groups)
Type of power analysis
A priori: Compute required sample size - given α, power, and effect size
Input parameters
Tail(s) Two
Determine
Effect size d 0.25
α err prob 0.05
Power (1-β err prob) 0.8
Allocation ratio N2/N1 1
Output parameters
Noncentrality parameter δ 2.8118055
Critical t 1.9646820
Df 504
Sample size group 1 253
Sample size group 2 253
Total sample size 506
Actual power 0.8013584
X-Y plot for a range of values
Calculate

Large effects are easy to find



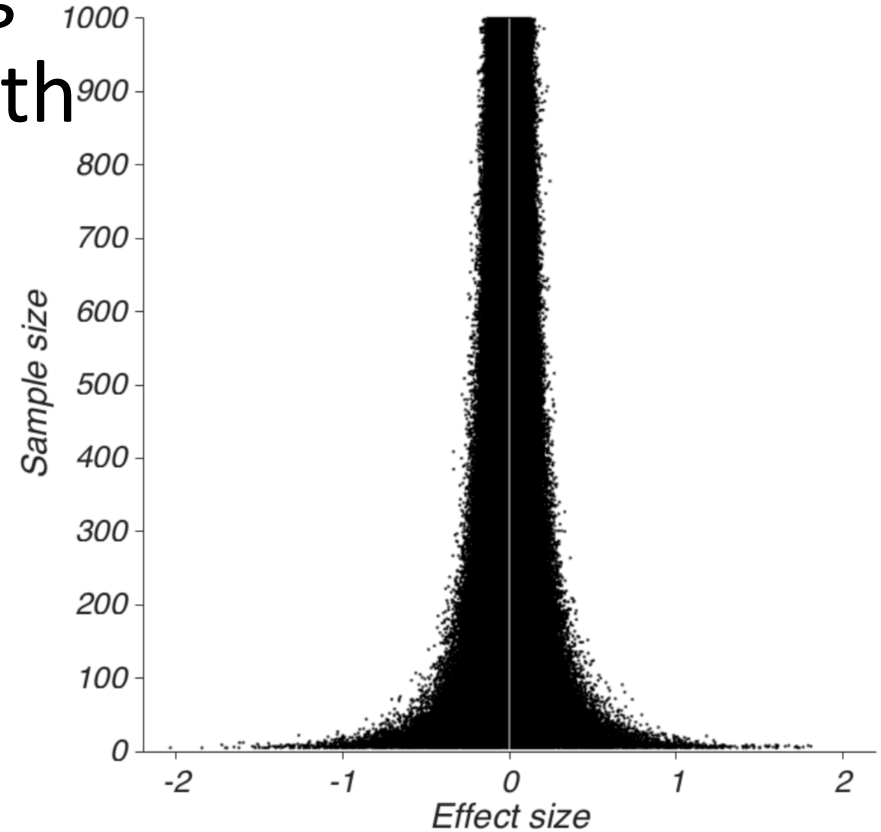


Small effects are not – particularly if you want to be sure



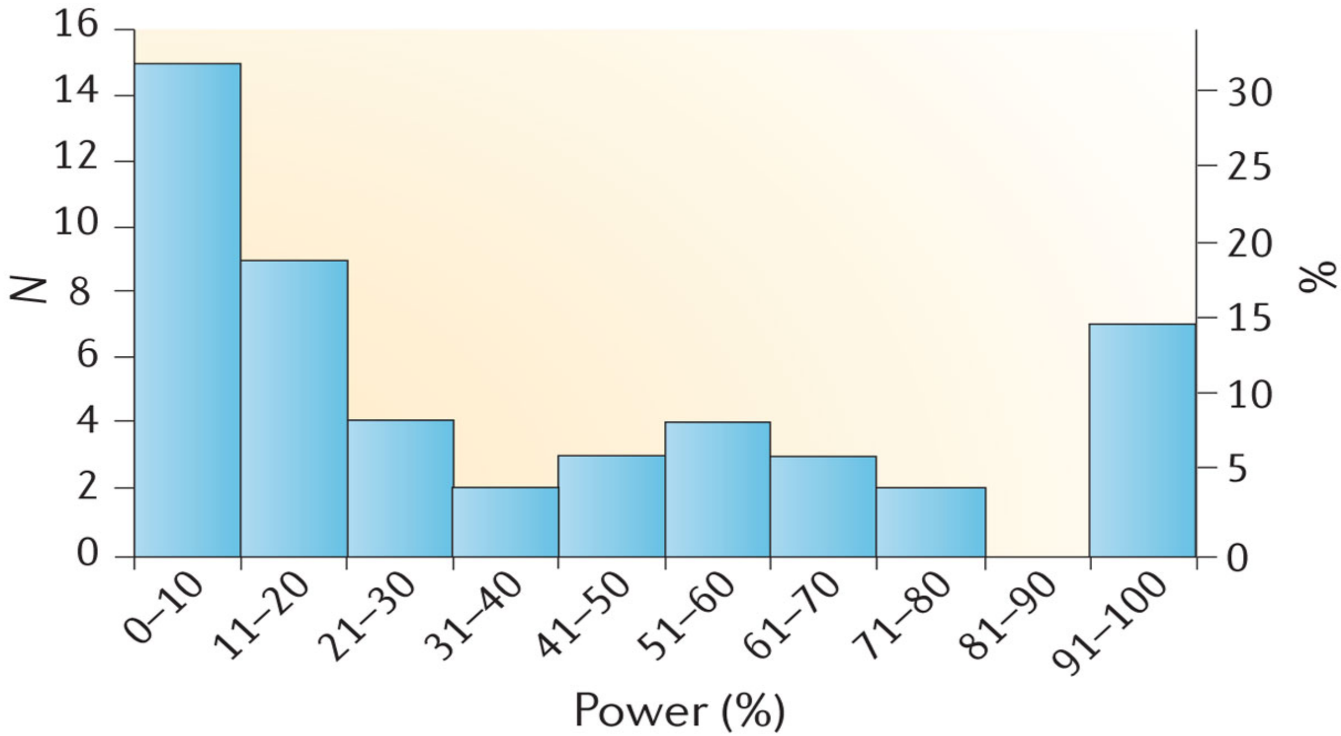
Only a priori power is real power

- Why?
- Because if one runs an underpowered study, one will likely have large sampling errors, in combination with publication bias, this overstates post-hoc or “observed” power.
- Observed power can be misleading:
- Witness the funnel plot



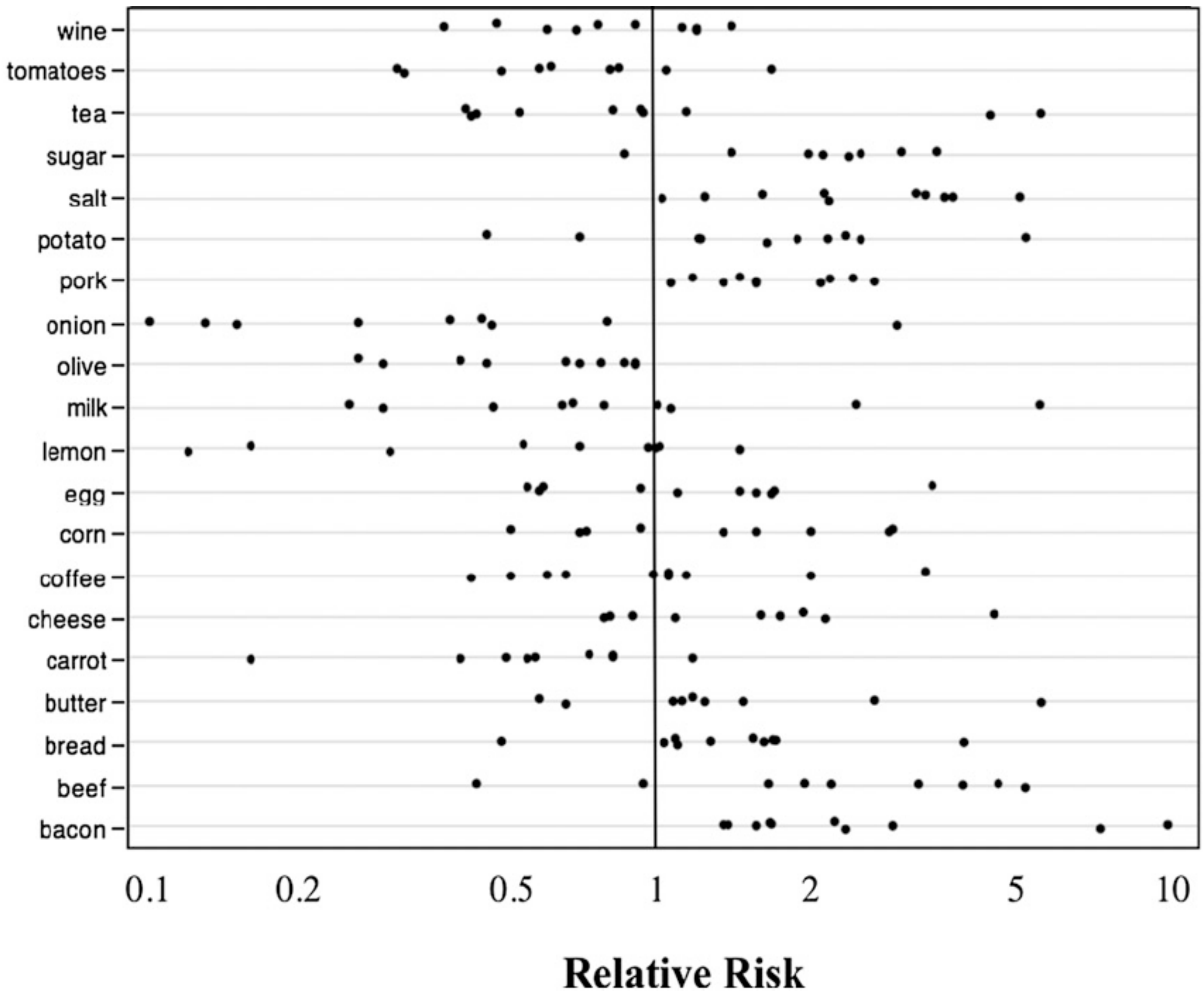
How is neuroscience doing, power-wise?

- Neuroscience research has an average power of between 8% and 31%. (Button et al., 2013).
- Median power is 21%



Nature Reviews | Neuroscience

It's not just us: *"Is everything we eat associated with cancer?"* (Schoenfeld & Ioannidis 2013)



Why power is linked to confidence

- What is the probability that a published result is actually true?
- If we set the significance level α to 0.05, does this mean that the false positive rate in the field is 5%?
- No!

Why most published research findings are false (Ioannidis, 2005)

- It also depends on the prior probability of a true effect in a given field as well as power.
- PPV = "positive predictive value" – *post* study probability that a result is true:
- R: Ratio of true to false effects in a field.
- PPV links power, alpha and R.

$$PPV = (1-\beta)R / (R - \beta R + \alpha)$$

$$PPV = (1-\beta)R / ((1-\beta)R + \alpha)$$

Deriving PPV

RESEARCH/REALITY	TRUE	FALSE	TOTAL
SIG	**$1-\beta$**	**α**	**$1-\beta + \alpha$**
NONSIG	**β**	**$1-\alpha$**	**$\beta + 1-\alpha$**
TOTAL	**1**	**1**	**2**

RESEARCH/REALITY	TRUE	FALSE	TOTAL
SIG	**$N_T\,(1-\beta)$**	**$N_F \bullet \alpha$**	**$N_T\,(1-\beta) + N_F \bullet \alpha$**
NONSIG	**$N_T \bullet \beta$**	**$N_F\,(1-\alpha)$**	**$N_T \bullet \beta + N_F\,(1-\alpha)$**
TOTAL	**N_T**	**N_F**	**$N_T + N_F = c$**

$R = N_T/N_F$ $\quad$ $R \bullet N_F = c - N_F$ $\quad$ $R = N_T/N_F$ $\quad$ $N_T = c \bullet R/(R+1)$

$N_T = R \bullet N_F$ $\quad$ $R \bullet N_F + N_F = c$ $\quad$ $R/(R+1) = (N_T/N_F)/c/N_F$

$N_T = c - N_F$ $\quad$ $R + 1 = c/N_F$ $\quad$ $R/(R+1) = N_T/c$ $\quad$ $N_F = c/(R+1)$

Deriving PPV

RESEARCH/REALITY	TRUE	FALSE	TOTAL
SIG	$N_T\ (1-\beta)$	$N_F \bullet \alpha$	$N_T\ (1-\beta) + N_F \bullet \alpha$
NONSIG	$N_T \bullet \beta$	$N_F\ (1-\alpha)$	$N_T \bullet \beta + N_F\ (1-\alpha)$
TOTAL	N_T	N_F	$N_T + N_F = c$

$R = N_T/N_F$	$R \bullet N_F = c - N_F$	$R = N_T/N_F$	$N_T = c \bullet R/(R+1)$
$N_T = R \bullet N_F$	$R \bullet N_F + N_F = c$	$R/(R+1) = (N_T/N_F)/c/N_F$	
$N_T = c - N_F$	$R + 1 = c/N_F$	$R/(R+1) = N_T/c$	$N_F = c/(R+1)$

RESEARCH/REALITY	TRUE	FALSE	TOTAL
SIG	$c(1-\beta) \bullet R/(R+1)$	$c \bullet \alpha/(R+1)$	$c \bullet (R \bullet (1-\beta)+\alpha)/(R+1)$
NONSIG	$c \bullet \beta \bullet R/(R+1)$	$c \bullet (1-\alpha)\ /(R+1)$	$c \bullet (1 + R \bullet \beta-\alpha)/(R+1)$
TOTAL	$c \bullet R/(R+1)$	$c/(R+1)$	c

Deriving PPV

RESEARCH/REALITY	TRUE	FALSE	TOTAL
SIG	$c(1-\beta)\bullet R/(R+1)$	$c\bullet\alpha/(R+1)$	$c\bullet(R\bullet(1-\beta)+\alpha)/(R+1)$
NONSIG	$c\bullet\beta\bullet R/(R+1)$	$c\bullet(1-\alpha)\ /(R+1)$	$c\bullet(1 + R\bullet\beta-\alpha)/(R+1)$
TOTAL	$c\bullet R/(R+1)$	$c/(R+1)$	c

$PPV = p(\text{effect true} \mid \text{significant}) = p(\text{true})/p(\text{sig})\bullet p(\text{sig}|\text{true})$

$p(\text{true}) = N_T/c = R/(R+1)$

$p(\text{sig}|\text{true}) = 1-\beta$

$p(\text{sig}) = p(\text{sig}|\text{true})\bullet p(\text{true})+p(\text{sig}|\text{false})\bullet p(\text{false})$

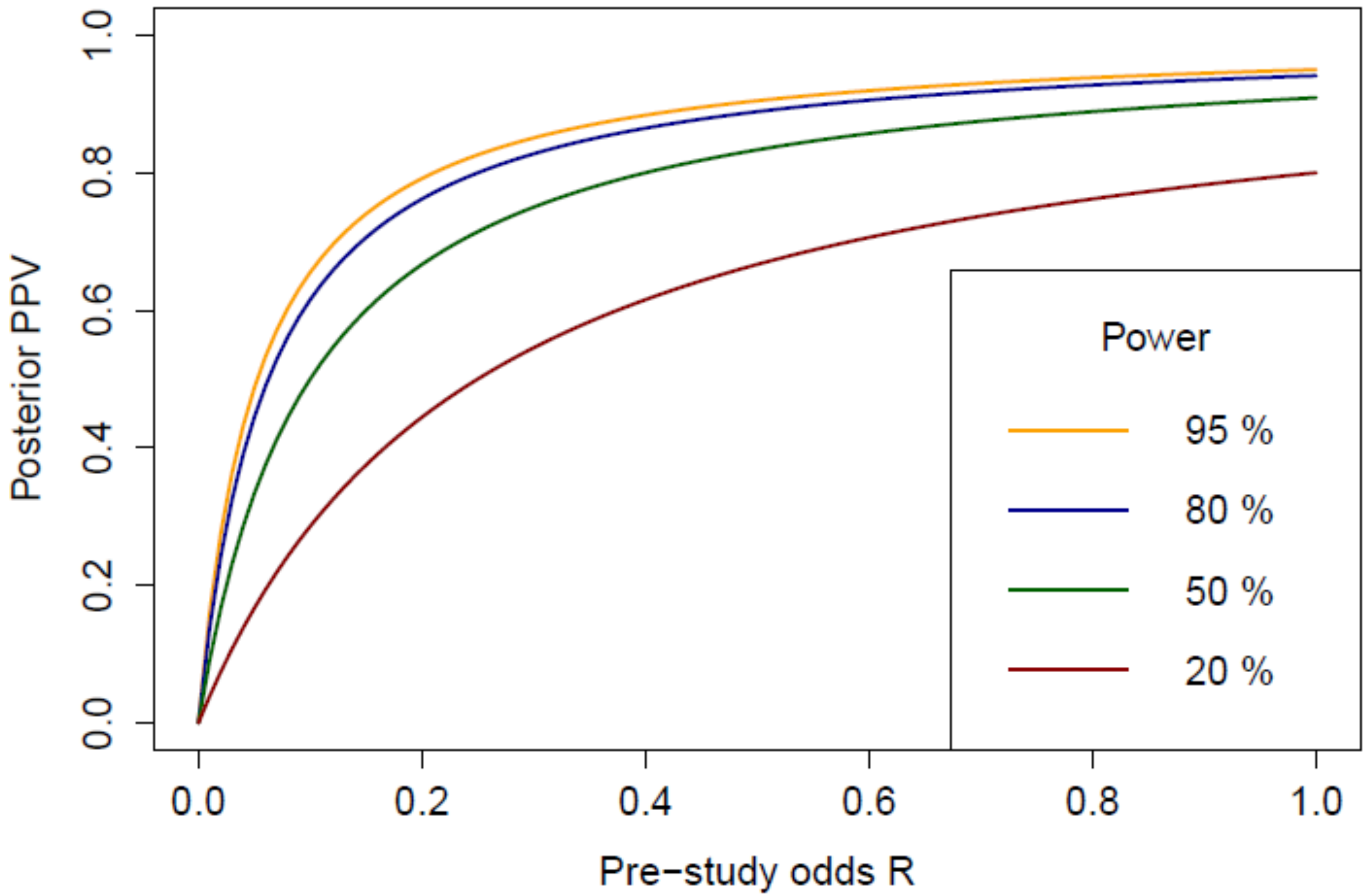
$p(\text{sig}) = 1-\beta\bullet R/(R+1)+\alpha\bullet N_F/c$

$p(\text{sig}) = 1-\beta\bullet R/(R+1)+\alpha\bullet(R+1)$

$PPV = R/(R+1)/1-\beta\bullet R/(R+1)+\alpha\bullet(R+1)\bullet 1-\beta$

$PPV = (1-\beta)\bullet R/((1-\beta)\bullet R+\alpha)$

Power critically affects PPV at low R



Higher power makes everything better

- Less misses (beta/type II errors)
- Less false positives (alpha/type I errors)
- Makes publication of null results legitimate (evidence of absence of an effect)
- Makes confidence intervals narrower
- Higher PPV