Power Calculations in R

Doctoral school 2024 --- Neurosciences week

Boulakis Paradeisios Alexandros, MSc
FNRS Aspirant
Physiology of Cognition Lab
GIGA CRC In vivo imaging
University of Liège

December 09, 2024

If you are not cheating, you are not trying

Cross Validated

ChatGPT

StackExchange

Google

B-STAT

CHU de Liège

LIÈGE université

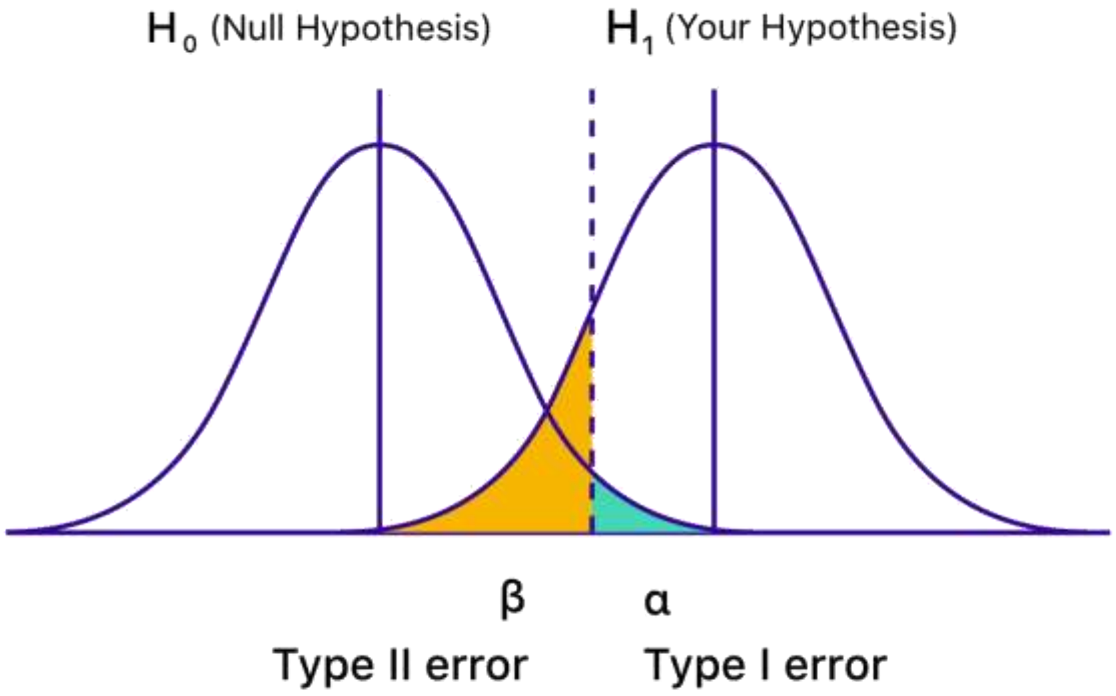
Statistical Concepts

Significance

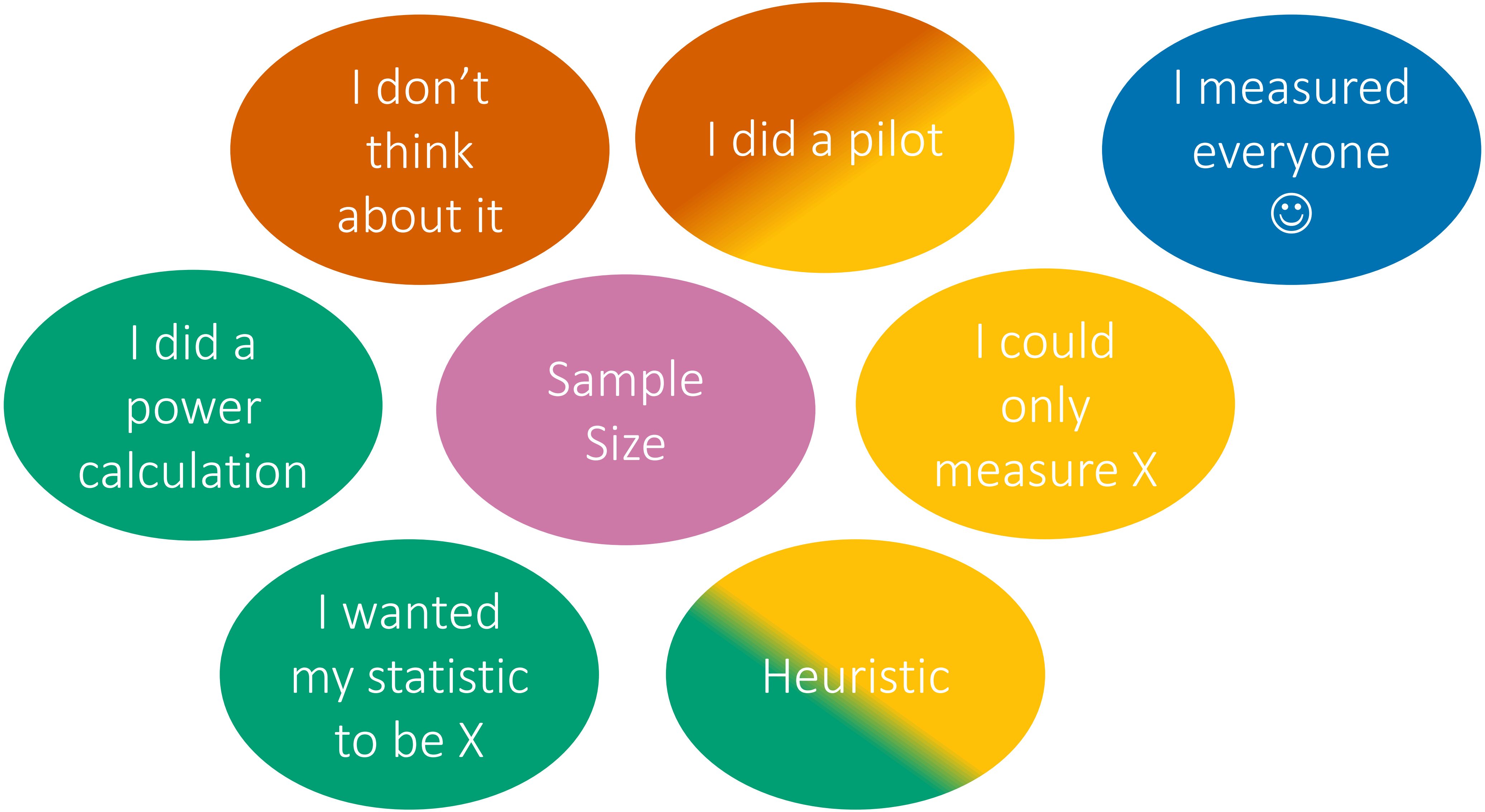
- Likelihood of results under H0
- α level = Type 1 error = False Positive
- Saying that something exists when it does not
- I tolerate finding a result that does not exist X% of the time
- p-value = How surprising my results are if H0 is true
- Heuristically around .05 *(God hates the number .051)*

Power

- Probability of correctly H0
- β level = Type 2 error = False Negative
- Saying that something does not exist when it does
- I tolerate not finding a result that exists X% of the time
- Heuristically around .2 *(God hates the number even more)*



Why did you select your sample size ?



I don’t think about it

I did a pilot

I measured everyone ☺

I did a power calculation

Sample Size

I could only measure X

I wanted my statistic to be X

Heuristic

Lakens, 2020

Why justify ?

- Experiments are expensive

  Minimum resources

- Reduce risk of random sampling variability

  Increases your confidence in your results.

- Forces you to think your analysis

  Improves statistical questions

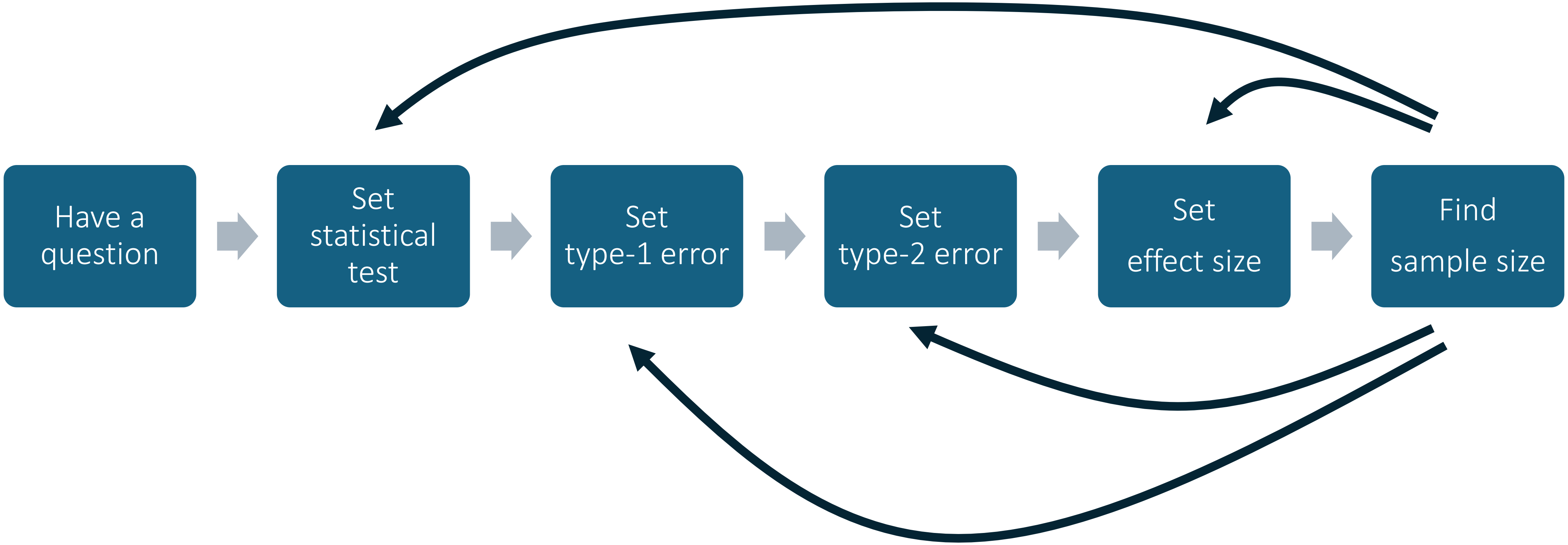
- Increase generalizability

  A justified sample size is easier to reproduce

- Preregistrations, registered reports and grant request it

  Helps you get funding and publications

The lifespan of a-priori power calculations



DON’T DO THIS

Practical 1: Understanding parameters of power calculations

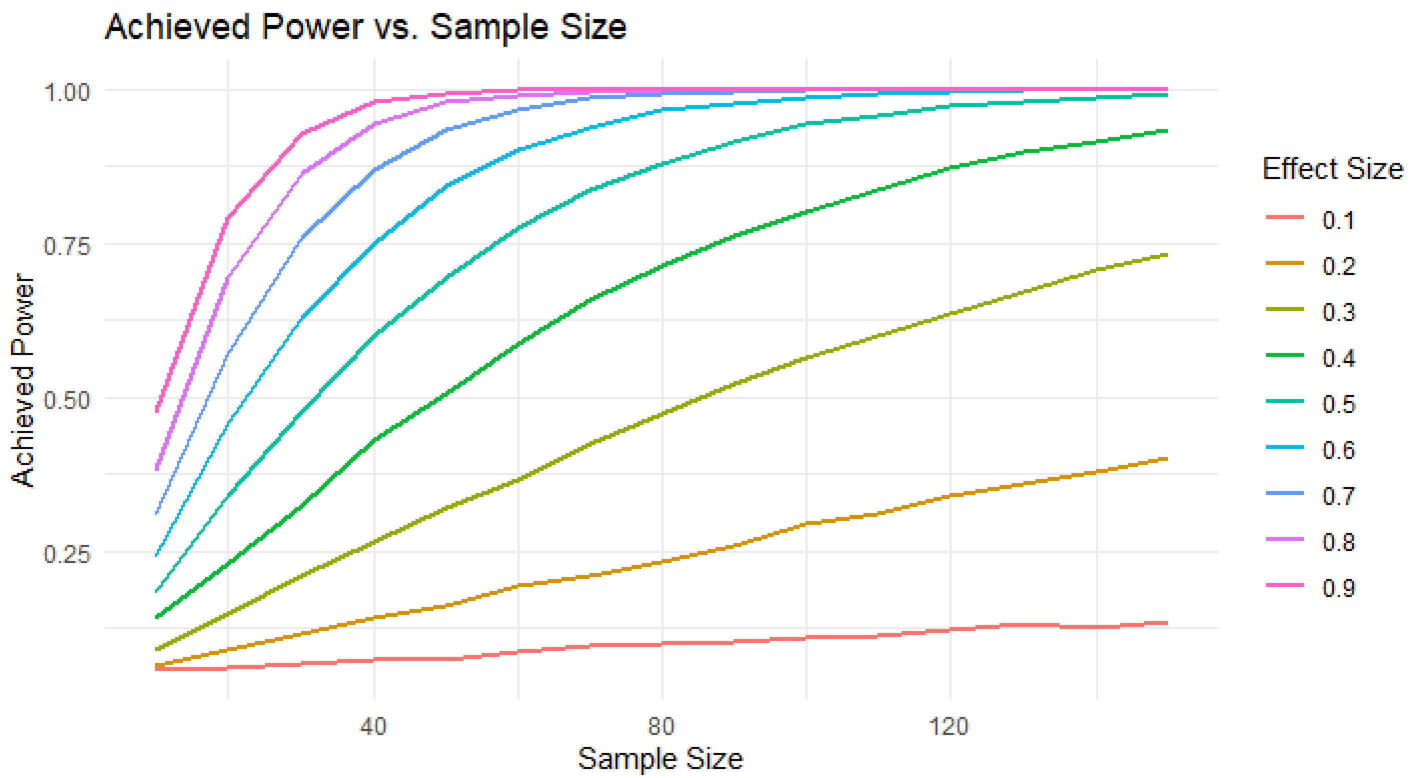
Paris wants to determine if there is a significant difference in systolic blood pressure between patients on a new antihypertensive drug and those on a placebo.

- Independent : Treatment group (new drug vs. placebo)
- Dependent: Systolic blood pressure (continuous variable)

- Open R.
- Load ‘ex_01_params.R’
- Using the exist parameters
- Explore how different params affect what sample size John needs.
- Check what happens if you use a paired sample test (before / after administration of drug)

Effect of parameters on power calculations

- Alpha level ↑ Sample ↑
- Beta level ↑ Sample ↑
- Effect size ↑ Sample ↑
- Design ~ Depends ....

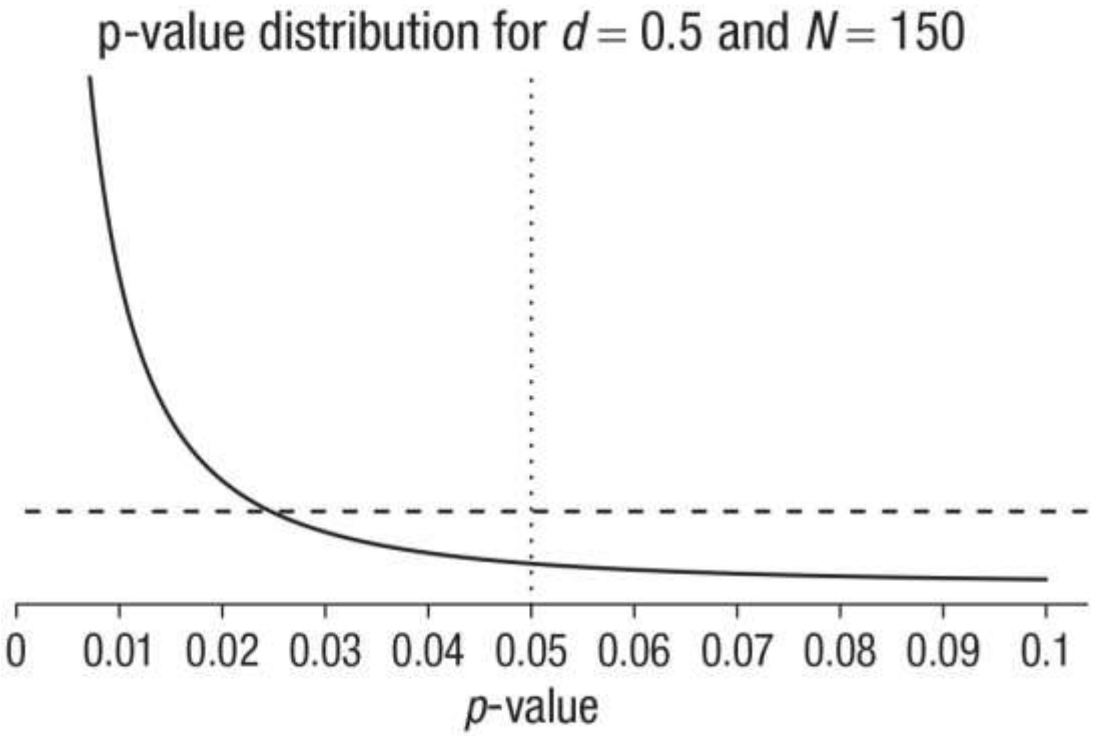


Why parameters affect us

- Alpha
  - The more results we consider false positives, the less studies we end up accepting as significant
- Beta
  - The more studies we consider false negatives, the less studies we end up accepting as significant
- Effect Size
  - The stronger the effect size, the smaller the sample size to detect it
  - The larger the difference in mean, the smaller size we need to reach it
  - The larger the variance of the means, the higher sample sizes we need
- Type of Test
  - Paired / Nested designs allow us to reduce the variance of our estimates
- Type of predictions we make
  - One sided tests require a smaller difference in means to reach significance, making the sample size necessary smaller

Selecting alpha

- 0.05 vs 0.02 vs 0.01
- The higher the statistical power of a test, the less likely it is to observe relatively high *p*-values (e.g., *p* > .02).
- If H0 = True, then p values are uniformly distributed (dotted line). So for high-powered studies, .01< p < .05 might be evidence for null (????)



Selecting effect sizes or parameters

- Pilot study -> requires high N of participants to be meaningful
- Heuristics (Simmons, 2011) -> use at least 50 (suspicious)
- Previous effect sizes -> approximation of relevant study
- Previous parameters -> approximation of relevant study
- Smallest effect size of interest (SESOI) -> theoretical minimum

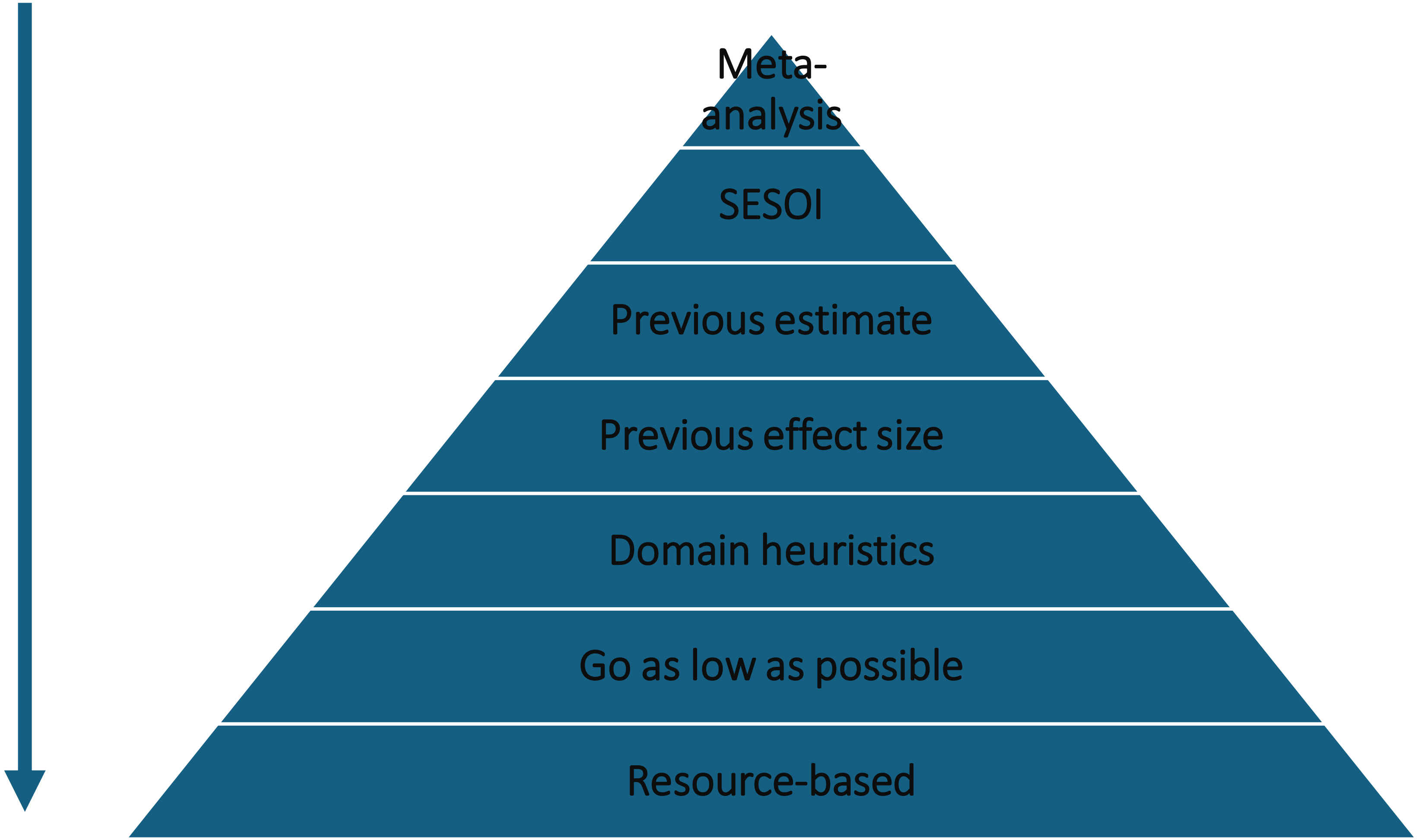
Simmons, 2011
Lakens, 2019
Simonsohn, 2015

SESOI

- Minimal Statistically Detectable Effect
  - Don’t ask what effect another study found, ask what is the minimum they could have found
  - If a study has found an effect, it might have been **inflated**. Being suspicious, I accept that an effect exists, but I will power for the minimal possible effect
  - Driven by sample size
    - If sample is large, your minimal effect is small
    - If sample is small, your minimal effect is large
  - Example: A study found a Cohen’s D = .6. With 50 participants in each group, the minimal possible effect size is .4
- Small telescopes
  - Don’t ask what effect another study found, assume what it would find if it was underpowered
  - Gamble: I give you 2:1 odds that the study will not have a result (33% power)
  - If that is true, what size can you find?
  - Driven by sample size
    - If sample is large, your minimal effect is small
    - If sample is small, your minimal effect is larger

Simonsohn, 2015
Laken, 2017

Effect size selection



Meta-analysis

SESOI

Previous estimate

Previous effect size

Domain heuristics

Go as low as possible

Resource-based

Easy / Arbitrary

Simonsohn, 2015
Laken, 2017

Practical 2: Selecting SESOI

- Consider a study where 25 smokers evaluated how many cigarettes they smoke per day before and after exposure to scare images on the health detriments of smoking.
- You want to replicate it. Find the smallest effect size of interest and estimate how many people you need.

- Load ‘ex_02_sesoi.R’

Software vs Simulations

	Software	Simulations
Easy to use	Maybe	Maybe
Analytical Solution	Yes	No
Any test	No	Yes
Reproducible	Maybe	Yes
Intuitive	No	Yes

Basic Simulation Structure

- For every sample size
    - Create storage for statistic
    - Create storage for p values
    - For every simulation
        - Simulate dataset with required parameters
        - Add noise to the dataset
        - Run statistical test of interest
            - Extract simulated statistic
            - Extract significance
        - Store results
    - Count how many tests were significant / How many simulations you run
    - Congratulations! You estimated power for tested sample size
- Plot power calculation curve (x axis = samples, y axis = achieved power)

Practical 3: Running your first simulation

Paris wants to calculate if smokers have higher rates of anxiety.

Previous literature suggests an effect size Cohen' D=.6 (suspicious for psychology). He wants to replicate this study. Help him!

- Set control group mean = 0
- Assume a noise level of SD=1
- Specify a=.01, b=.05
- Test sample size from 10 to 200, in increments of 5
- Run 500 simulations per sample size

Hierarchical Designs

- Factorial Designs
- Random effects (multiple measurements per run / subject)
- Random slopes (multiple measurements per run / subject)

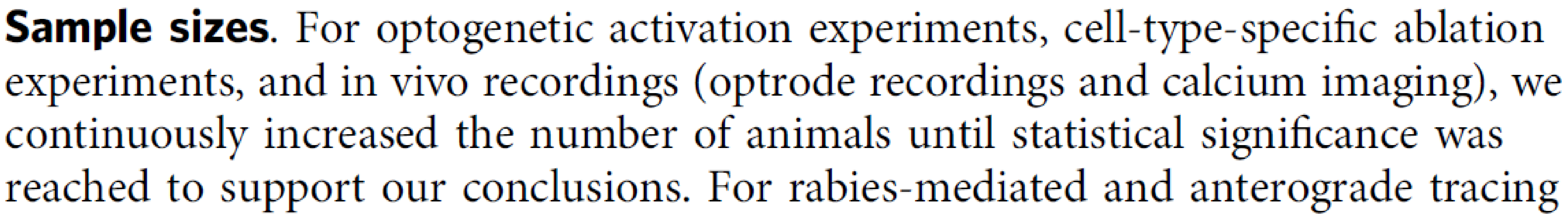
- No analytical solution
- Yet easy to conceptualize using simulations

Basic Hierarchical Simulation Structure (and many more …)

- For every sample size
    - Create storage for statistic
    - Create storage for p values
    - For every simulation
        - Simulate dataset with required parameters
        - Add noise to the dataset
        - Run statistical test of interest
            - Extract simulated statistic
            - Extract significance
        - Store results
    - Count significant / tests run
    - You estimated power for tested sample size
- Plot power curve

- For every sample size
    - Create storage for statistic
    - Create storage for p values
    - For every simulation
        - Simulate dataset
            - For every subject
                - Simulate noisy, multi-trial
                - Add constant noise per subject.
        - Run statistical test of interest
        - Extract simulated statistic
        - Extract significance
        - Store results
        - Run an equivalence test to test whether H0 stands
    - Count significant / total tests you run
    - Congratulations! You estimated power for tested sample size
- Plot power curve

Optional Stopping vs Sequential Sampling

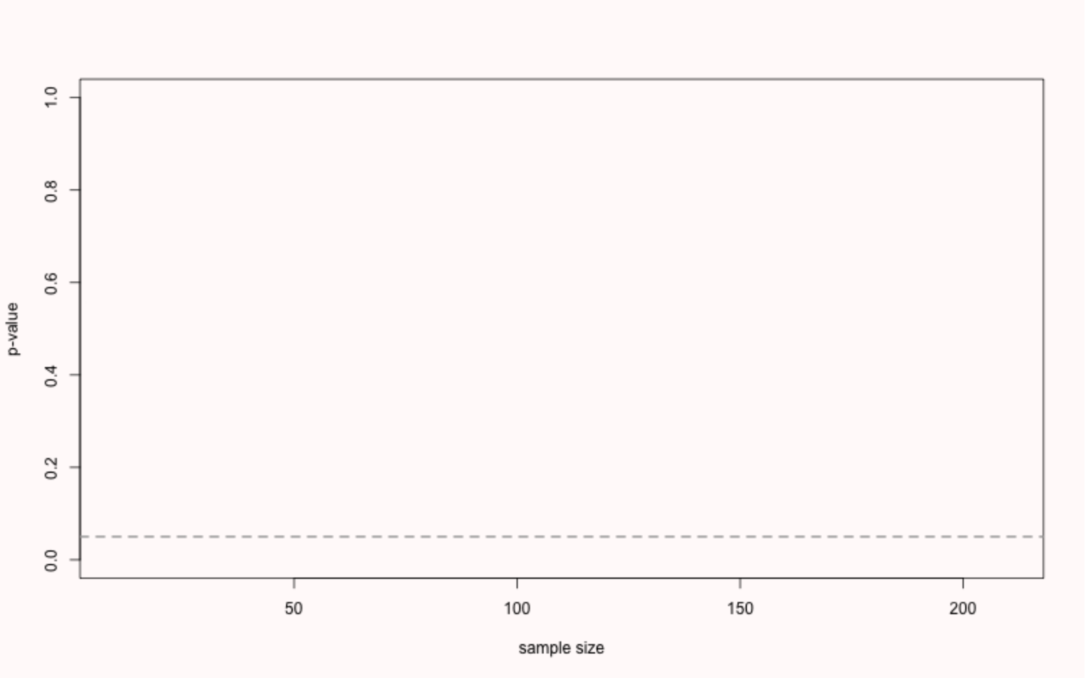


Sample sizes. For optogenetic activation experiments, cell-type-specific ablation experiments, and in vivo recordings (optrode recordings and calcium imaging), we continuously increased the number of animals until statistical significance was reached to support our conclusions. For rabies-mediated and anterograde tracing

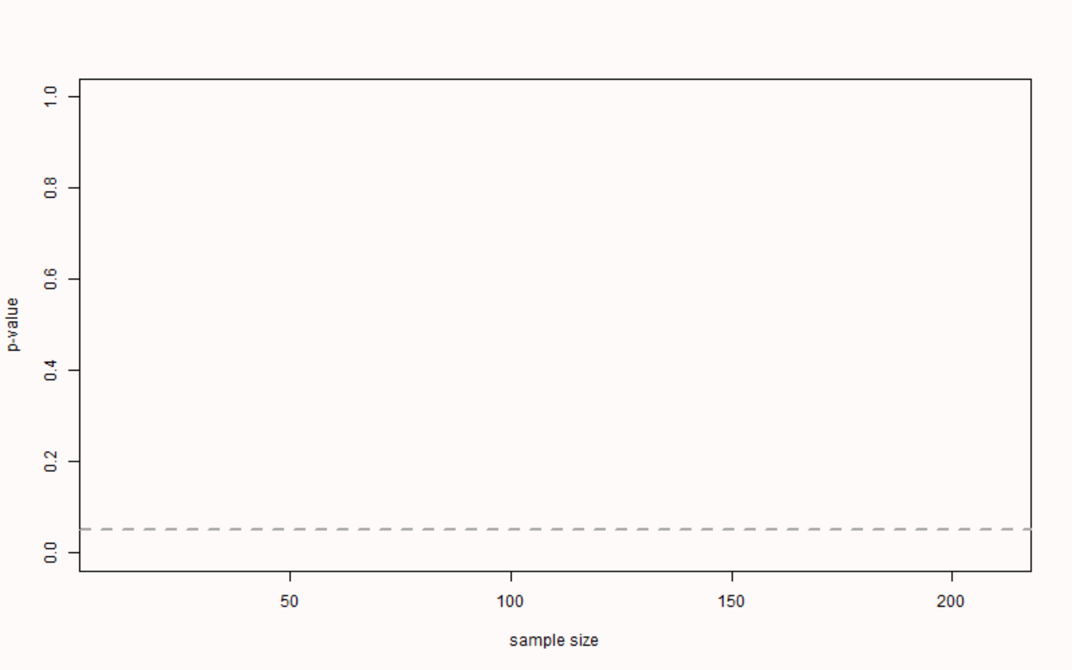
Heart is the right place …
Is it possible to do ?

Optional Stopping vs Sequential Sampling

Optional Stopping: Cohen’s D=0



Sequential Sampling: Cohen’s D=.3

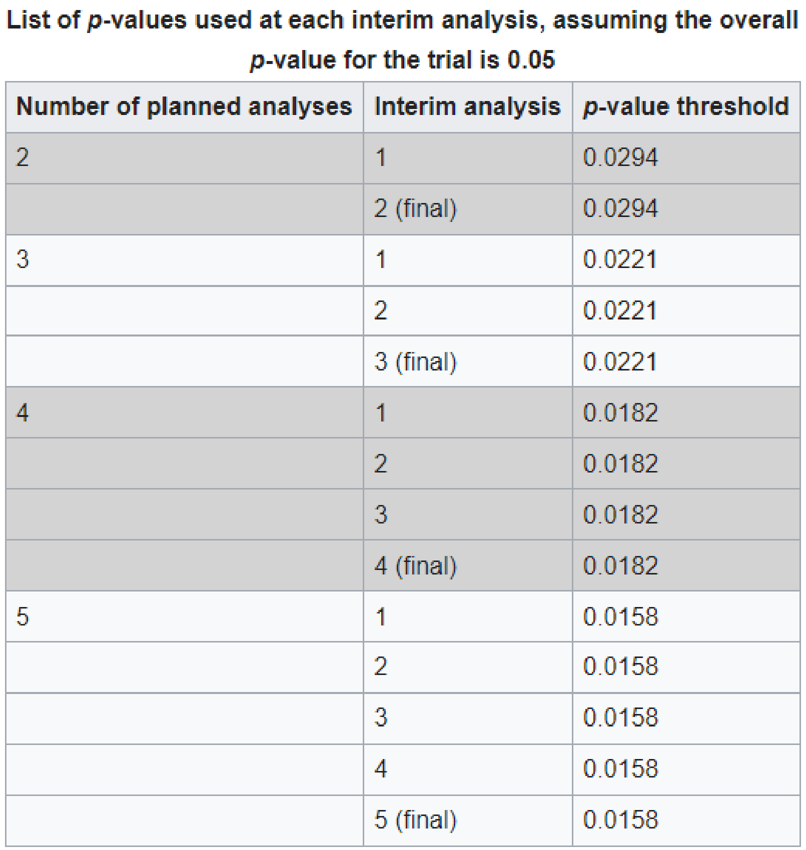


Lakens, MOOC

Sequential Sample Size Estimation

- You power analysis said you need 150 people ☹
- You can do interim analysis at intervals (50,100 participants) *IF* you control for your Type 1 error.
- You have a set amount of error budget (.05) and you need to spend it across all your interim analysis.
- GOAL = Find how to split the bill

Pocock algorithm



List of *p*-values used at each interim analysis, assuming the overall *p*-value for the trial is 0.05

Number of planned analyses	Interim analysis	*p*-value threshold
2	1	0.0294
	2 (final)	0.0294
3	1	0.0221
	2	0.0221
	3 (final)	0.0221
4	1	0.0182
	2	0.0182
	3	0.0182
	4 (final)	0.0182
5	1	0.0158
	2	0.0158
	3	0.0158
	4	0.0158
	5 (final)	0.0158

Jennison & Turnbull, 2000
Wassmer & Brannath, 2016

WP3: Help Paris create the best study ever

Paris is a psychologist who studies how reaction times track arousal level. Going over the literature, he found that a previous study with 50 well-rested participants in one group and 50 sleep-deprived participants in the other. There was an effect size of Cohen’s D = .4, where well-rested participants were faster in detecting a familiar faces compared to sleep deprived participants.

- 1. Find the smallest effect size of interest
- 2. Simulate how many people Paris needs to achieve a power of .95 at an error rate of .01 (or just use an analytic solution …).
- 3. Find a way to adjust the error rate so that Paris can acquire data with 4 interim analysis.

Tutorial Inspirations

- Improving your Statistical Inferences by Daniel Lakens
- Improving your Statistical Questions by Daniel Lakens
- Statistical Rethinking by Richard McElreath
- Being bullied at Stack Overflow and Cross Validated

Time for you to design the best study ever