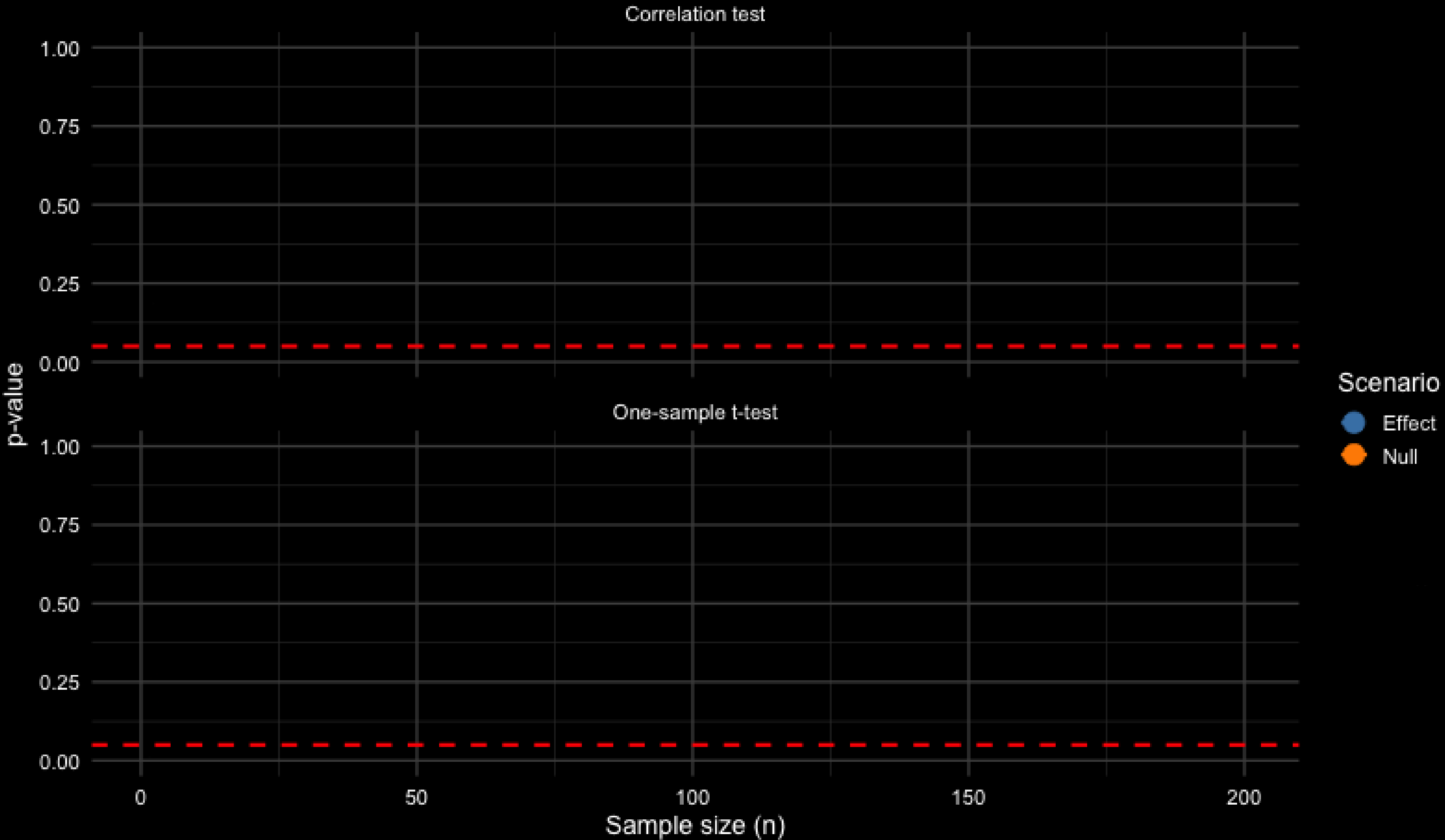
Inference, Statistical Power and Experimental Design

Robin Welsch
robin.welsch@aalto.fi
human-ai-interaction.com


Correlation test
One-sample t-test
p-value
1.00
0.75
0.50
0.25
0.00
0
50
100
150
200
Sample size (n)
Scenario
Effect
Null

We can do better!

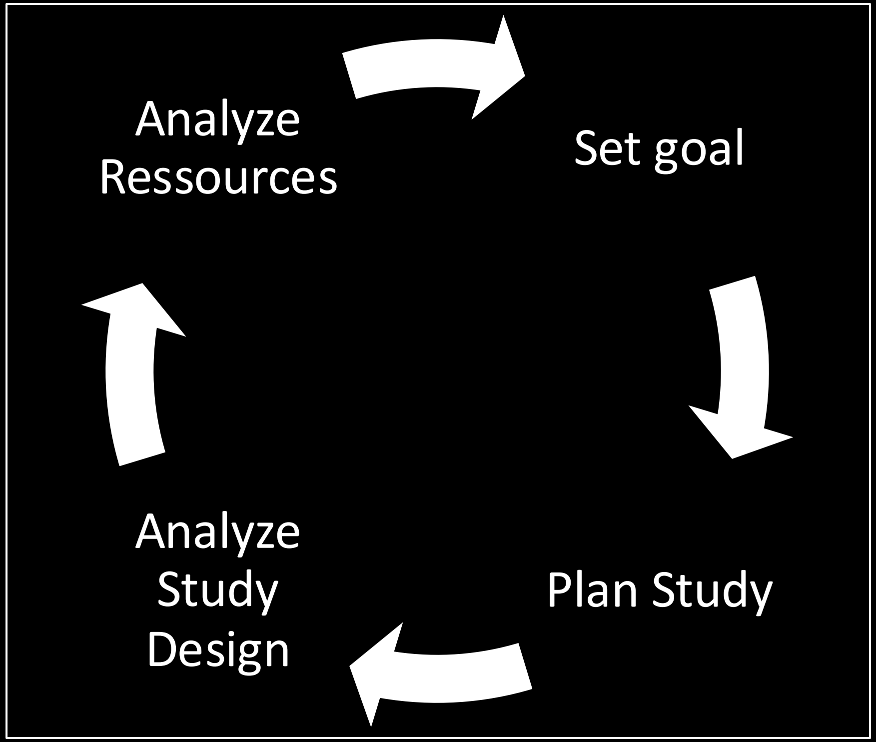
Let‘s power up!

Outline

1. Sample size planning and experimental design for statistical power
2. Determining a minimal effect size of interest
3. Thinking beyond *p*-values
4. Simulating data to estimate statistical power

Checklist for sample size planning of experimental studies

- ❑Is the study's goal defined?
  - ❑ Parameter estimation
  - ❑ Exploration
  - ❑ Hypothesis testing → Power analysis
- ❑Ressources known?
  - ❑ Time
  - ❑ Money
  - ❑ Personnel
  - ❑ no limits → simple power analysis
  - ❑ …
- ❑Decision criterion for study success
  - ❑ *p-value* → Poweranalysis
  - ❑ Bayes Factor
  - ❑ Interval
  - ❑ Data collection as an end in itself
  - ❑ Amount of information (e.g., for qualitative studies)

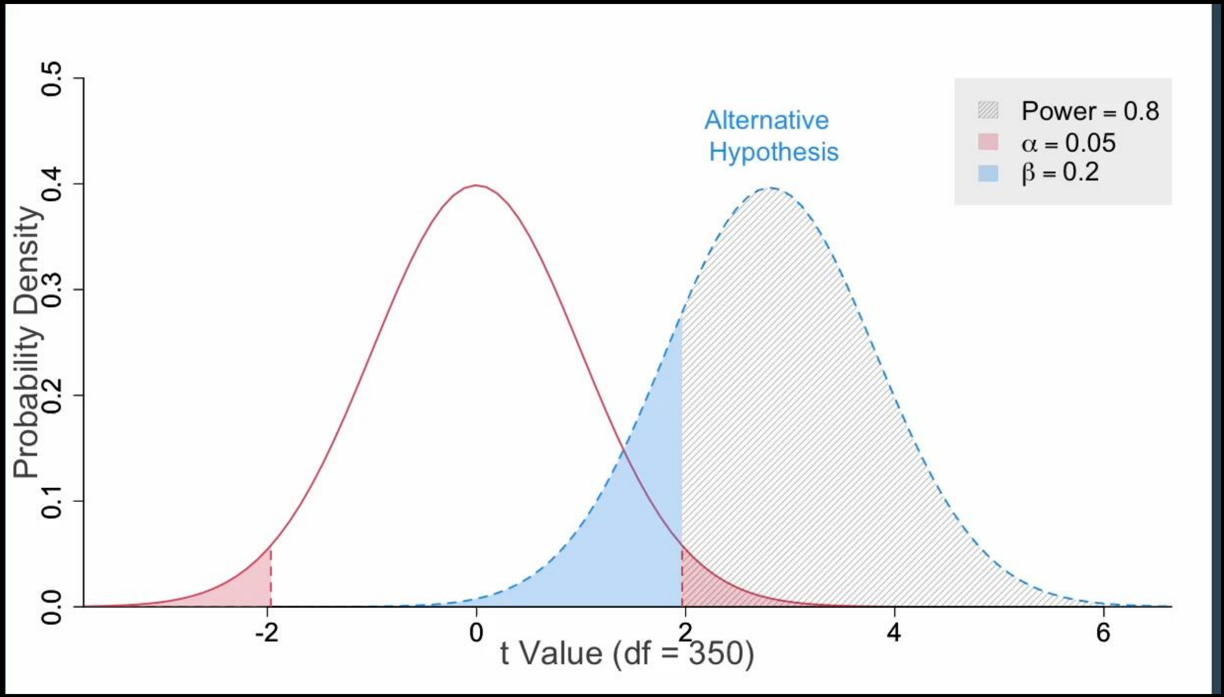


Recap: Power (test strength) is central to hypothesis testing in experiments.

The goal is to achieve high power without wasting resources → reaching 100 % power would require infinitely many data points.

- Power is the ability to statistically detect an effect that truly exists.
- Higher power reduces the risk of committing a Type II error.
  - (incorrectly accepting the null hypothesis)

  → Studies with low power and no significant results are not informative.
- Variables influencing power (aiming for >80%)
  - Typ 2 error (alpha) → set to 5 %
  - **Sample Size**
    - set before the study
  - **Effect Size** (z.B. cohen's *d*)
    - has to be estimated



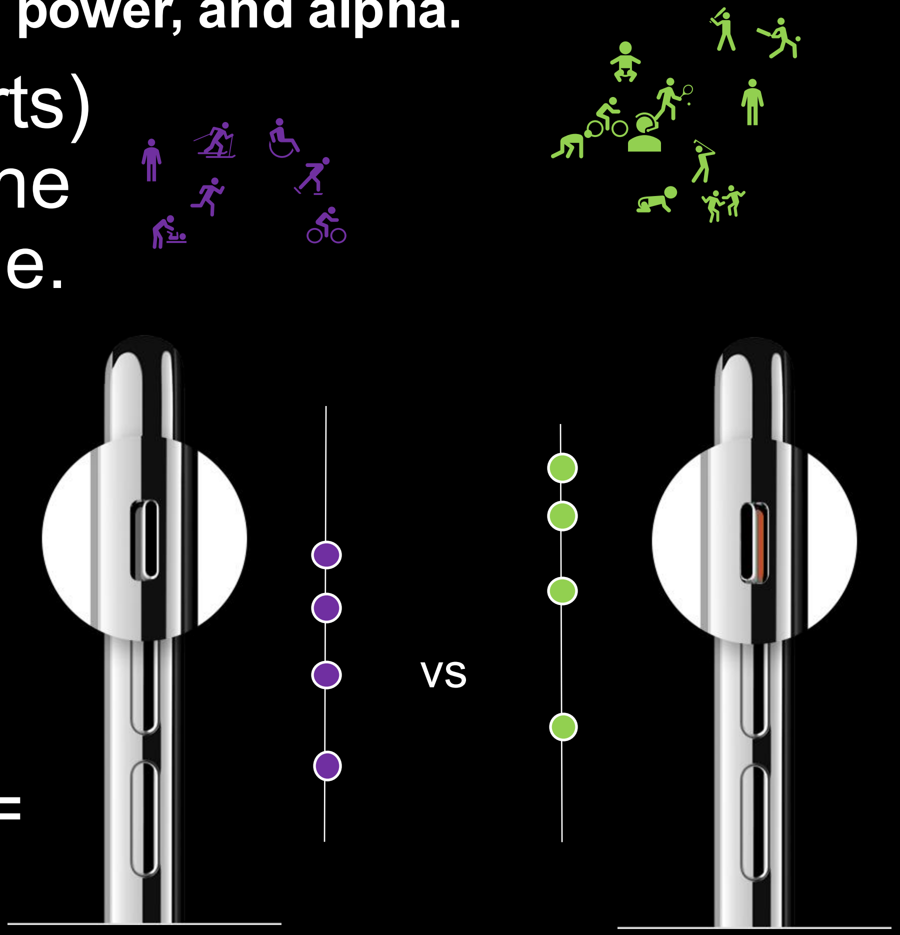
How do I plan my study to find an effect*?

*,if there is a true effect in the world

A priori: Power analysis for a between-subjects study

Determination of N given a specified minimum effect size, power, and alpha.

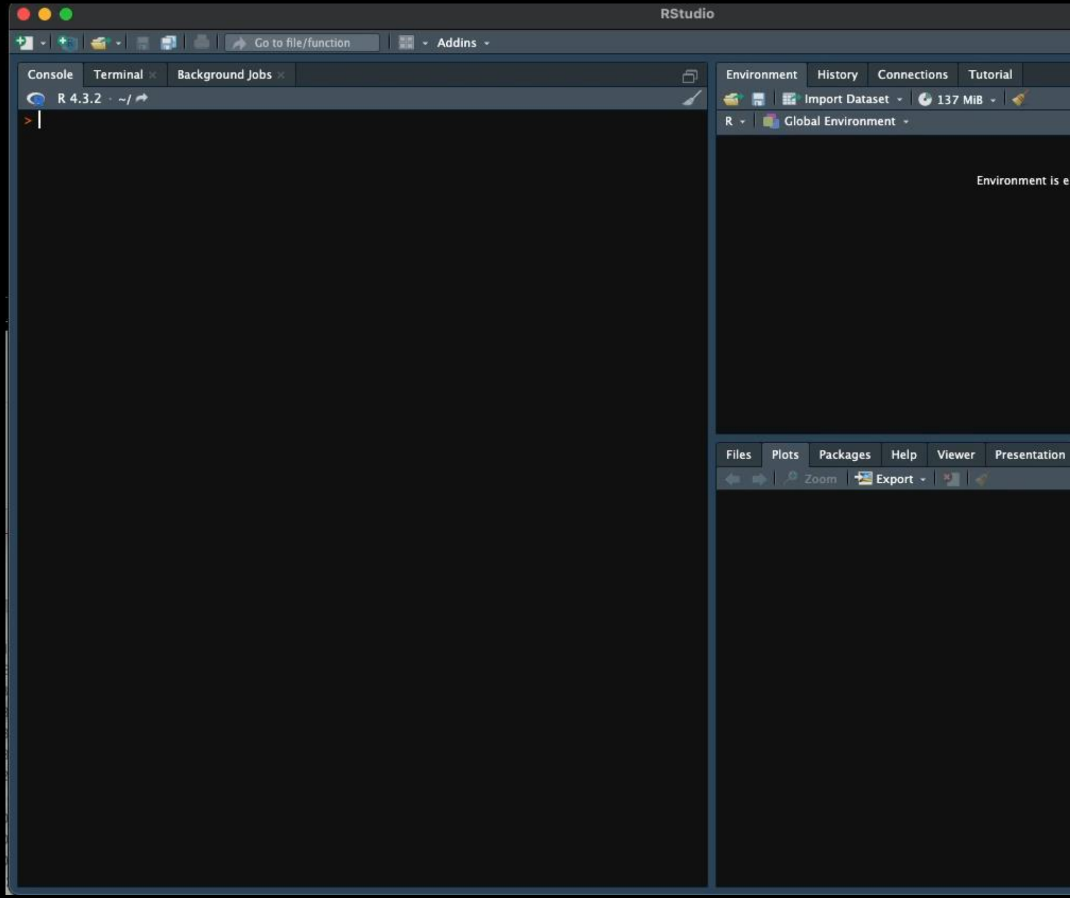
- Hypothesis: Notification tones (audible alerts) lead, on average, to longer daily smartphone usage (in minutes) compared to silent mode.
- Assumptions *t*-test:
  - Comparison of two groups
  - alpha = 5%
  - Power = 80%
  - $SD_{between}$ = 40 Minutes
- Minimal effect size of interest→ d = 10/40 = .25



A priori: Power analysis for a between-subjects study

Determination of N given a specified minimum effect size, power, and alpha.

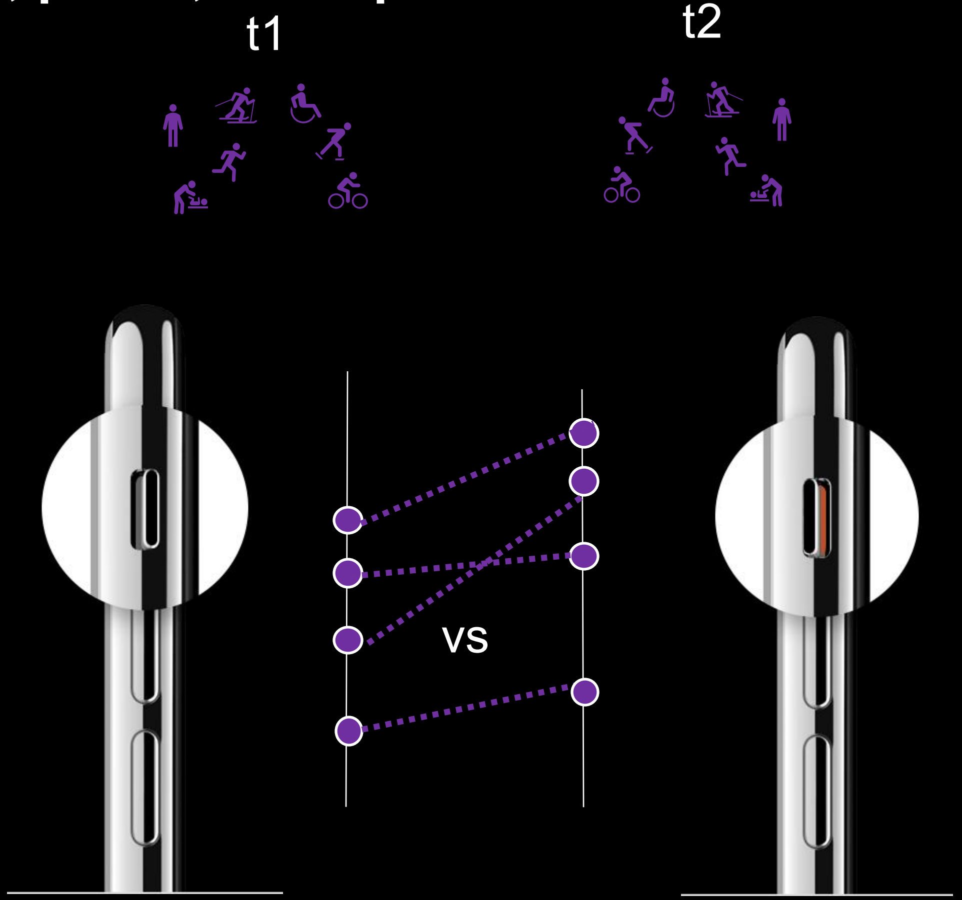
- Assumptions *t*-test:
  - Comparison of two groups
  - alpha = 5%
  - Power = 80%
  - SD_{between} = 40 Minutes
- Minimal effect size of interest→ d = 10/40 = .25
- *n* for each group = 252



A-Priori: Poweranalyse für Within-Subjects Studie

Determination of N given a specified minimum effect size, power, and alpha.

- Comparison of smartphone usage (in minutes) with notification tones versus silent mode **within** a single sample
- Assumptions:
  - Comparison within the sample
  - alpha = 5%
  - Power = 80%
  - SD_{within} = 15 minutes



A-Priori: Poweranalyse für Within-Subjects Studie

Control of person-specific nuisance variables

- Assumptions:
  - Comparison within the sample
  - alpha = 5%
  - Power = 80%
  - SD_{within} = 15 minutes
- Minimal effect size of interest 10 minutes→ d = 10/<u>15</u> = .67
- 20 participants are enough

```
> pwr::pwr.t.test(d=.67,sig.level = .05,power=.80,type="paired",alternative = "two.sided")

     Paired t test power calculation 

              n = 19.49243
              d = 0.67
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

NOTE: n is number of *pairs*
```

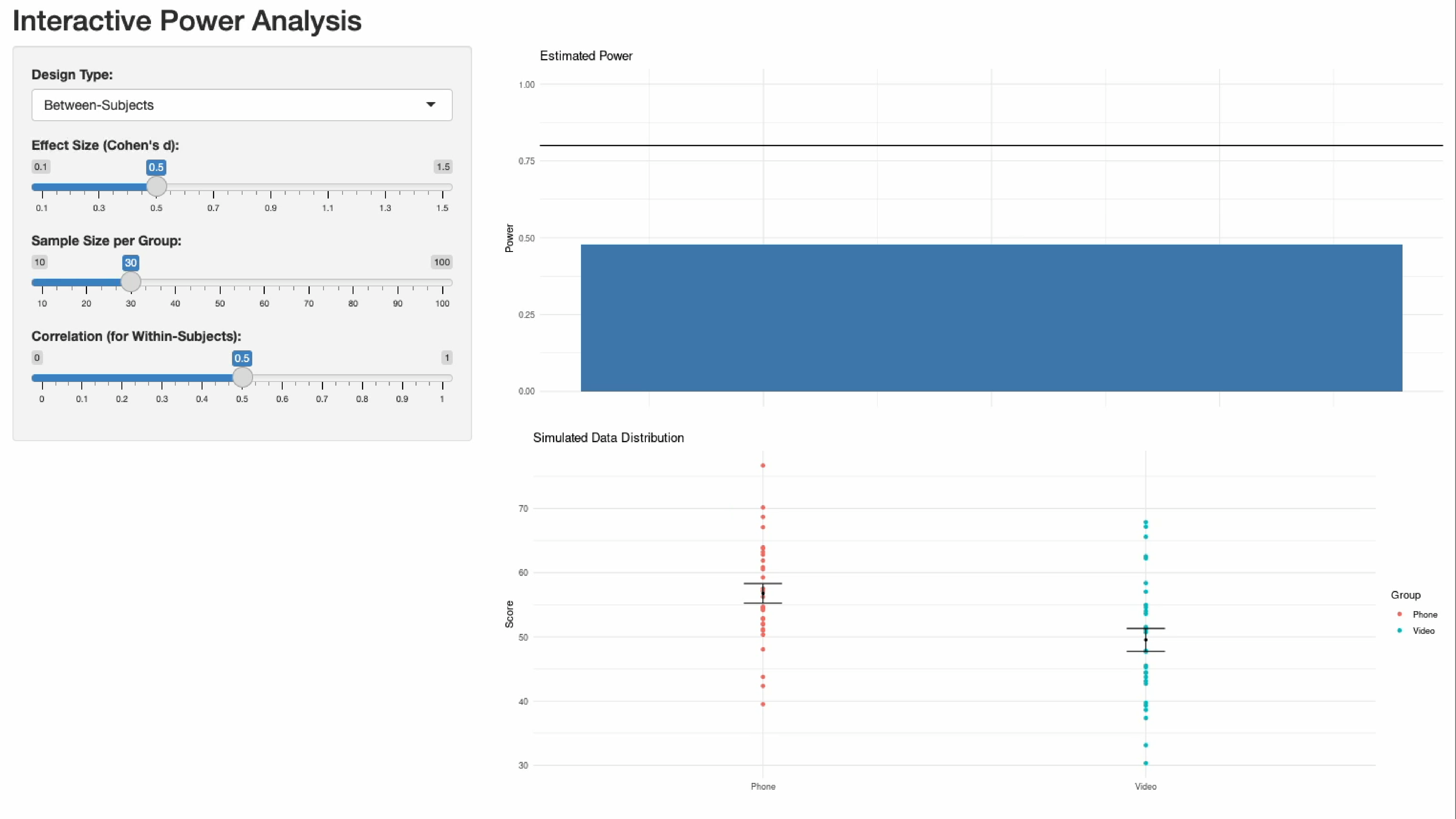
Interactive Task

Go to: https://robwel.shinyapps.io/power_interactive/

Adjust the effect size, sample size, design, and, if applicable, the correlation to achieve over 80% power..

1. You are investigating a medium-sized effect (d = .5) in a between-subjects design. How many participants do you need at minimum?
2. You are investigating a medium-sized effect (d = .5) using a within-subjects design with a limited sample of 50 participants. How good must your measure be—that is, what minimum correlation is required?

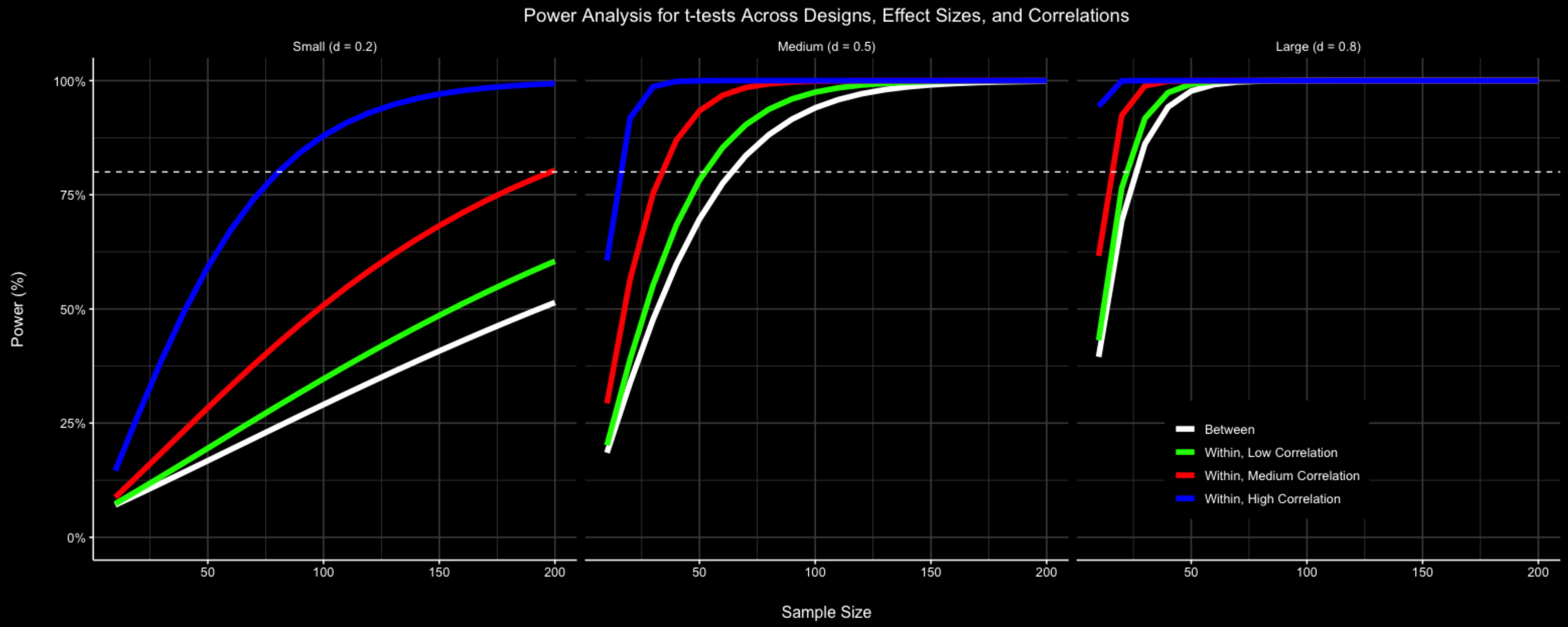
→ How do correlation, sample size, and effect size relate to statistical power?


Interactive Power Analysis
Design Type:
Between-Subjects
Effect Size (Cohen's d):
0.1
0.5
1.5
0.1 0.3 0.5 0.7 0.9 1.1 1.3 1.5
Sample Size per Group:
10
30
100
10 20 30 40 50 60 70 80 90 100
Correlation (for Within-Subjects):
0
0.5
1
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
Estimated Power
Power
1.00
0.75
0.50
0.25
0.00
Simulated Data Distribution
Score
70
60
50
40
30
Phone
Video
Group
Phone
Video

Whats the relation of power, effect size and correlation within subjects?

Sample size for within- and between-subjects designs

Higher power in a within-subjects design: due to reduced variability from individual differences, since each participant serves as their own control. Power in a between-subjects design: may require larger sample sizes to achieve similar power levels, because variability between individuals can mask effect sizes.



Interim conclusion I

Power analysis for classical statistical inference

- Getting the right sample size and design is important
- Power analysis can calibrate Type-II error
- We can adjust the design of a study to increase power

Questions?

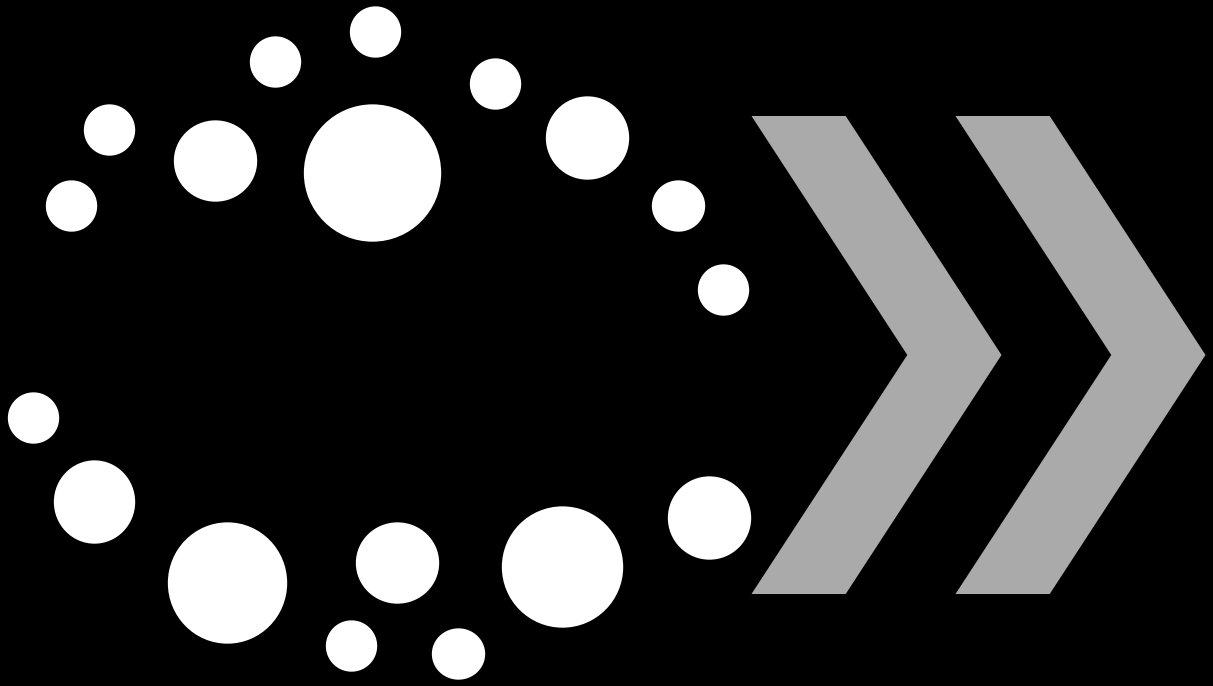
Limitations of power analysis

- focuses on classical statistical inference with p-values
- Complex study study designs(e.g., RCT)
  - Analytical solutionis not possible
- Ressources are unrealistic for small effects
  - Use repeated measures
- Effect size has to be estimated
  - Hard in novel research areas

Limitations of power analysis

- focuses on classical statistical inference with p-values
- Complex study study designs(e.g., RCT)
  - Analytical solutionis not possible
- Ressources are unrealistic for small effects
  - Use repeated measures
- <u>Effect size has to be estimated</u>
  - Hard in novel research areas

How do I estimate the size of the effect for power analysis?



Smallest effect of interest

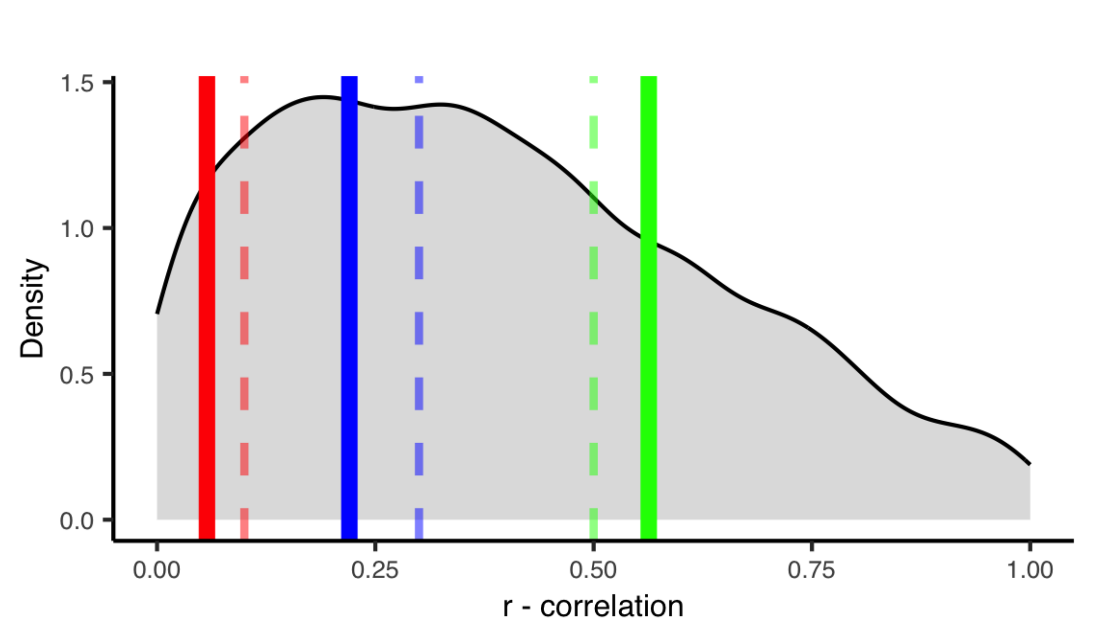
Identify hypothesis with the smallest effect

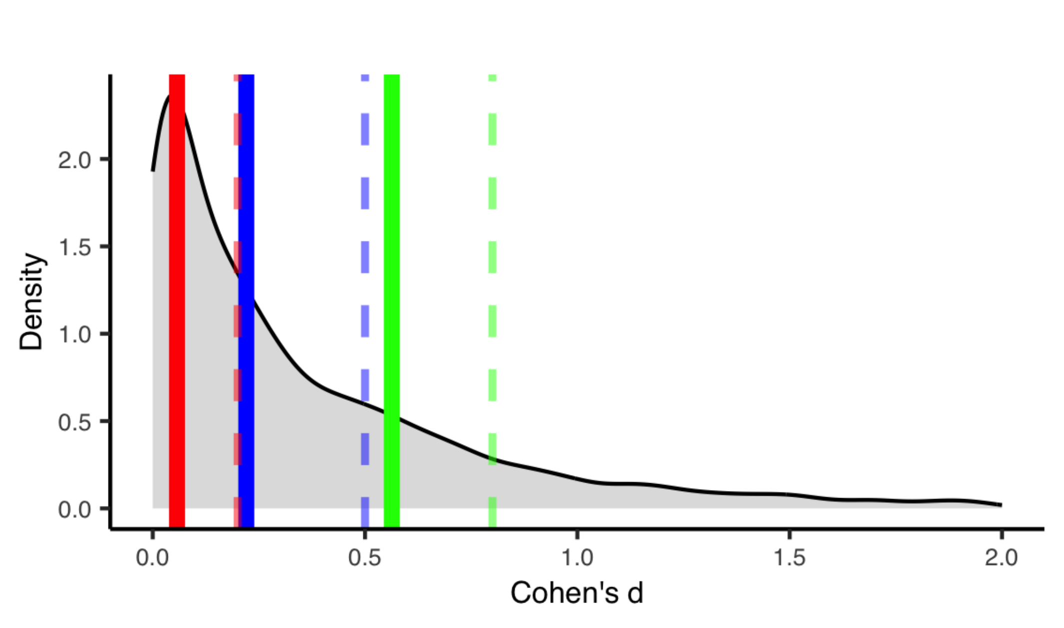
Run power analysis for statistical model that tests this effect

My data

>1600 data points

CHI has smaller effect sizes than cogntive science





Robin‘s rule of thumb

1. Find published effect sizes in Cognitive Science and HCI
2. <u>Halve the effect size – be conservative (accounts for publication bias)</u>
3. Run power analysis
4. See if you can still run the study within the scope of your resources

Interim conclusion II

Minimal effect of interest

- Expecting too large effects is unrealistic
- Prior studies can guide the way
- We can use conservative estimates in the absence sufficient information

Questions?

Limitations of power analysis

- focuses on classical statistical inference with p-values
- Complex study study designs(e.g., RCT)
  - Analytical solutionis not possible
- Ressources are unrealistic for small effects
  - Use repeated measures
- Effect size has to be estimated
  - Hard in novel research areas

Thinking beyond *p*-values

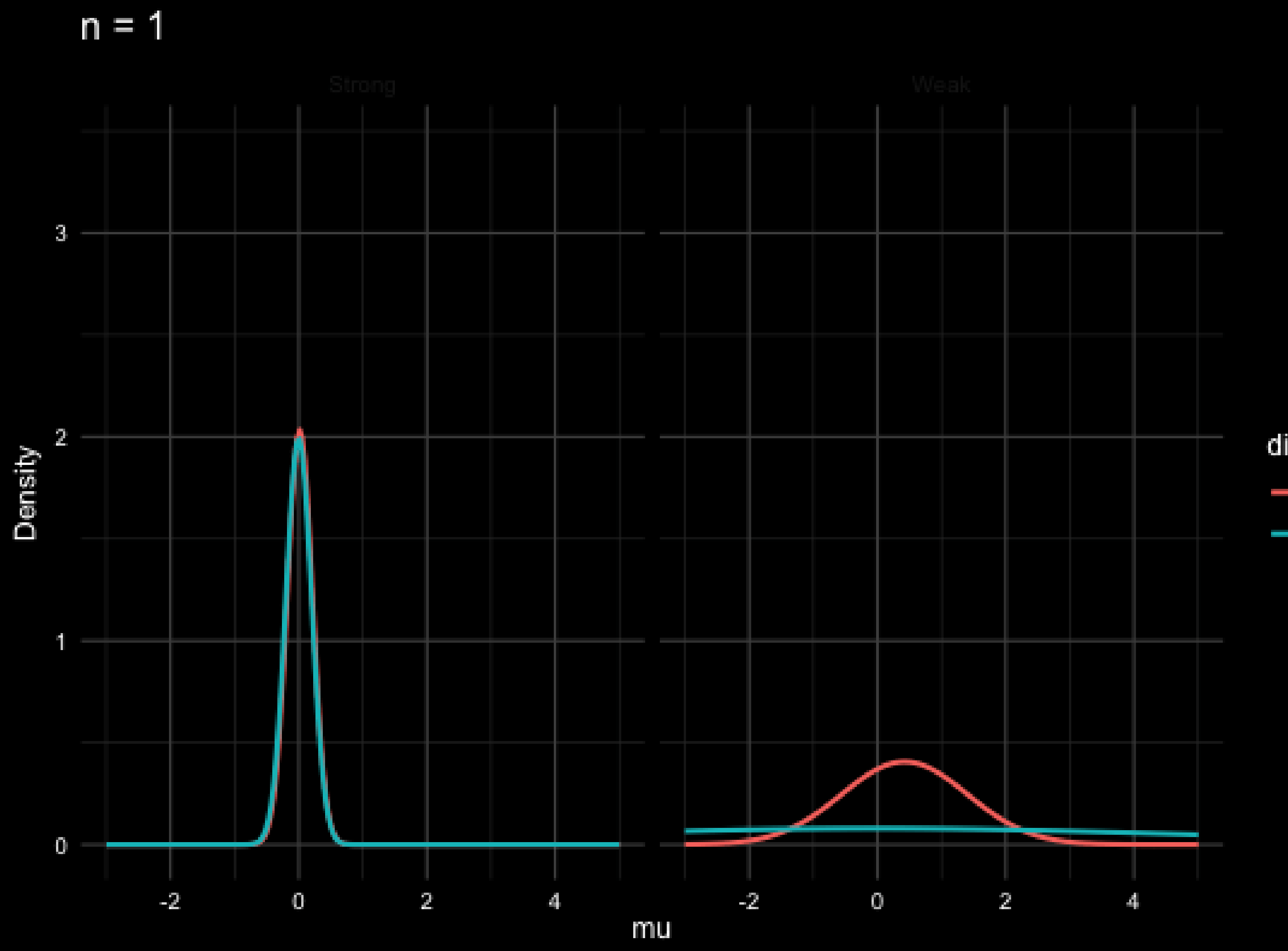
Bayesian Learning

Prior → data → Posterior

Bayesian and classical inference comparison

Recap 1

- Classical power analysis isn't directly applicable to Bayesian methods
- Bayesian statistics
  - Bayes Rule: Posterior = Likelihood×Prior
    - **Prior**: Initial beliefs about parameters before seeing data
    - **Likelihood**: Probability of observing the data for a given parameter value
    - **Posterior**: Updated beliefs after incorporating the data
  - Produces full distributions, not just point estimates → degree of hypothesis confirmation after seeing the data
  - Allows direct probability statements (e.g. "95% chance effect > 0")
  - Enables evidence quantification



Bayesian and classical comparison

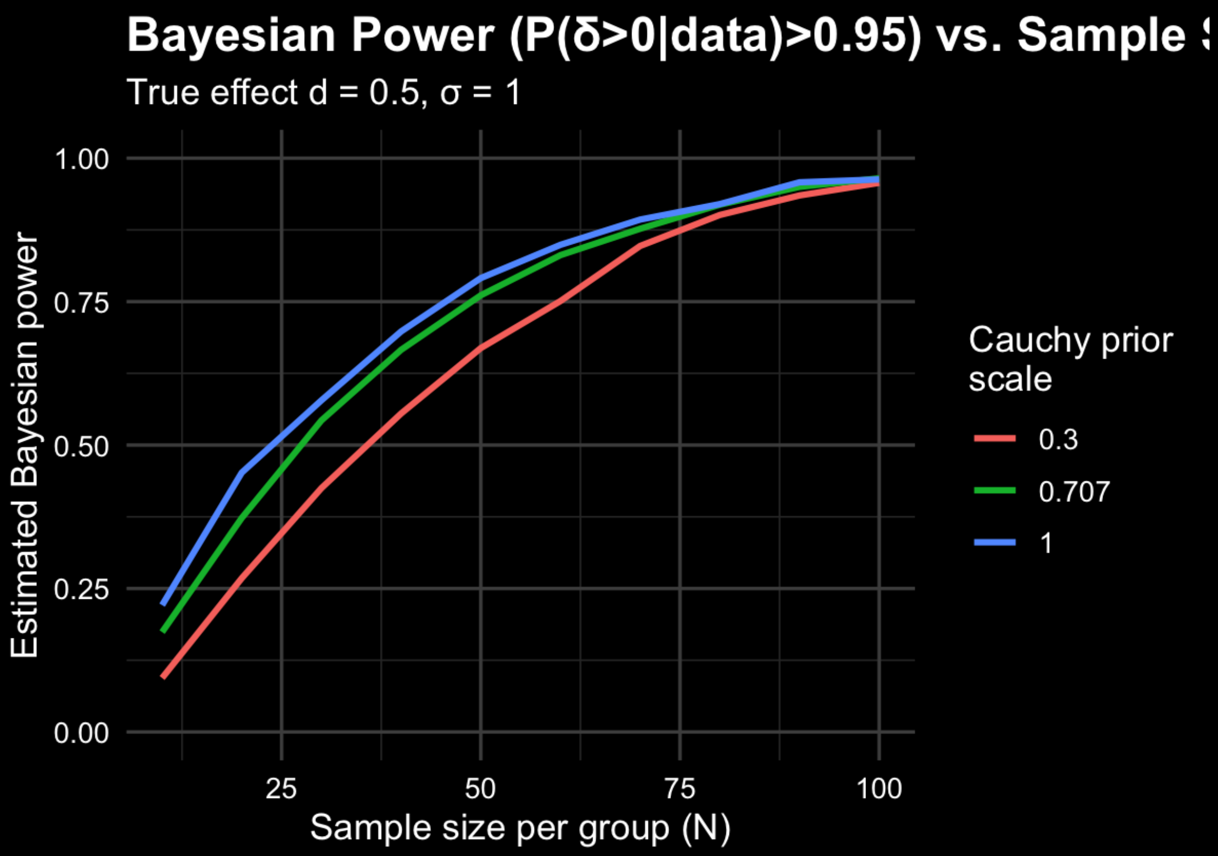
Recap 2

- **Question Asked**
  - **Classical (Frequentist):** “If the null hypothesis is true, how surprising are our data?”
  - **Bayesian:** “Given the data we observed, how plausible is each hypothesis?”
- **Core Probability**
  - **Frequentist:** P(data|H0) → Compute p-value to assess rarity under H_0.
    - Reject H_0 if p < α (fixed Type I error rate).
    - “Assuming no effect, there’s only a 1% chance of seeing data this extreme.”
    - Power analysis checks how likely it is to detect a true effect given the threshold
  - **Bayesian:** P(H|data) ∝ P(data|H) P(H) → Derive posterior probability of hypotheses.
    - Declare support if posterior probability > threshold (e.g. 0.95) or Bayes factor above cutoff (e.g. BF_{10} > 3) but also gradual statements are possible
    - “Given these data, there’s a 95% chance the effect is positive.”
    - Bayesian assurance/power simulates how much data is needed to learn enough from the data to make a decision

Bayesian Power?

T-test with different priors how quickly we reach decision criterion of posterior >.95

- As the curves rise, you see how quickly each prior reaches our 95 % posterior-probability cutoff:
- **Steeper curve** → less data needed: the stronger the prior (larger Cauchy scale), the faster P(δ>0) crosses 0.95.
- **Rightward shift** → more data needed: with weaker priors or smaller N, you need larger samples to hit 95 % certainty.



Interim conclusion III

Power analysis for bayesian statistical inference

- We can update the prior based on simulated data to see how updating informs our decision
- Priors are critical for „power“ of bayesian testing
- Gradual statements are also possible effects do not exist or not but after seeing the data, the probability of effect existence is xx%

Questions?

Limitations of power analysis

- focuses on classical statistical inference with p-values
- Complex study study designs(e.g., RCT)
  - Analytical solution is not possible
- Ressources are unrealistic for small effects
  - Use repeated measures
- Effect size has to be estimated
  - Hard in novel research areas

Using simulation for estimating power in classical inference

Tutorial

- *t*-test (simulation and analytical solution)
- Wilcoxon-test simulation and kruskal-wallis ANOVA
- Maybe: Nested repeated measures data

Questions?