Simulation-Based Power Analysis in (Generalized) Linear Mixed Models

Wokshop with the Open Science Initiative Frankfurt

Leah Kumle

24.03.2021

Today's roadmap

- **Introduction**
  - What is power and why do we need it
  - (G)LMMs and power
  - Simulation-based power analyses R
- **Power analyses in (G)LMMs**
  - Choosing simulation parameters
  - R-package mixedpower R
- **Questions, discussion, ..**

Have you ever conducted a power analysis for one of your studies?

Have you ever conducted a power analysis for one of your studies?

If yes – what was your goal? How did you do it?

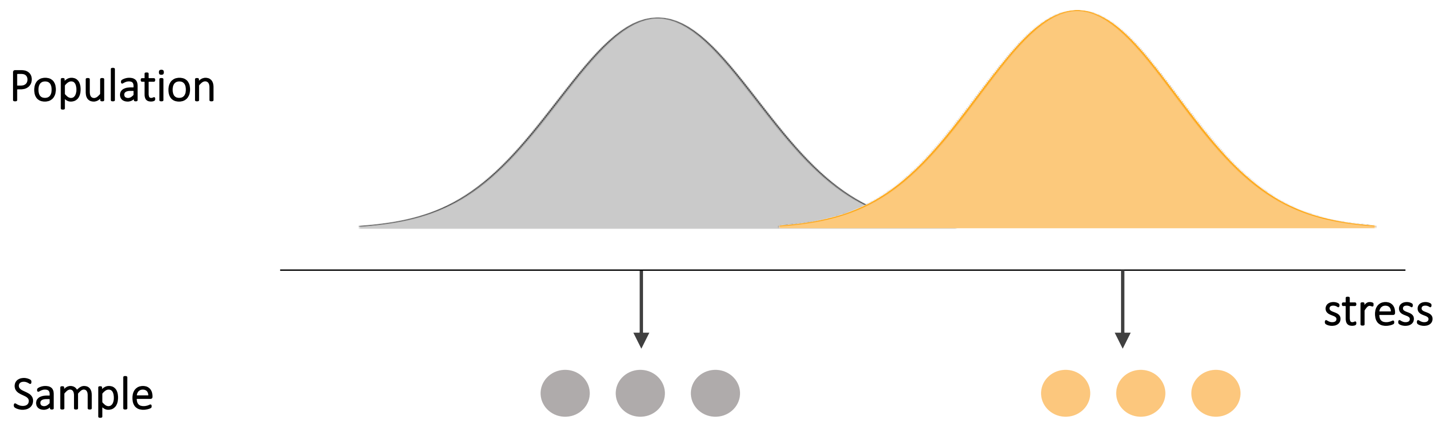
If not – what stopped you? What were difficulties in the process?

What is power and why do we need it?

Submission guidelines of Behavior Research Methods:

> ***"It is important to address the issue of statistical power***. [...] Studies with low statistical power often produce ambiguous results. Thus it is **highly desirable to have ample statistical power** for effects that would be of interest to others and to report a priori power at several effect sizes (not post hoc power) for tests of your main hypotheses. [...] The main points here are to (a) do what you reasonably can to **design an experiment that allows a sensitive test** and (b) explain **how the number of participants was determined"**

Power

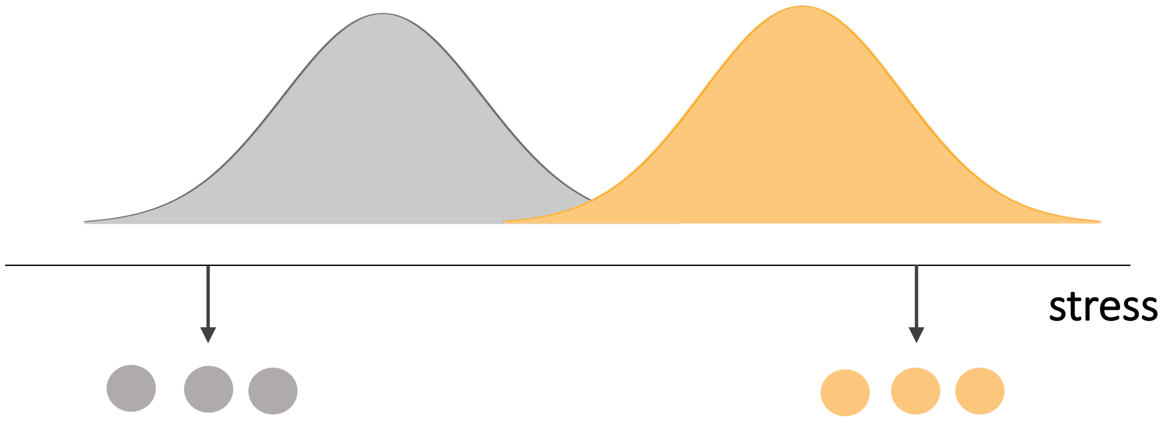


H_0: The two groups do not differ!

REJECT!

Power

Population

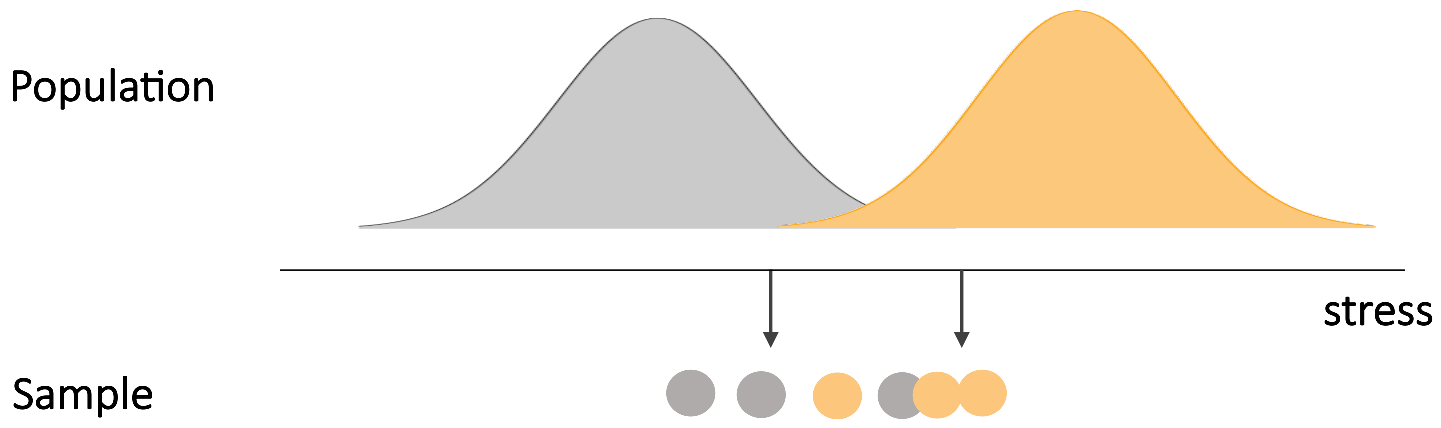


Sample

H_0 : The two groups do not differ!

REJECT!

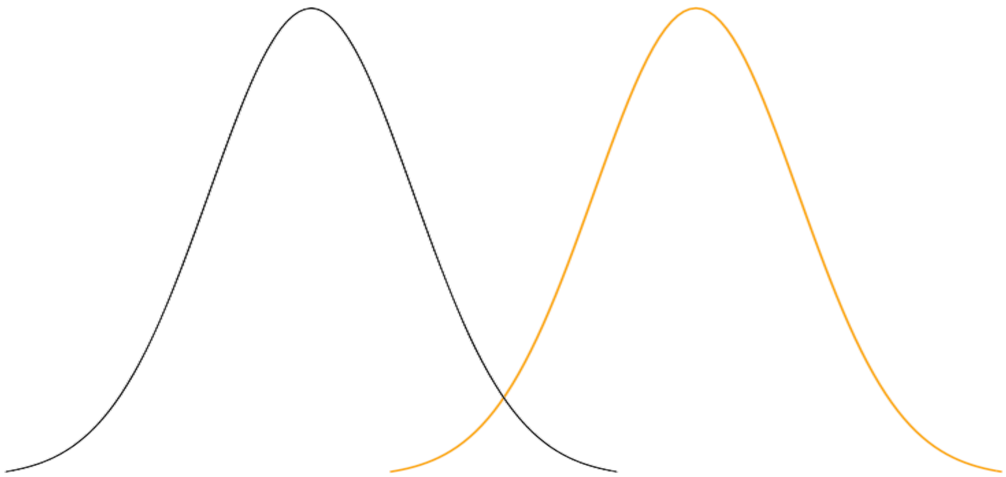
Power



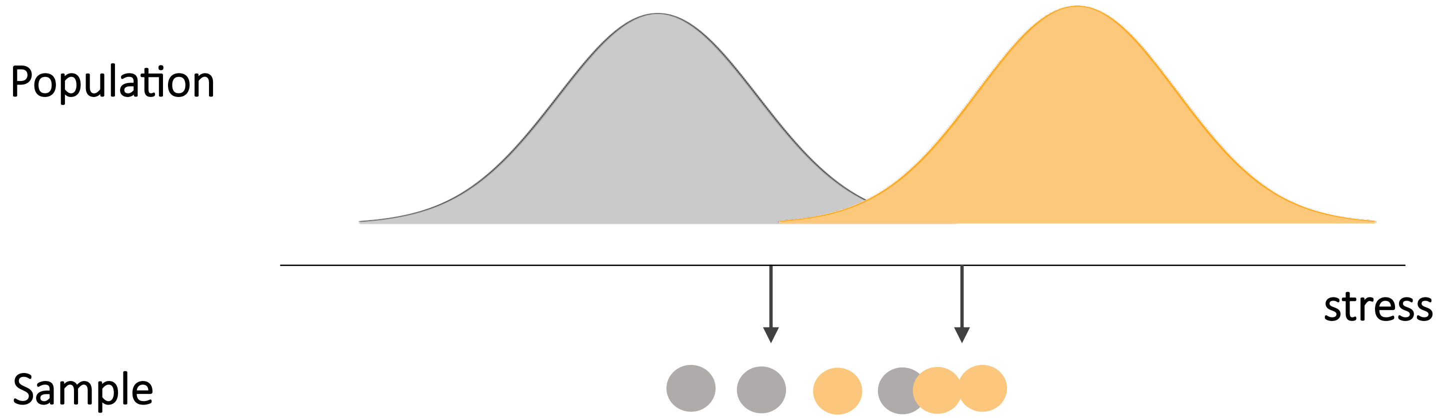
H_0 : The two groups do not differ!

NOT REJECTED

Sampling distributions



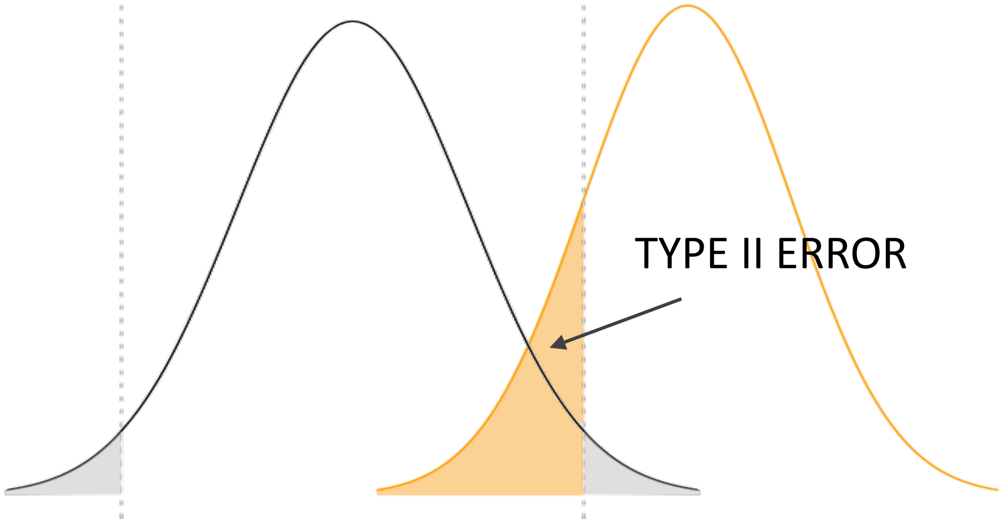
Power



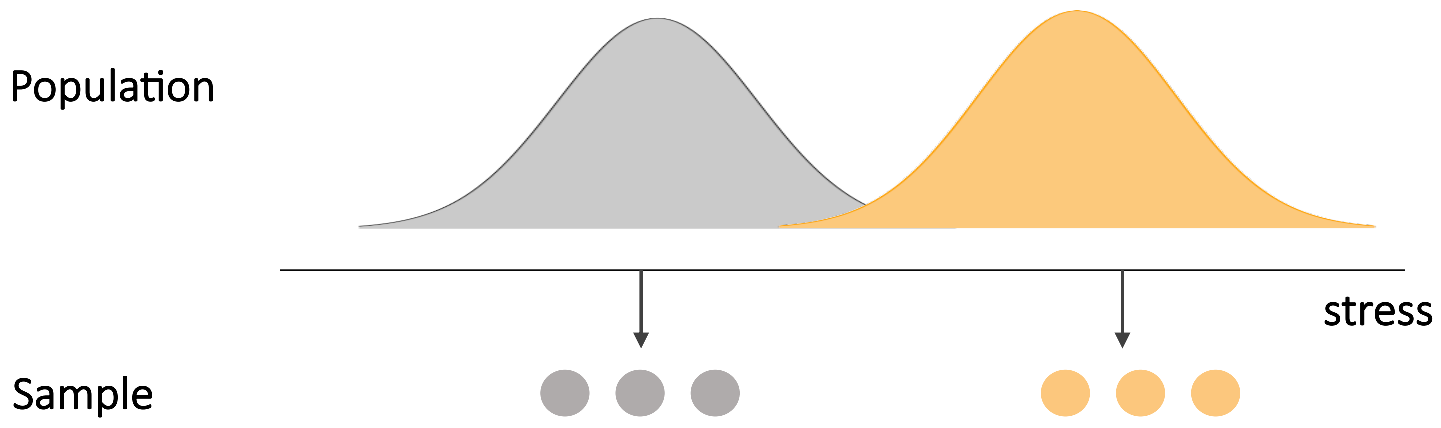
H_0 : The two groups do not differ!

NOT REJECTED

Sampling distributions



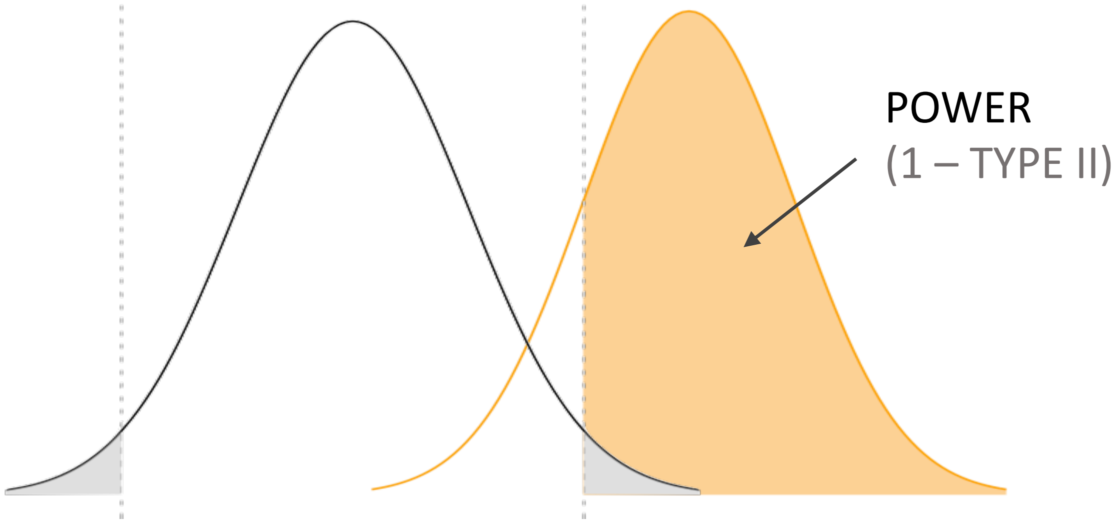
Power

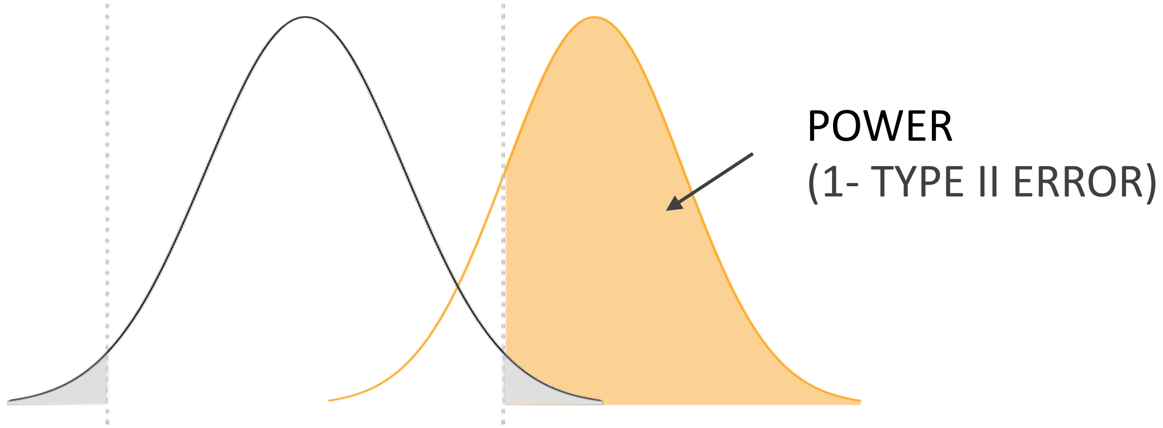


H_0 : The two groups do not differ!

REJECT!

Sampling distributions



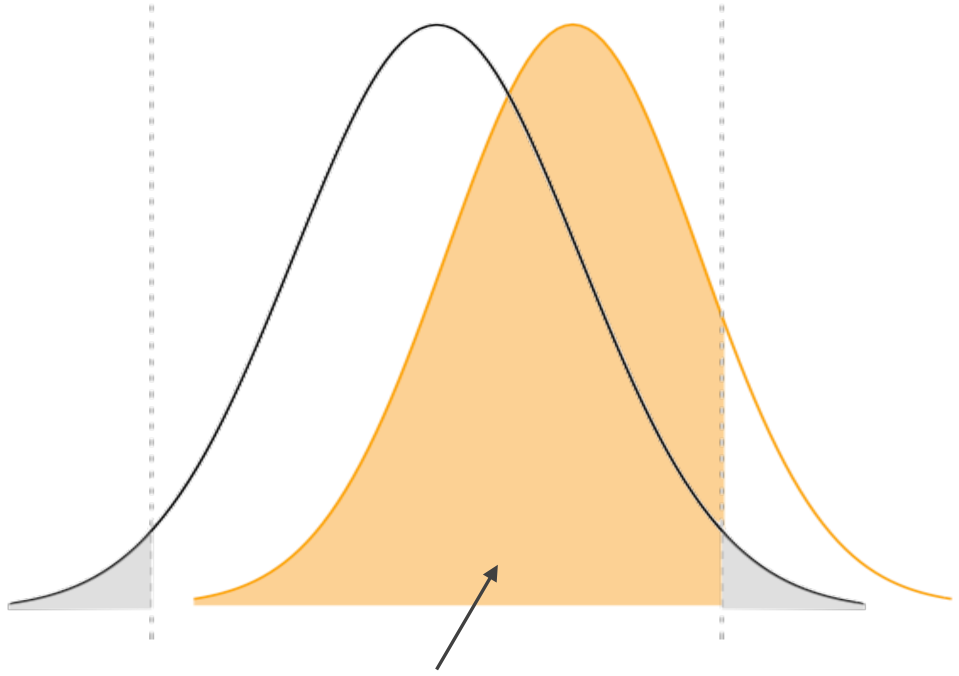


POWER = Probability of rejecting H_0 when H_0 is indeed false.

or

"If there really is an effect of a certain size – what is the probability that my study will detect it?"

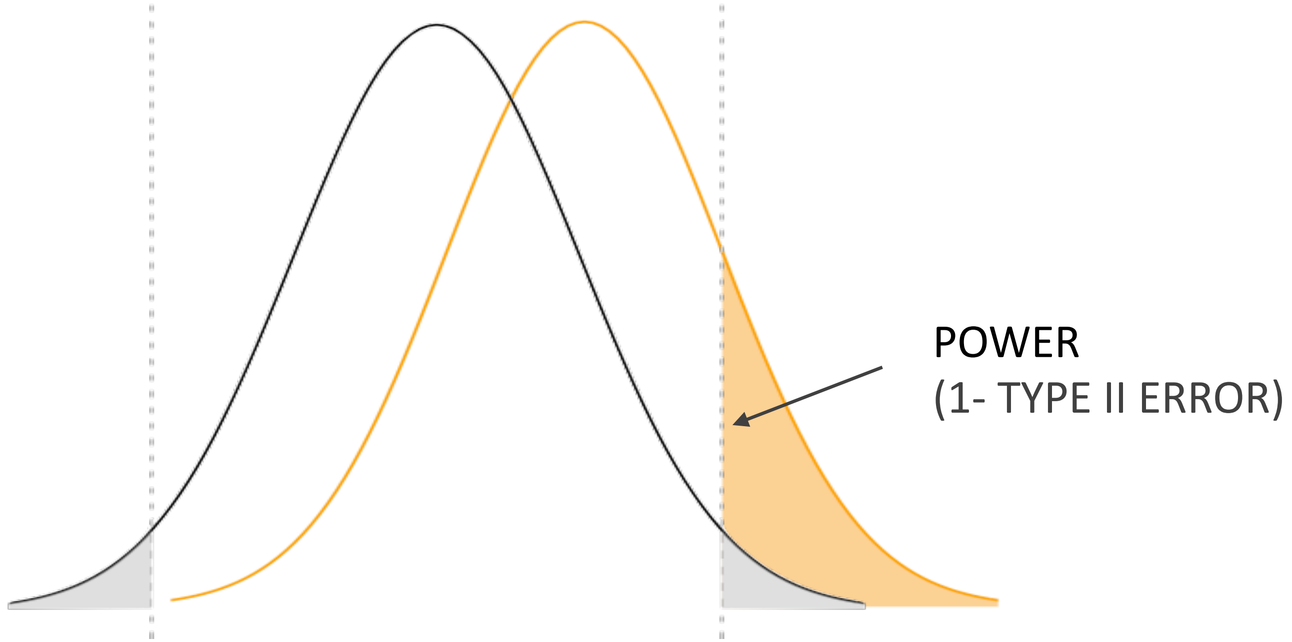
Danger of Low Power



TYPE II ERROR
(1- POWER)

WE MISS A LOT OF TRUE EFFECTS!

Danger of Low Power

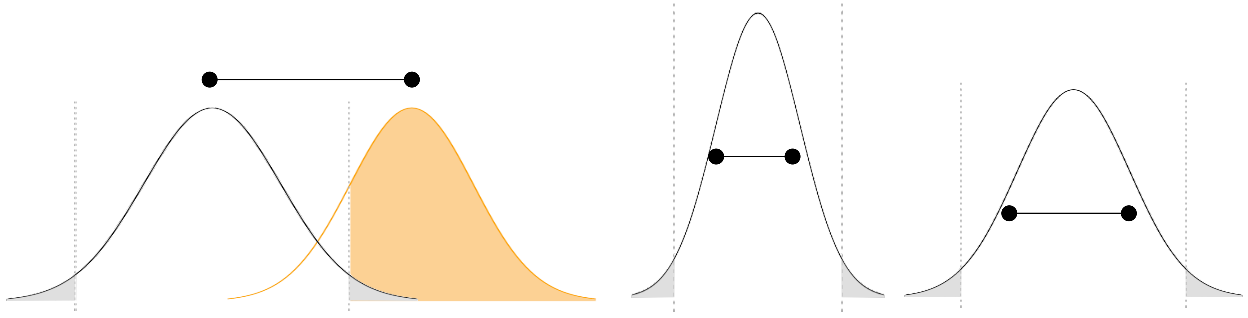


EFFECTS WE FIND ARE MUCH LARGER THAN THE TRUE EFFECTS!

→ not reliable, hard to replicate

What influences power?

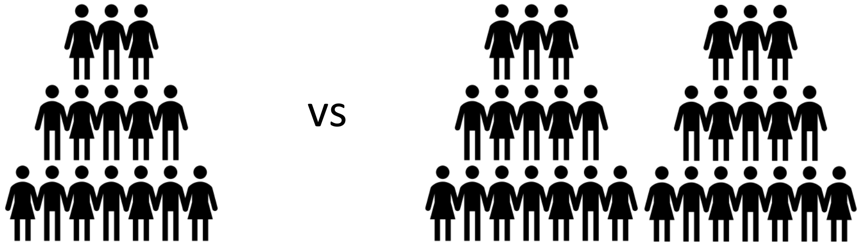
EFFECT SIZE



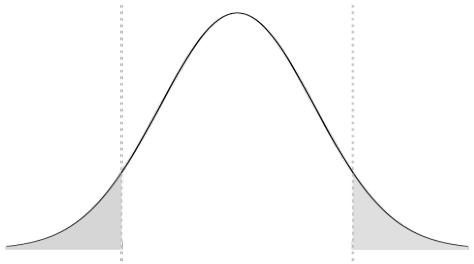
difference of means

variance

SAMPLE SIZE/ NUMBER OF OBSERVATIONS



ALPHA LEVEL



How can a power analysis help?

EFFECT SIZE + SAMPLE SIZE → POWER

EFFECT SIZE + POWER → SAMPLE SIZE

POWER ANALYSIS AS A TOOL FOR SAMPLE SIZE/DESIGN PLANNING

How can power analysis help?

POWER ANALYSIS AS A TOOL FOR SAMPLE SIZE/DESIGN **PLANNING**

Needs to be done **a-priori**

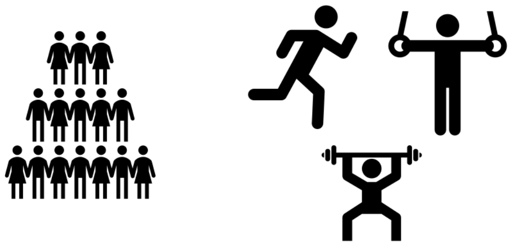
- based on the expected effect size
- Post-hoc power should not change the interpretation of *p*-values

Hoenig & Heisey, 2001

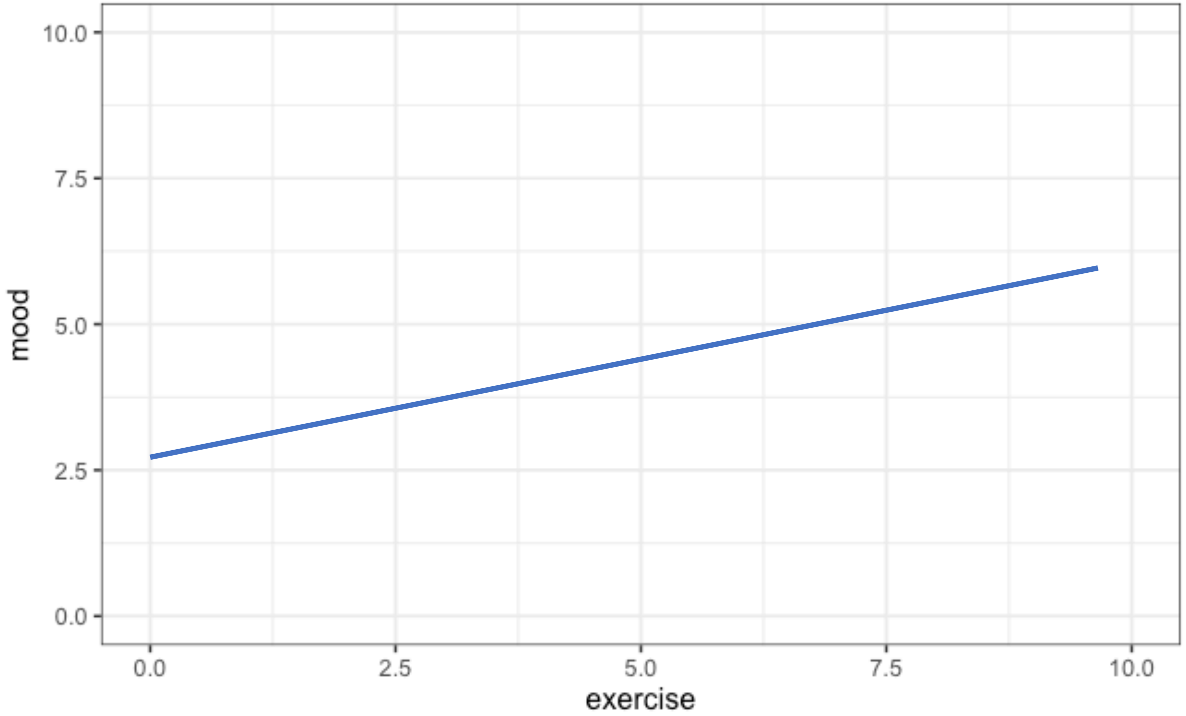
any questions so far?

What if designs/analyses get more complicated?

“How does 10 min of exercise change mood?”



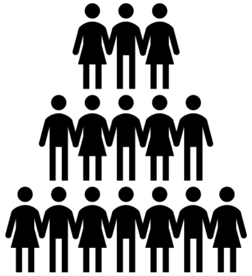
Average over subjects/exercise



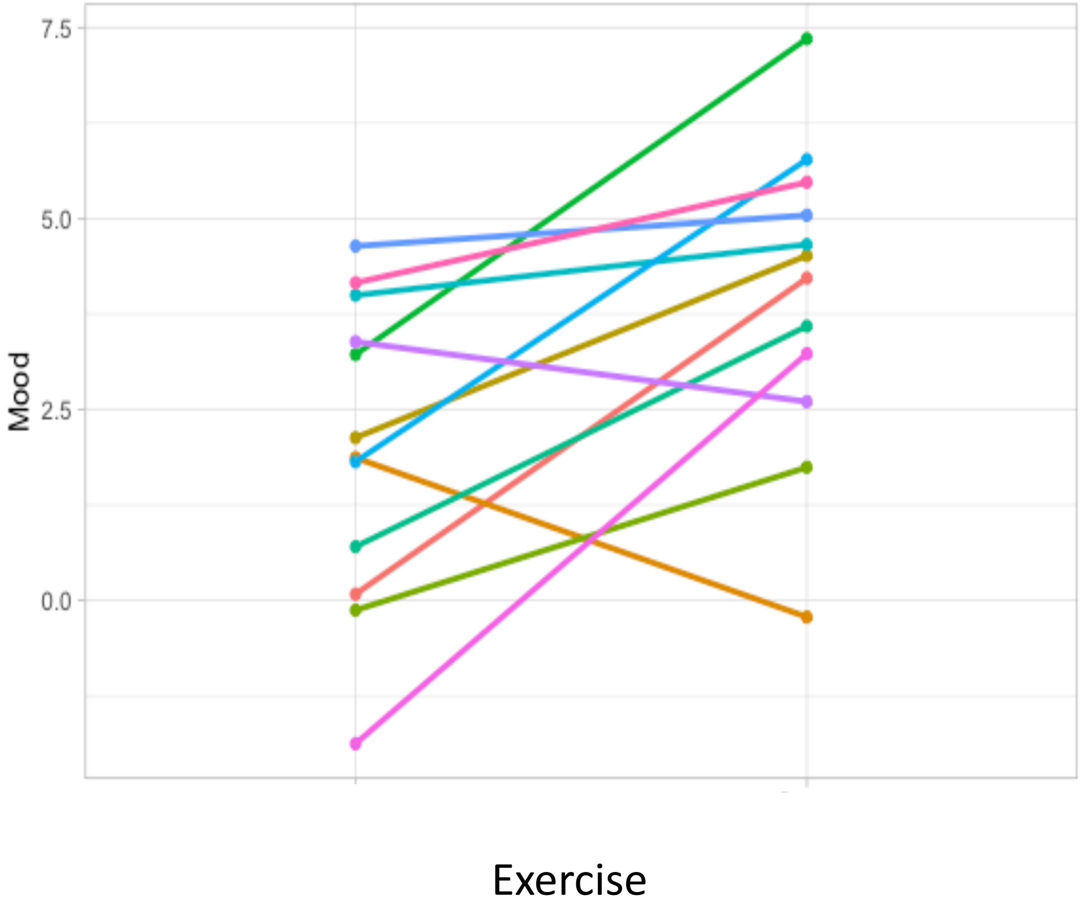
mood ~ exercise

What if designs/analyses get more complicated?

“How does 10 min of exercise change mood?”

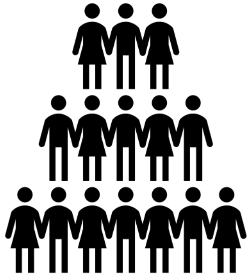


RANDOM FACTOR

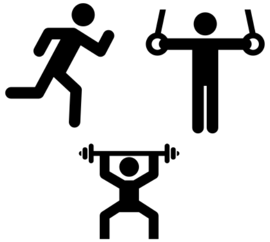


(Generalized) Linear Mixed Models

“How does 10 min of exercise change mood?”



RANDOM FACTOR



RANDOM FACTOR

```
mood ~ exercise + (1 | subject) + (1 | exerciseType)
```

(Generalized) Linear Mixed Models

```
mood ~ exercise + (1 | subject) + (1 | exerciseType)
```

- provides **flexibility** in not only modeling the means of the data, but their variance and covariance as well
- allows for estimating **between-subject and between-stimuli** variance simultanously (crossed random factors) (Baayen,Davidson & Bates, 2008)
- **advantages**: greater power in unbalanced designs, can deal with clustered data, <u>super flexible</u>!

Power and (G)LMMs

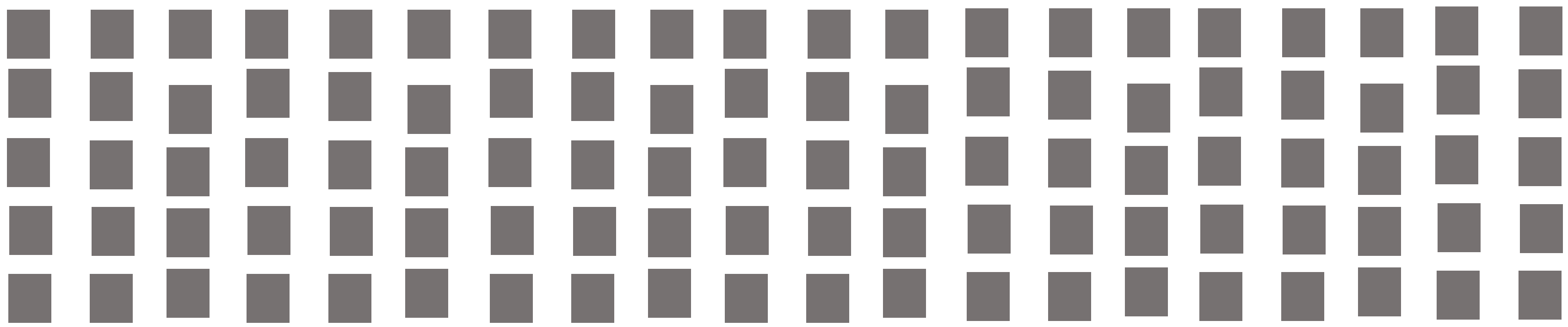
- power is influenced by variability of responses:
  - all factors that have an influence on the variability of responses need to be accounted for when estimating power
- close caption formulas for calculating power are not suitable for most (G)LMMs
- **exception:** Westfall, Kenny, and Judd (2014) for LMMs with one fixed effect (2 levels)

…we need an approach that can keep up with the flexibility of (G)LMMs!

simulation-based approach

POWER = If there really is an effect of a certain size – what is the probability that my study will detect it?

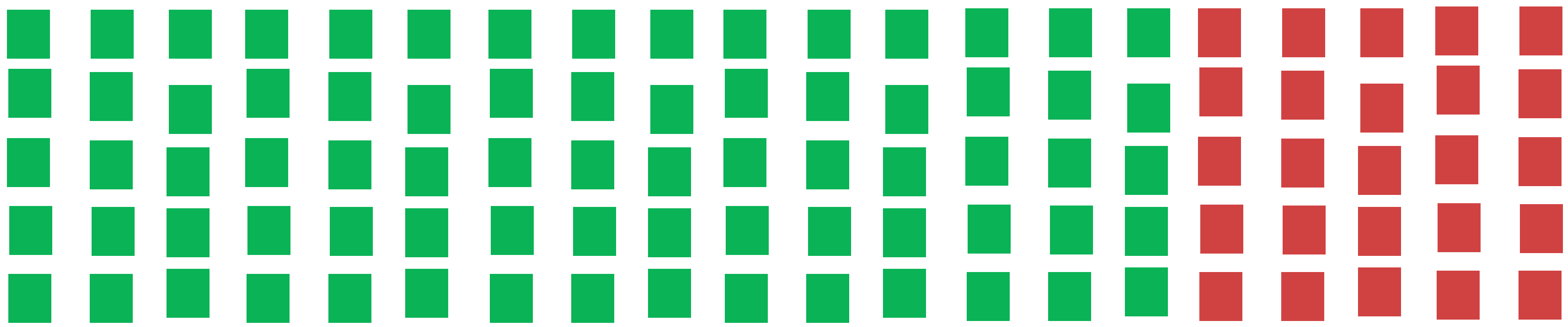
If I repeat my study 100 times – how many time will I find the effect?



simulation-based approach

POWER = If there really is an effect of a certain size –
what is the probability that my study will detect it?

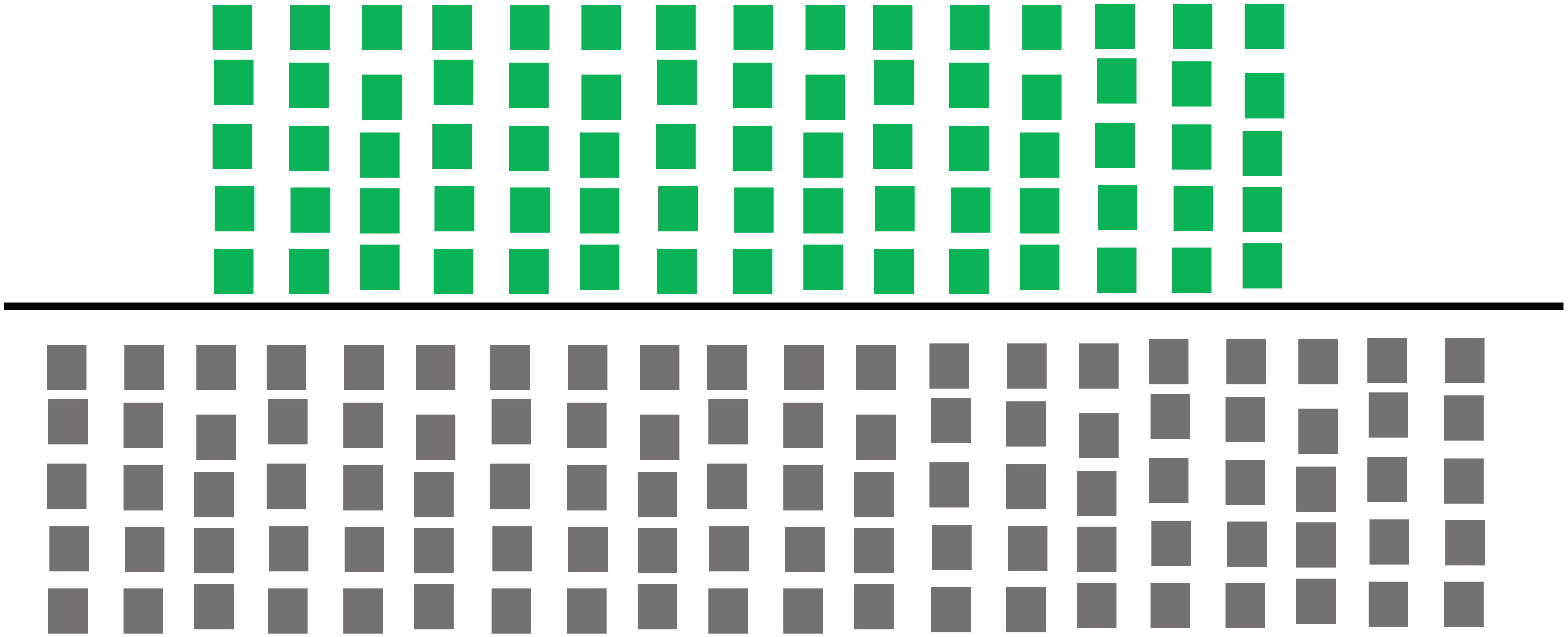
If I repeat my study 100 times – how many time will I find the effect?



simulation-based approach

If I repeat my study 100 times – how many time will I find the effect?

POWER =



simulation-based approach

If I repeat my study 100 times – how many time will I find the effect?

$$\textbf{POWER} = \frac{\text{Number of significant outcomes}}{\text{Number of all outcomes}} = 80\%$$

simulation-based approach

simulate

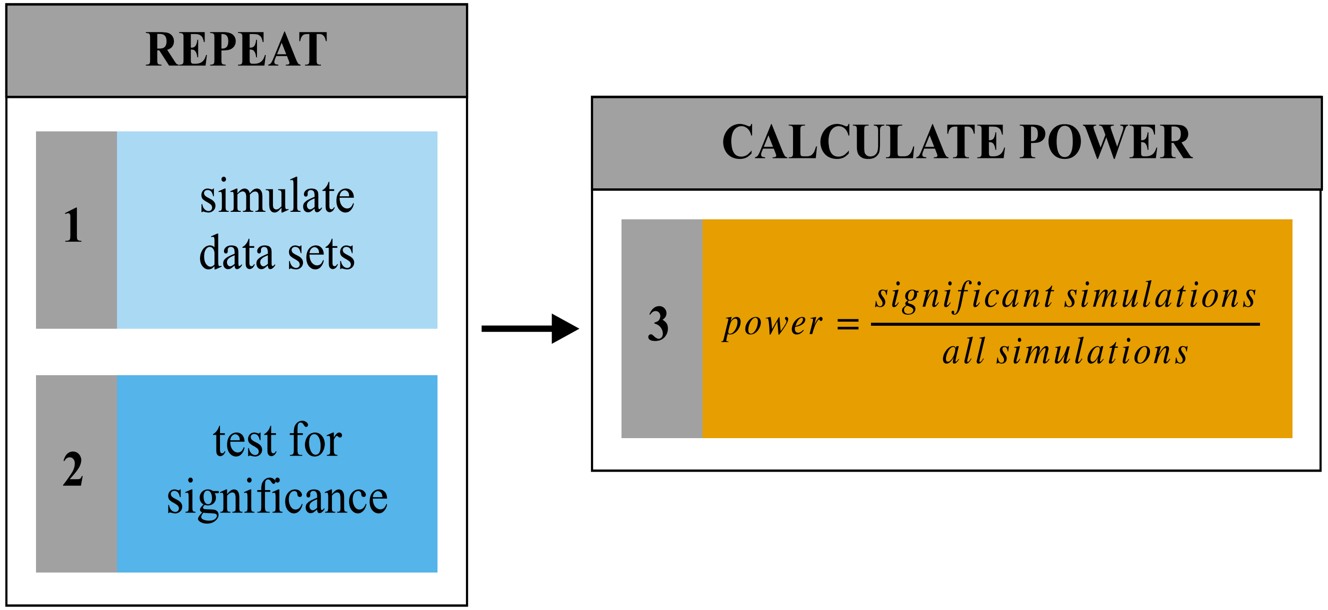
If I ~~repeat~~ my study 100 times – how many time will I find the effect?

simulations

$$\textbf{POWER} = \frac{\text{Number of significant } \cancel{\text{outcomes}}}{\text{Number of all } \cancel{\text{outcomes}}}$$

simulations

simulation-based approach



R

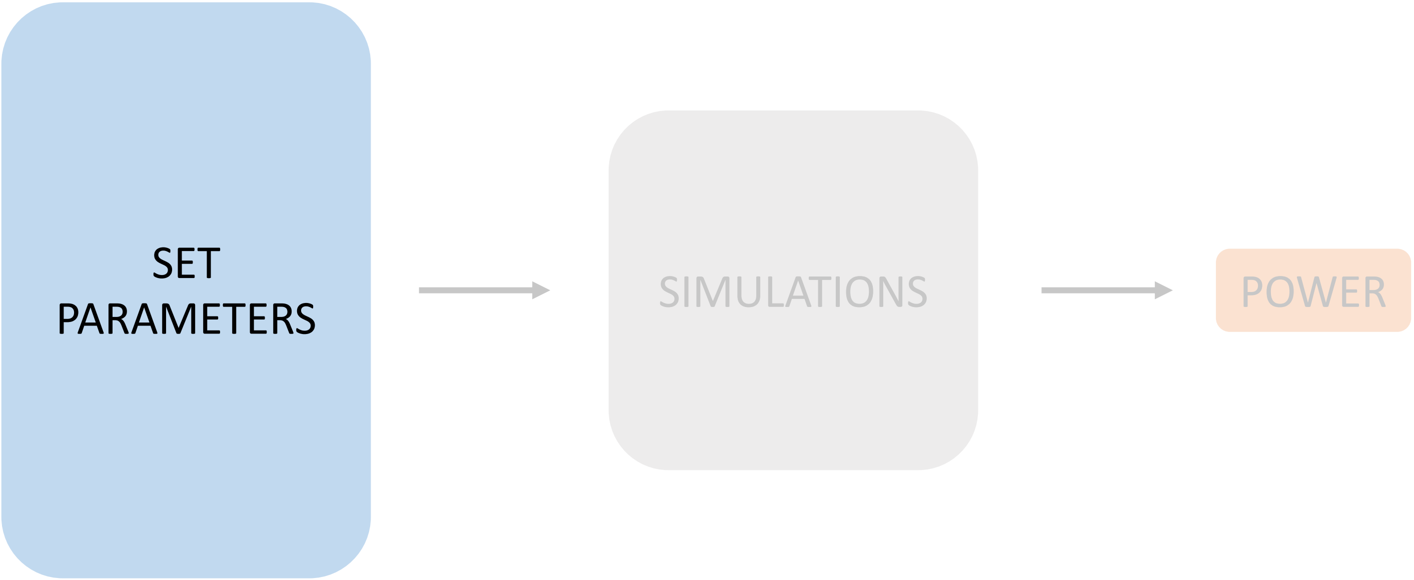
R-Time!
Introduction_simulations.R

+ short break: Let's continue at 11:00!

Challenges

SET PARAMETERS → SIMULATIONS → POWER

Challenges



1 Which parameters should we choose?

! Accuracy of the power estimate heavily depends on how accurate the simulation mimics the real data/experiment

(Thomas & Juanes, 1996)

Which parameters should we choose…?

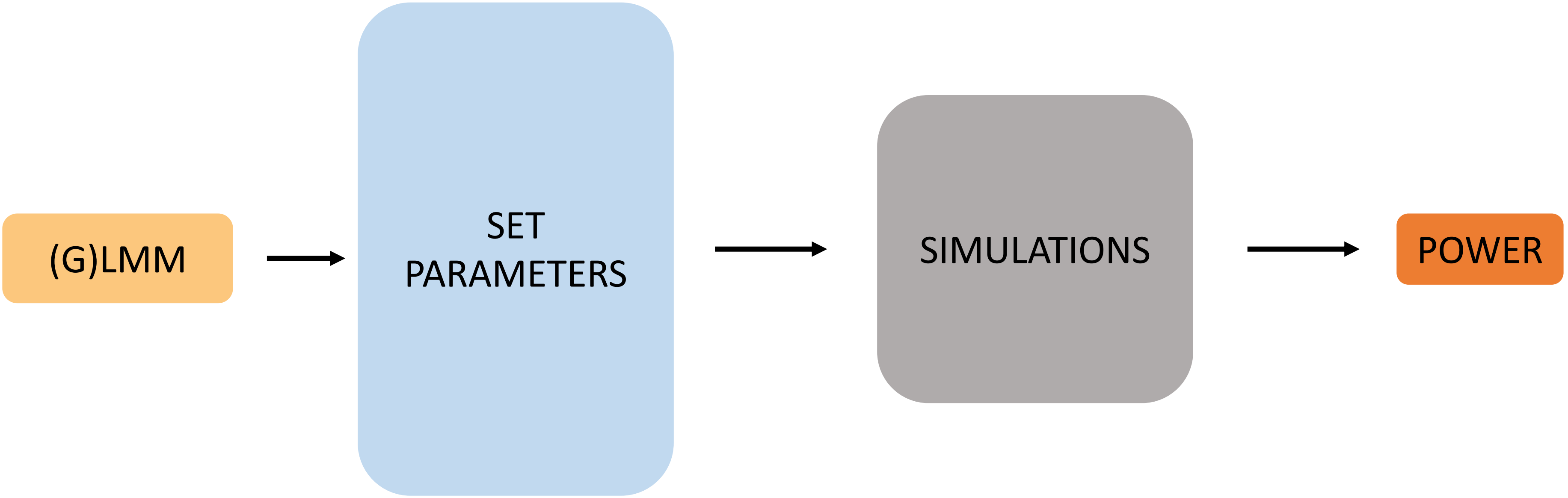
… and how can we justify them?

Difficult anyways: we will never know the true effect size/ population parameters

Even more difficult in (G)LMMs: so many more parameters to specify!

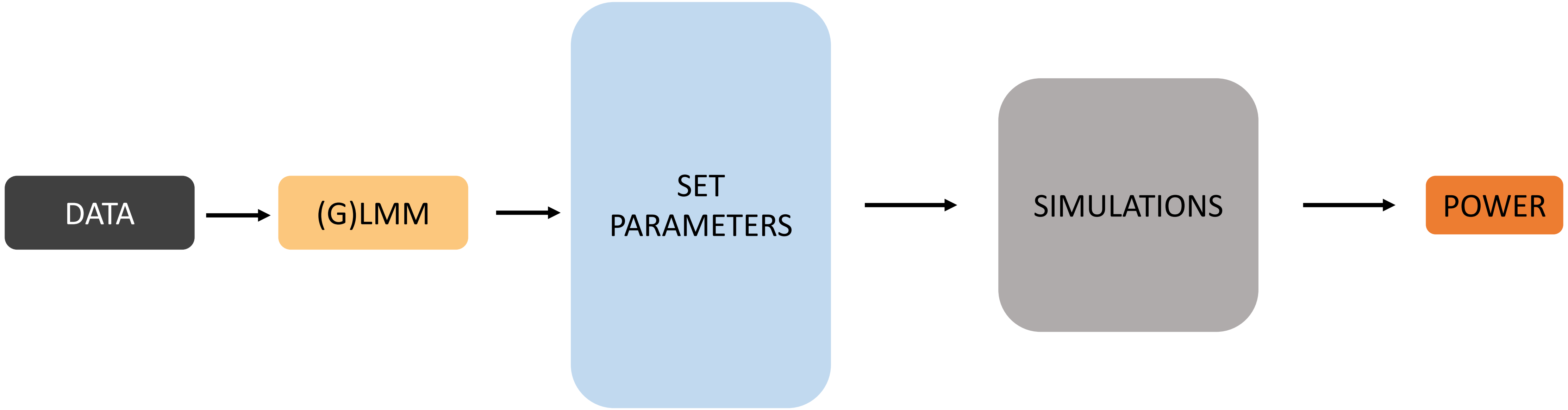
- We need a **fitted (G)LMM** to ensure that we account for parameters capturing all variation in the data

Challenges



Where from can we get a fitted model?

Challenges



Where from can we the data from?

Where from can we get the data?

prior/published studies

- **Best case scenario**: data from preceding, highly powered (reliable) studies
- Provides us with parameters for random effect structures
- **But:** We know that published effect sizes are not necessarily reliable

pilot data

- "Can't we just use pilot data?" – Yes and no
- unbiased pilot data can provide us with estimates for our random effect structure
- **But:** Small N pilot studies do not necessarily provide reliable estimates for effect sizes

simulated data

- "What if I don't have any data available?"
- Simulating data offers the greatest flexibility
- **But:** we need to be able to justify <u>a lot</u> of parameter decisions to ensure that the simulated data is accurate

How can we deal with unreliable effect sizes?

... this is where it gets very difficult!

- **Smallest Effect Size of Interest** (SESOI): “Which is the smallest effect worth discovering?”

  (Albers & Lakens, 2018)

- **But:** effect sizes in (G)LMMs are indicated through unstandardized beta coefficients
- more work is needed to establish informed decision making for SESOIs in (G)LMMs

- **Until then:** repeating power simulations over a range of plausible effect sizes/beta coefficients can give us an idea of how much power changes if effect sizes where e.g. 15% smaller

Some parameters we want to be unsure about!

... we conduct power simulations to figure out those parameters!

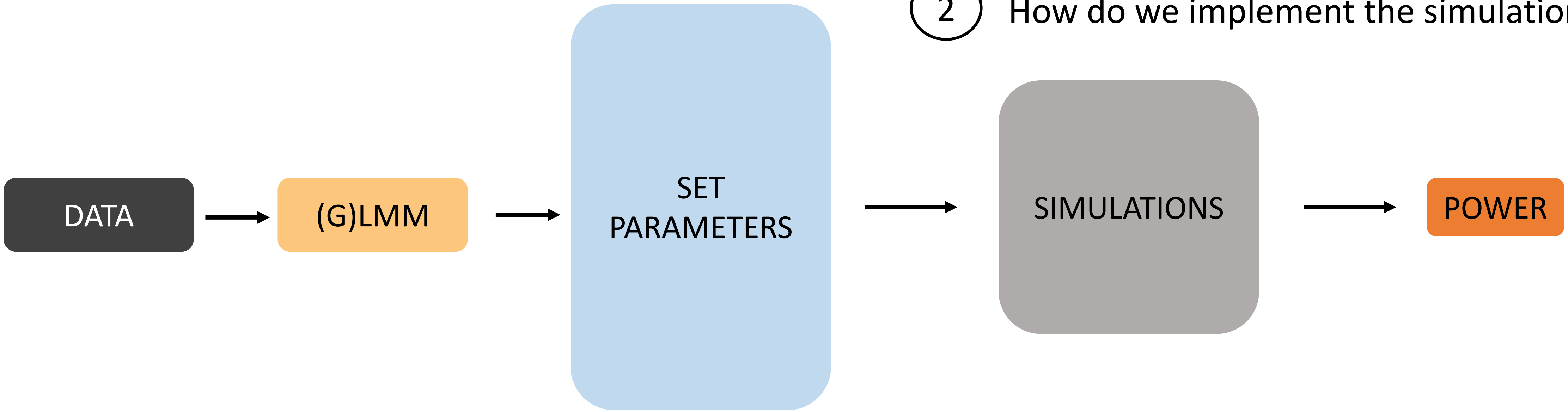
- Sample size & number of trials/stimuli

EFFECT SIZE + **POWER** → **SAMPLE SIZE**

- Simulate power over a **range of different parameters** (sample sizes/ number of stimuli) to find to make informed design choices which ensure ample power any questions so far?

Challenges

2 How do we implement the simulation in R?



Different R-packages available (**mixedpower**, SIMR, …)

R-package mixedpower

…How do we implement the simulation in R?

- **all we need**: fitted (G)LMM and corresponding data
- mixedpower uses lme4- function *simulate.merMod()*
    - if model is misspecified, simulations will not be accurate

- **What are the general conditions for my study?**
    - Are there any restrictions? (e.g. how much time/funding do I have to pay participants? Do I have a fixed stimulus list and can't increase the number of stimuli?)
    - Which power are do I want to aim for?

R-package mixedpower

How does power change if I change the levels of…

…one random factor?

mixedpower()

e.g. sample size *or* number of stimuli

…two random factor?

R2power()

e.g. sample size *and* number of stimuli

mood ~ exercise + (1 I subject) + (1 I exerciseType)

R-package mixedpower

```
mixedpower(model = moodModel, data = moodData,
           fixed_effects = c("exercise"),
           simvar = "subject", steps = c(20,30,40,50,60),
           critical_value = 2, n_sim = 1000,
           SESOI = SESOI, databased = T)
```

R-package mixedpower

Lme4 model:
mood ~ exercise + (1 I subject) + (1 I exerciseType), data = moodData

```
mixedpower(model = moodModel, data = moodData,
           fixed_effects = c("exercise"),
           simvar = "subject", steps = c(20,30,40,50,60),
           critical_value = 2, n_sim = 1000,
           SESOI = SESOI, databased = T)
```

R-package mixedpower

Lme4 model:
mood ~ exercise + (1 I subject) + (1 I exerciseType), data = moodData

```
mixedpower(model = moodModel, data = moodData,
           fixed_effects = c("exercise"),
           simvar = "subject", steps = c(20,30,40,50,60),
           critical_value = 2, n_sim = 1000,
           SESOI = SESOI, databased = T)
```

! Rule-of-Thumb: Transform all variables **before** fitting the model, not in the model formula

R-package mixedpower

Lme4 model:
mood ~ exercise + (1 I subject) + (1 I exerciseType), data = moodData

```
mixedpower(model = moodModel, data = moodData,
           fixed_effects = c("exercise"),
           simvar = "subject", steps = c(20,30,40,50,60),
           critical_value = 2, n_sim = 1000,
           SESOI = SESOI, databased = T)
```

R-package mixedpower

Lme4 model:
mood ~ exercise + (1 I subject) + (1 I exerciseType), data = moodData

```
mixedpower(model = moodModel, data = moodData,
           fixed_effects = c("exercise"),
           simvar = "subject", steps = c(20,30,40,50,60),
           critical_value = 2, n_sim = 1000,
           SESOI = SESOI, databased = T)
```

Which random factor to I want to vary?
Which levels of this random factordo I want to simulate power for?

R-package mixedpower

Lme4 model:
mood ~ exercise + (1 I subject) + (1 I exerciseType), data = moodData

```
mixedpower(model = moodModel, data = moodData,
           fixed_effects = c("exercise"),
           simvar = "subject", steps = c(20,30,40,50,60),
           critical_value = 2, n_sim = 1000,
           SESOI = SESOI, databased = T)
```

! lme4 does not provide p-values, critical value is used to establish significant

- z –value (GLMMs), t-value (LMMs) of 2 ~ alpha = 0.05

(Baayen et al., 2008)

R-package mixedpower

Lme4 model:
mood ~ exercise + (1 I subject) + (1 I exerciseType), data = moodData

```
mixedpower(model = moodModel, data = moodData,
           fixed_effects = c("exercise"),
           simvar = "subject", steps = c(20,30,40,50,60),
           critical_value = 2, n_sim = 1000,
           SESOI = SESOI, databased = T)
```

! **Good practice:** Include at least 1000 single simulations

➢ The more single runs we include, the more stable our simulation estimate becomes!

R-package mixedpower

Lme4 model:
mood ~ exercise + (1 I subject) + (1 I exerciseType), data = moodData

```
mixedpower(model = moodModel, data = moodData,
           fixed_effects = c("exercise"),
           simvar = "subject", steps = c(20,30,40,50,60),
           critical_value = 2, n_sim = 1000,
           SESOI = SESOI, databased = T)
```

allows for modifying effect sizes (beta coefficients) of our simulation

Do we want to include a simulation based on the “original” beta coefficients?

R-package mixedpower

Lme4 model:
mood ~ exercise + (1 I subject) + (1 I exerciseType) data = moodData

```
R2power(model = moodModel, data = moodData,
          fixed_effects = c("exercise"),
          simvar = "subject", steps = c(20,30,40,50,60),
          R2var = "exerciseType", R2level = 10,
          critical_value = 2, n_sim = 1000,
          SESOI = SESOI, databased = T)
```

Interpreting simulation results

Question: “Which sample size do we need for our study?”

	20	30	40	50	mode
exercise	0.42	0.54	0.73	0.89	databased
fitnessLevel	0.51	0.65	0.82	0.94	databased
exercise	0.27	0.41	0.65	0.79	SESOI
fitnessLevel	0.41	0.56	0.75	0.82	SESOI

Lme4 model:
mood ~ exercise + **fitnessLevel** + (1 I subject) + (1 I exerciseType), data = moodData

Interpreting simulation results

Question: “Which sample size do we need for our study?”

	20	**30**	**40**	**50**	**mode**
exercise	0.42	0.54	0.73	0.89	databased
fitnessLevel	0.51	0.65	0.82	0.94	databased
exercise	0.27	0.41	0.65	0.79	SESOI
fitnessLevel	0.41	0.56	0.75	0.82	SESOI

fixed effects

Lme4 model:
mood ~ exercise + **fitnessLevel** + (1 I subject) + (1 I exerciseType), data = moodData

Interpreting simulation results

Question: “Which sample size do we need for our study?”

steps of *simvar* → sample sizes

	20	**30**	**40**	**50**	**mode**
exercise	0.42	0.54	0.73	0.89	databased
fitnessLevel	0.51	0.65	0.82	0.94	databased
exercise	0.27	0.41	0.65	0.79	SESOI
fitnessLevel	0.41	0.56	0.75	0.82	SESOI

Lme4 model:
mood ~ exercise + **fitnessLevel** + (1 I subject) + (1 I exerciseType), data = moodData

Interpreting simulation results

Question: "Which sample size do we need for our study?"

Which effect size is power based on?

	20	30	40	50	mode
exercise	0.42	0.54	0.73	0.89	databased
fitnessLevel	0.51	0.65	0.82	0.94	databased
exercise	0.27	0.41	0.65	0.79	SESOI
fitnessLevel	0.41	0.56	0.75	0.82	SESOI

Lme4 model:
mood ~ exercise + **fitnessLevel** + (1 I subject) + (1 I exerciseType), data = moodData

Interpreting simulation results

Question: “Which sample size do we need for our study?”

	20	30	40	50	mode
exercise	0.42	0.54	0.73	0.89	databased
fitnessLevel	0.51	0.65	0.82	0.94	databased
exercise	0.27	0.41	0.65	0.79	SESOI
fitnessLevel	0.41	0.56	0.75	0.82	SESOI

- Decision should be based on **fixed effect with lowest power** to ensure ample power for the whole design

R-package mixedpower

R-Time! ~ 20 min

In case you want to work with your own data:

- Introduction_mixedpower.R

Otherwise:

- Introduction_mixedpower_example.R

Reporting simulation results

"Sample size was planned in accordance with **effect sizes and samples from Josephs et al. (2016).** Simulation-based estimation of sample size was conducted by running **1000** linear mixed-models on simulated **data of that study** using the mixedpower library (Kumle, Võ, & Draschkow, 2018). Twenty-one participants yielded **> 90% power** for the detection of the **search vs. memorization comparison** using the reported non- standardized effect sizes of Joseph et al.'s Experiment 4 (**condition β = 0.25, gaze duration β = 0.76**) [...]"

- **Good practice:** Make transparent which parameters you used in your simulation and where those parameters are coming from!

Helbing, Draschkow & Võ (2020)

Resources

Workshop is based on this tutorial paper:

Kumle, Vo & Draschkow (in press). Estimating power in (generalized) linear mixed models: an open introduction and tutorial in R.

- accompanying **R-Notebooks**: https://lkumle.github.io/power_notebooks/
- mixedpower **R-package**: https://github.com/DejanDraschkow/mixedpower

Further Resources:

Brysbaert, M., & Stevens, M. (2018). Power Analysis and Effect Size in Mixed Effects Models: A Tutorial. *Journal of Cognition*, *1*(1), 1–20. https://doi.org/10.5334/joc.10

Green, P., & MacLeod, C. J. (2016). SIMR: An R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution*, *7*(4), 493–498. https://doi.org/10.1111/2041-210X.12504

Westfall, J., Kenny, D. A., & Judd, C. M. (2014). Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology.* https://doi.org/10.1037/xge0000014

And **THANK YOU** to everyone involved in the work featured in this workshop!

Kumle, L., Võ , M. L., & Draschkow, D. (in press). Estimating power in (generalized) linear mixed models: an open introduction and tutorial in R. *Behavior Research Methods,* https://doi.org/10.31234/osf.io/vxfbh

Melissa Võ

Dejan Draschkow

Scene Grammar Lab

And **THANK YOU** for listening!

Any Questions?

References

Albers, C., & Lakens, D. (2018). When power analyses based on pilot data are biased: Inaccurate effect size estimators and follow-up bias. *Journal of Experimental Social Psychology*. https://doi.org/10.1016/j.jesp.2017.09.004

Baayen, Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Bates, Mächler, M., Bolker, B., & Walker, S. (2014). Fitting Linear Mixed-Effects Models using lme4. *Journal of Staistical Software*, *67*(1). https://doi.org/10.18637/jss.v067.i01

Helbing, J., Draschkow, D., & Võ, M. L. H. (2020). Search superiority: Goal-directed attentional allocation creates more reliable incidental identity and location memory than explicit encoding in naturalistic virtual environments. *Cognition*, *196*, 104147.

Hoenig, J. M., & Heisey, D. M. (2001). The abuse of power: The pervasive fallacy of power calculations for data analysis. *American Statistician*, *55*(1), 19–24. https://doi.org/10.1198/000313001300339897

Kumle, L., Võ, M. L.-H., & Draschkow, D. (2018). Mixedpower: a library for estimating simulation-based power for mixed models in R. https://doi.org/https://doi.org/10.5281/ zenodo.1341047

Thomas, L., & Juanes, F. (1996). The importance of statistical power analysis: An example from Animal Behaviour. *Animal Behaviour*, *52*(4), 856–859. https://doi.org/10.1006/anbe.1996.0232