# Workshop power analysis / sample size estimation

## Duration: 3-4 h

## White board / group discussion / case studies

## Set up:

### Part A (1 h): group discussion: broading horizons

### Part B (0.5-1 h): case study together

### Part C (0.5-1 h): case study yourself

### Part D: remaining questions

## Feel free to ask questions during the presentation

# General setting:

## Based on randomized controlled trial

## Conditions: intervention / control

## But concepts and most of methods can be applied to regression model (coefficients) on observational data as well

## unless otherwise: 80% power, significance level alpha=0.05 two-sided test

# Part A: power and sample size broadening horizons

# What is power ?

## Originally from hypothesis testing

### Type 1 error, type 2 error (=1-power)

## Example: sample size z-test (approximation to t-test)

# Power > success criterion

## Generally based on obtaining a “success” (Yes/No) of study

## Original: success criterion= stat. sig. for a test

### Equivalently: position of the confidence interval compared to 0 (additive effect) or 1 (relative effect)

# Power > more \*general\* success criteria

## Based on combination of two or more endpoints (confidence intervals)

### Superiority for performance OR satisfaction

### Superiority for performance AND satisfaction

### Superiority for performance AND non-inferiority for satisfaction

### Superiority for one of performance and satisfaction and at least non-inferiority in the other (combination of the previous and vice versa)

## References: Donkers et al. (2017), Borm et al. (2007)

# Power > more \*general\* success criteria

## Based on combination of analysis results: e.g. success is:

### 1) Effect directly after program is stat.sign. AND 2) at least half of the effect is maintained numerically 6 month later

### Note that 1) is a condition on the confidence interval and 2) a condition on a point estimate.

## Based on Bayesian analysis: e.g. success is:

### Posterior probability that [the effect is larger than 0.2ES] is larger than 80%

### Bayesian 95%-credibility interval is above 0.24 (for example)

# What is the use of power?

## Funding requirement

## Sample size planning

# Are there other ways to plan sample size?

### Feasibility / convenience sample

### so large so as to be representative

#### unbiased but may be not precise

### Explorative study (if not invasive often 10-20 subjects approvable by Ethical Committee)

### Precision !

# Precision calculation (alternative for power)

## If feasibile sample sizes do not allow sufficient power

## Precision calculation: width of confidence interval

## Argumentation that this precision is meaningful

### For interpreting the effect (“if effect is such and such large, then with this precision, we can …”)

### For basing new studies on

## Advantage:

### transparent,

### no need to reverse engineer power (based on the sample size calculate, the effect for which one as sufficient power)

# Which factors influence power?

## There is not “the” power of a study

## Depends, apart from sample size, on scenario:

### Size of effect(s)

### Variance in the data

### Study design

### …

## Which scenarios to cover….?

# What is post-hoc power?

## post-hoc = after seeing the data

### point estimate and confidence interval (can be) known!

## There is not “the” post-hoc power calculation, has many appearances, amongst others:

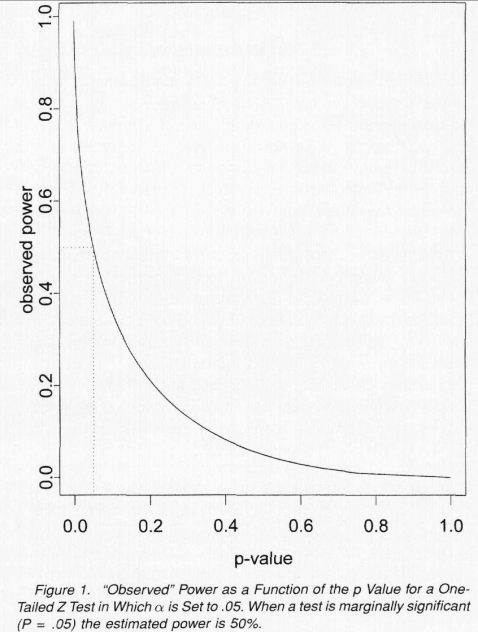
### What is the power for the observed effect?

### What is the detectable difference / is the effect ‘near-null’?

### What is the power for a given effect (not the observed or the a priori assumed effect)

### Other …(see Hoenig & Heissy, 2001)

# Power for observed effect is directly related to p-value



# Why do some people want to know post-hoc power?

## ….. often in a failed study!

## Argue that a relevant effect existed but this study was stat.sign. due to low power

## Argue that the effect is actually zero (absent)

## Confidence Intervals better: upper/lower limit show:

### Can the real effect be a relevant effect?

### Is the real effect close to nil?

## Confidence intervals better for planning a new study

# Power for Replicate studies

## Replicate study: exactly the same design as original study

### apart from sample size

## P-value of the original study determines power of replicate study

## P=0.05: power of replicate study of same size is 50%

### Sample size 2x as large for 80% power

## P=0.005: power of replicate study of same size is 80%

# Reason: observed p-value vs observed power relation

## Borm et al 2010

# Does 1-sided instead of 2-sided testing make sense to improve power?

## Sometimes one-sided testing is proposed to improve power

## One-sided (e.g.): H0: delta ≤ 0 vs H1: delta > 0

## Two-sided (e.g): H0: delta = 0 vs H1: delta ≠ 0

## Often: power gain is small

# Example one-sided testing:

## We test experimental vs control with H0: delta ≤ 0 vs H1: delta > 0

### Thus superiority testing only (“is experimental better than control”)

## Sample estimate of delta=-5 with standard error of 2.

## Then a two-sided 95%-CI is approximately -5 ± 1.96x2= ( -8.92 , -1.08 ), so could have concluded that the experimental is worse than control.

## However, we did one-sided testing and the corresponding confidence is approximately -5 – 1.64x2 =(-8.28 , ∞), so basically we have a failed experiment

### Cannot show experimental is better than control nor the opposite.

### Note: 1.64 is the percentile for a one-sided 95%-confidence interval

## Key: only can draw a conclusion when the null hypothesis is rejected

## Probably: one-sided testing often has more disadvantages than advantages

# How to choose the size of effect?

## Expected effect (replicate something you have seen or anticipate)

## Relevant effect (for a purpose):

### often smaller sample size needed

### but: would you really \*not\* be interested in a smaller effect?

## Measurement scale?

### Standardized, e.g. Cohen’s d or …

#### Distribution based measure

### original scale

#### may lend itself better for assessing relevance

# Size of effect

## Expected effect (replicate something you have seen or anticipate)

## Relevant effect (for a purpose):

### often smaller sample size needed

### but: would you really \*not\* be interested in a smaller effect?

## Choose depending on

### previous studies,

### whether expected or relevant effect is aimed

### biomedical often more original scale for interpretation, social sciences often more standardized measures

# How to determine relevance of effect?

## effect is on population level

### e.g. for a mean: some may improve more than the mean and others less

### therefore often responder analyses

## relevant for who?

### Payers of the intervention, patients/subjects, care givers, …

## relevant effect:

### effect relevant on individual level, then often also on population level

### but: also small effect on population level may be relevant in big population

### 0.1% mortality reduction from 5% to 4.9% is relevant in 100.000 patients: 100 patients

### Minimally important difference (MID, MCID)

# Size of effect: minimally important difference

## anchor-based

### comparison with an external indicator (e.g. subject’s assessment or professional’s assessment) of change

### construction provides relevance/importance

## statistical distribution-based

### 0.5 x SD

### measurement error of in one measurement of the outcome

### Not clear why these should capture ‘relevance’

## Some anecdotal findings that statistical distribution methods are similar to anchor-based

## See Revicki et al. 2008

# Ways to improve power by design (measurements within subjects)

## more subjects

## more measurements (but power may plateau)

## more ‘informative’ comparisons (when intervention/control)

### cross-over design (within-subject comparisons) more power than

### one-sided cross-over designs (like stepped-wedge) more power than

### parallel group designs

# Ways to improve power by testing (when multiple outcomes)

## Multiple outcomes:

### different endpoints: performance and satisfaction

### different time points: effect directly after program and one year later

### different interventions : individual therapy and group therapy

### …

## Powerful testing procedures

# Powerful testing procedures

## Three principles to test multiple outcomes:

### 1 Bonferroni splitting

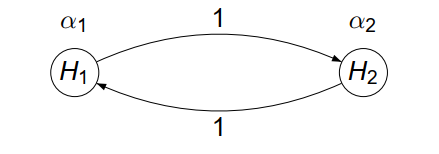
### 2 Fixed sequence testing

### 3 Closed testing

## Combinations give powerful testing procedures

# Powerful testing procedures: examples

## Weighted Bonferroni-Holm procedure



## Many more examples and introduction in the introduction by Bretz et al. 2010

# What about non-continuous/normal outcomes?

## In multilevel / repeated measurement models:

## consider averaging to higher level units to obtain more normally distributed outcomes

## Example:

### taking the average of a binary outcome in clusters gives a percentage.

### Unless close to 0 or to 100% percentages have reasonably constant variance

# Part B: casus

# Setting

## Randomization to two groups (equal size)

## Measurements

### at randomization

### one after end intervention (fu1)

### one long-term measurement (fu2)

## Continuous outcome with SD: σ

## Intervention with size of effect: δ

## How many subjects per group?

# KISS (keep it stupid simple)

## T-test on one followup measurement

### per group

## If ensuing sample size is feasible, then you are done:

### using more measurements will give more power

# KISS-2

## Assume the effect is constant over the follow-up measurements (or interest is in the average effect)

### more or less an area under the curve

## Take the average of the two followup measurements

## Then SD is reduced (so power increased):

## Apply t-test again

### per group

## Note: if *r* is close to 1 (i.e. fu1 and fu2 are very correlated), then you do not gain compared to t-test on one followup measurement

# KISS-3 (keep it stupid simple)

## Analyze one follow-up measurement and correct for baseline

## ANCOVA analysis

### per group, where *r* is correlation between baseline and followup measurement

# But what about using all three measurements?

## For “full” repeated measurements designs with equal correlation *r* between all pairs of measurements and constant effect:

## Describe the design as a matrix

### rows determine groups/arms in the trial,

### columns determine measurements

### 0 for control, 1 for intervention

## Then apply a calculation as in Teerenstra et al. to get

### per group, where *r* is correlation between baseline and followup measurement

# Details (full repeated, equal correlation, constant effect)

## Make the matrix describing the design

## Calculate

## where

### *I* the total number of subjects, *T* the total number of measurements,

### *f=S I – C, g= S2 – I R + f T*,

### *S* is the sum of the 1s in the matrix above, *C* the sum of the squared colum sums, *R* the sum of the squared row sums

# But what about using all three measurements and possibly non-equal correlation, and possibly different effects?

## In principle, this can be calculated for general correlations between and variances of the measurements, but tedious

## Alternatives

### Changing to another effect measure, e.g. difference in average slope

### Simulations

# Difference in average slopes

## Power analysis for multilevel trials (Moerbeek & Teerenstra) equation (9.11): formula but also with orthogonal contrast

## See power program SPA-ML

# Simulation

## Needed:

## 1 data generating model

## 2 analysis method

## 

* 1 uur casus
  + 3 repeated measurement
    - To model: time trend/course, correlation, intervention effect
    - What is effect measure?
      * Summary measure (single time, AUC, slope,
    - Where is effect meaningful (or largest)?
    - Effective sample / cluster size
    - Expected control time trend linear or quadratic (given three points)
    - Randomization assumption:
    - Correlation structure
      * Simplified Repeated measurement (article with Hooper)
      * Autocorrelation vs compound symmetry: somewhere in the middle
      * Damped exponential decay?
    - KISS (Keep It Simple)
    - Formulas:
      * Ready sample size formula Book for longitudinal measurements (see Mirjam)
    - Simulation program
      * Only at one regression model coefficient (e.g. time slope or single time point\*interaction)?
* half uur zelf (2x?) op casus Peter
* Also important
  + type I error
* formula voor repeated measurements:
  + equal correlation
  + AR(1)
  + ertussenin

Power simulation in R

* General idea for R: <https://htmlpreview.github.io/?https://github.com/jwb133/SimStudiesR/blob/master/SimulationStudiesinR.html>

<https://rpsychologist.com/introducing-powerlmm>

<https://raw.githubusercontent.com/OscarOlvera/R-code-for-publications/master/heteroskedasticity>

of tot nu toe de beste

linear mixed models:

<https://stats.idre.ucla.edu/other/mult-pkg/introduction-to-linear-mixed-models/>