Alexander Strobel
Faculty of Psychology

Workshop Open Science Practices

Open Science and Preregistration

Power Analyses

Introduction
Key terms

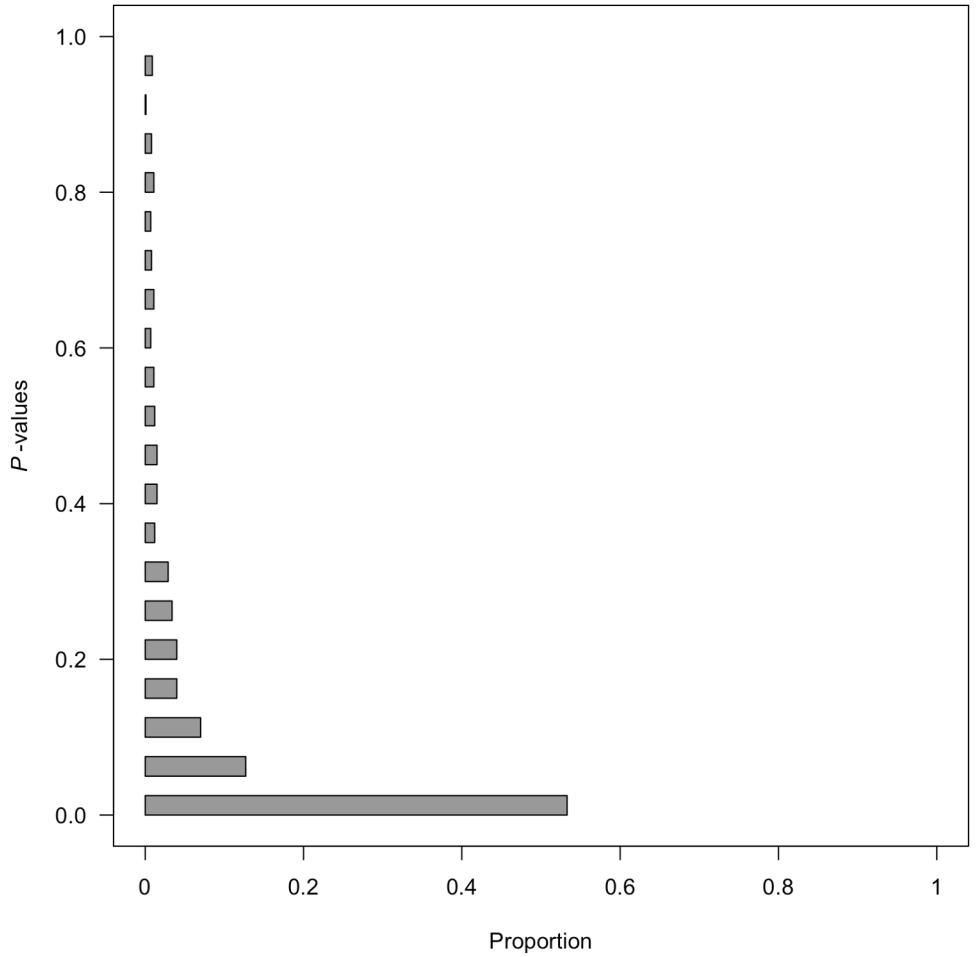
Starting point

We have a population of N = 1.000.000 where two variables are correlated with ρ = .20 (= true effect)

We draw 1000 random samples of n = 100 participants and test the correlation for significance

What we find is that in around 53% of the samples, we get a significant result at the α = .05 level

That is, we have a power (1-β) of .53 to reject the null hypothesis, if the alternative hypothesis is true



Introduction
Key terms

Starting point

We have a population of N = 1.000.000 where two variables are correlated with ρ = .20 (= true effect)

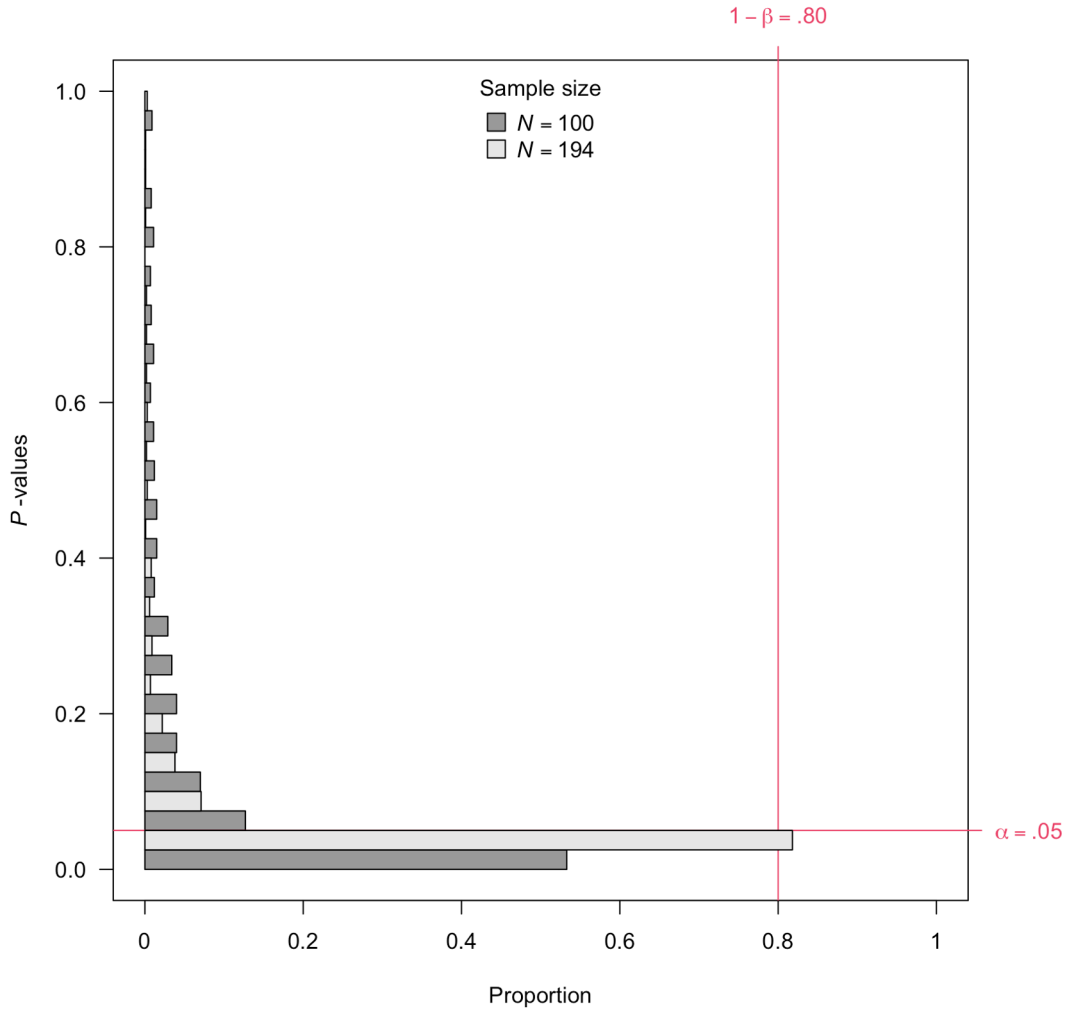
We draw 1000 random samples of n = 100 participants and test the correlation for significance

What we find is that in around 53% of the samples, we get a significant result at the α = .05 level

That is, we have a power (1-β) of .53 to reject the null hypothesis, if the alternative hypothesis is true

Power calculation

Based on the assumed effect size, we can calculate the required sample size to detect a true effect with a desired power (say 1-β = .80 $\rightarrow$ $n_{required}$ = 194)



Introduction
Key terms

Starting point

We have a population of N = 1.000.000 where two variables are correlated with ρ = .20 (= true effect)

We draw 1000 random samples of n = 100 participants and test the correlation for significance

What we find is that in around 53% of the samples, we get a significant result at the α = .05 level

That is, we have a power (1-β) of .53 to reject the null hypothesis, if the alternative hypothesis is true

Power calculation

Based on the assumed effect size, we can calculate the required sample size to detect a true effect with a desired power (say 1-β = .80 $\rightarrow$ $n_{required}$ = 194)

What do I need to understand?

An effect size is the magnitude of a relationship between variables, a group difference, explained variance etc.

In a population in which an effect of some size exists, power is the percentage of samples from that population that yield "significant results"

The smaller the population effect size, the larger is the sample size you need to have reasonable power (≥ .80) to detect this effect

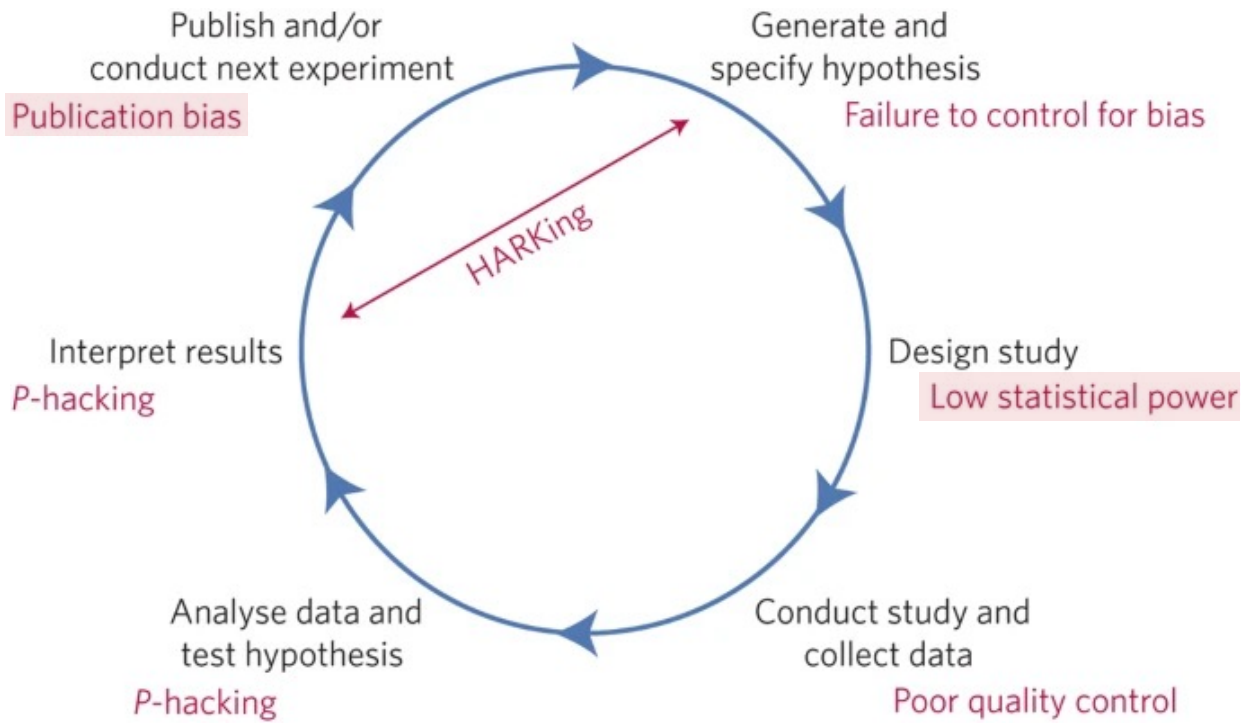
Outline

Issues to deal with
- Sample size
- Effect size
- Limited resources

Power analysis using …
- jamovi
- G*Power
- R package pwr

Further recommendations
- Further small sample disadvantages
- Power determination via simulations
- Bayesian power analysis

Summary



For cited literature and R code for examples and figures, see *Exercsies* folder!

Issues to deal with

Issues to deal with
Sample size

What are typical sample sizes?

N varies between psychological disciplines and ranges from
- two dozens (mostly in experimental psychology/imaging)
- hundreds or (rarely) thousands (correlational research)

Example from imaging research

Sample sizes in highly cited imaging studies nowadays typically between 20 to 30 (Szucz & Ioannides, 2020)

Example from social/personality psychology

Sample sizes in top journals in social/personality psychology have a median of *N* ~ 100 (Fraley & Vazire, 2014)

Numbers increase (even to several hundreds or thousands for studies with existing data), but being on your own …

What can I afford?

given a 1h experiment with two questionnaires …

an fMRI study at the NIC with 20–30 participants and scanning fees of 150 €/h and 10 € payment costs 3200-4800 € and (given a scanning time of 1h/week due to restrained access) takes 20–30 weeks to complete

a behavioral study with 100 participants in the own lab costs 1000 € and – with 10 experiments per week – takes 10 weeks

Szucz & Ioannides (2020). *NeuroImage, 221*, 117164. | Fraley & Vazire (2014). *PLoS ONE, 9*(10), e109019.

Issues to deal with
Effect size

What are typical effect sizes?

Cohen (1988) suggested that small/medium/large effects are
- for correlations: .10-.30/.30-.50/>.50
- for Cohen's *d*: 0.20-0.50/0.50-0.80/>0.80

Empirically, correlations reported in meta-analyses are much smaller with the 25th, 50th, and 75th percentiles being
- Personality psychology: .11/.19/.29 (Gignac & Szodorai, 2016)
- Social psychology: .12/.24/.41 (Lovakov & Agadullina, 2021)

Empirically informed classification of correlations
- small: .10-.20
- medium: .20-.30
- large: > .30

What can I detect?

Given an *N* of 30 or 100, α = .05 and 1–β = .80, one can detect
- *N* = 30: *r* ≥ .49 (a large effect)
- *N* = 100: *r* ≥ .28 (a medium effect)

As (small to) medium effects are far more likely than large effects, typical sample sizes in (not only) neuroimaging have low power

If you detect a large effect in a small sample, *it had to be large to be detected* (and likely is smaller in the population)!

Gignac & Szodorai (2016), *Pers Ind Diff, 102*, 74-78. | Lovakow & Agadullina (2021). *Eur J Soc Psychol.*

Issues to deal with
Effect size

Which size of an effect can one expect?

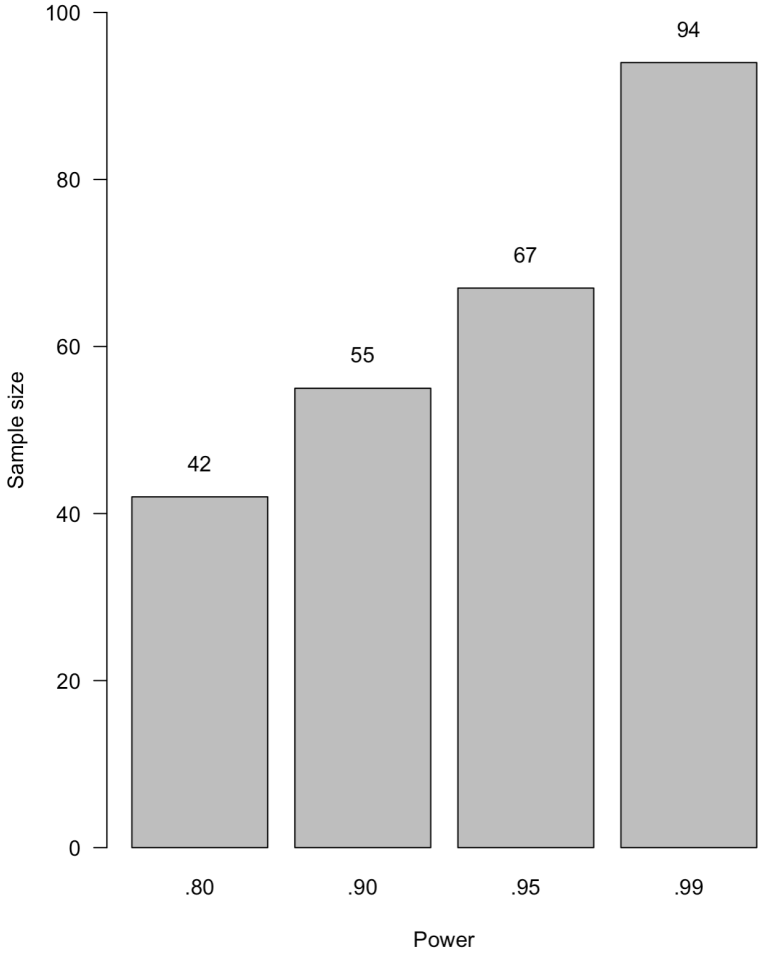
If one has no idea what size an effect could have, it is wise to assume that it is small to medium, i.e., for correlations .10 to .30, or .20 as a rule of thumb (see also Fraley & Vazire, 2021)

What if we have an established effect of a certain size?

One could (and should) use this effect to estimate the sample size required to detect an effect of this size

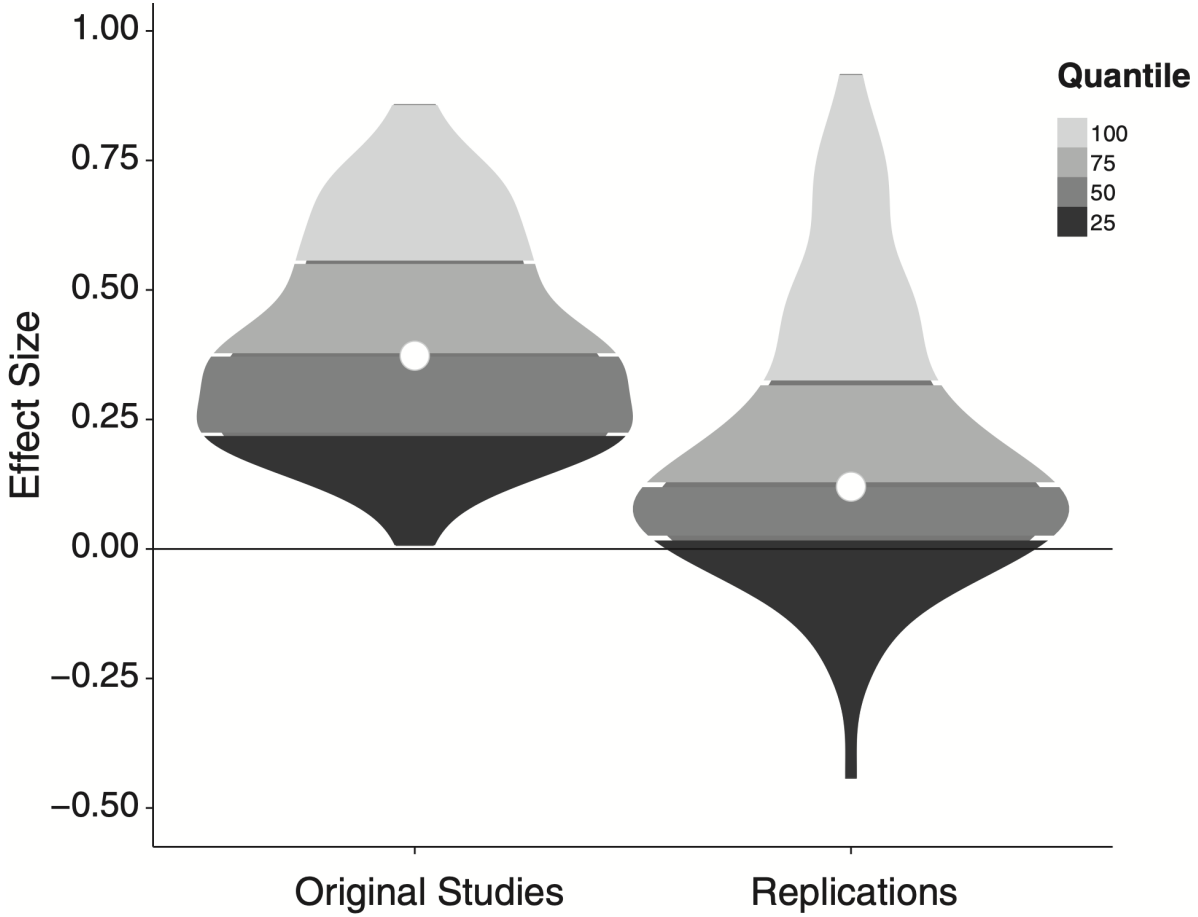
Example

An effect in question has been reported to be r = .42. Using the R package *pwr* (see below), a sample size of N = 42 is needed to detect a correlation of this size at α = .05 and 1–β = .80. With a higher power, say, 1–β = .99, sample size increases to N = 94



Fraley & Vazire (2014). *PLoS ONE, 9*(10), e109019.

Issues to deal with
Effect size



What can I rely on?

Evidence from replication research (e.g., Open Science Collaboration, 2015) suggests that even replicable effects are about half of the size as those originally reported

Therefore, one should use half of the reported effect size for power calculation. On our example, with an estimated r = .42/2 = .21, the sample sizes would then increase to N = 175 (80% power) or N = 288 (95% power)

Fig,. 1B from Open Science Collaboration (2015). *Science, 349*(6251), aac4716.

Issues to deal with
Limited resources

What is affordable within a (PhD) project?

Third-party funded PhD projects usually are financed for two (DFG projects) to four (CRC projects) years, you can expect funding for at least one further year from the same or other funding sources

For PhD projects funded from core support (i.e., *Haushaltsstelle*), there may be more time available (depending on your supervisor's policies)

Doing two to three studies with adequate power in this short time may be unrealistic (even for a six-year period). Also, the available financial and personnel resources may not be sufficient. This is a more severe problem for imaging than for behavioral of questionnaire studies

What should I do in this case?

Try to convince your supervisor to re-allocate resources to run at least one adequately powered study within your PhD project

If you are obliged to run studies that have too low power (and therefore come at the cost of low replicability), do at least employ open science practices such as preregistration, open data and code etc. (thereby ensuring research transparency) and discuss the power issue!

Power analysis using
jamovi, G*Power, and pwr

Power analysis
jamovi

jamovi isn't the first instance to consult for power analysis

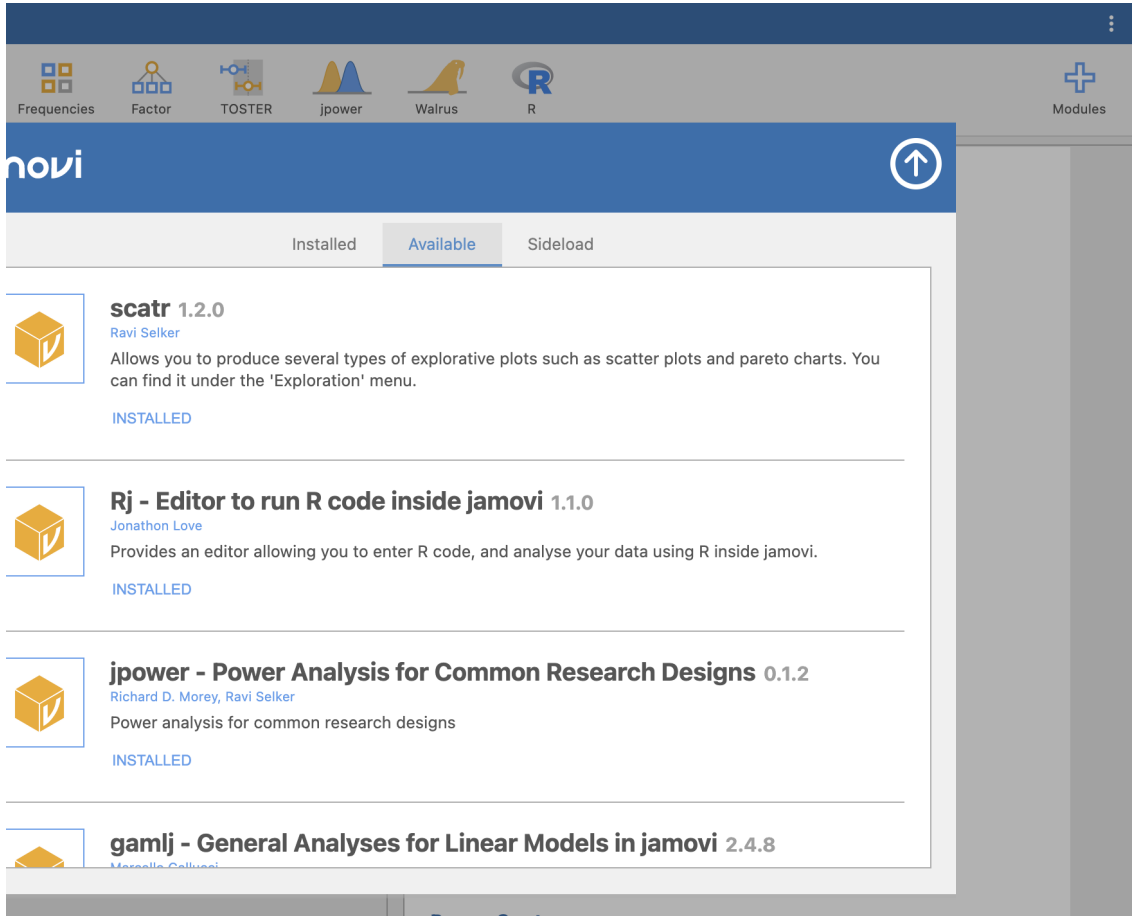
So far (Sept 2021) it can handle only *t*-tests

Yet, it has an intuitive GUI and …

- explains the results of the power analyses performed in a comprehensible manner
- provides easy-to-grasp plots
- leaves you with the (rightful) impression that you finally understood power analysis

How to run power analysis with jamovi?

In the *Analysis* panel on the right there is a + symbol, click on it and then choose *jamovi library* and among the options choose *jpower*



https://www.jamovi.org

Power analysis
jamovi

How to run it?

Just download it (it's free) from

https://www.jamovi.org/download.html

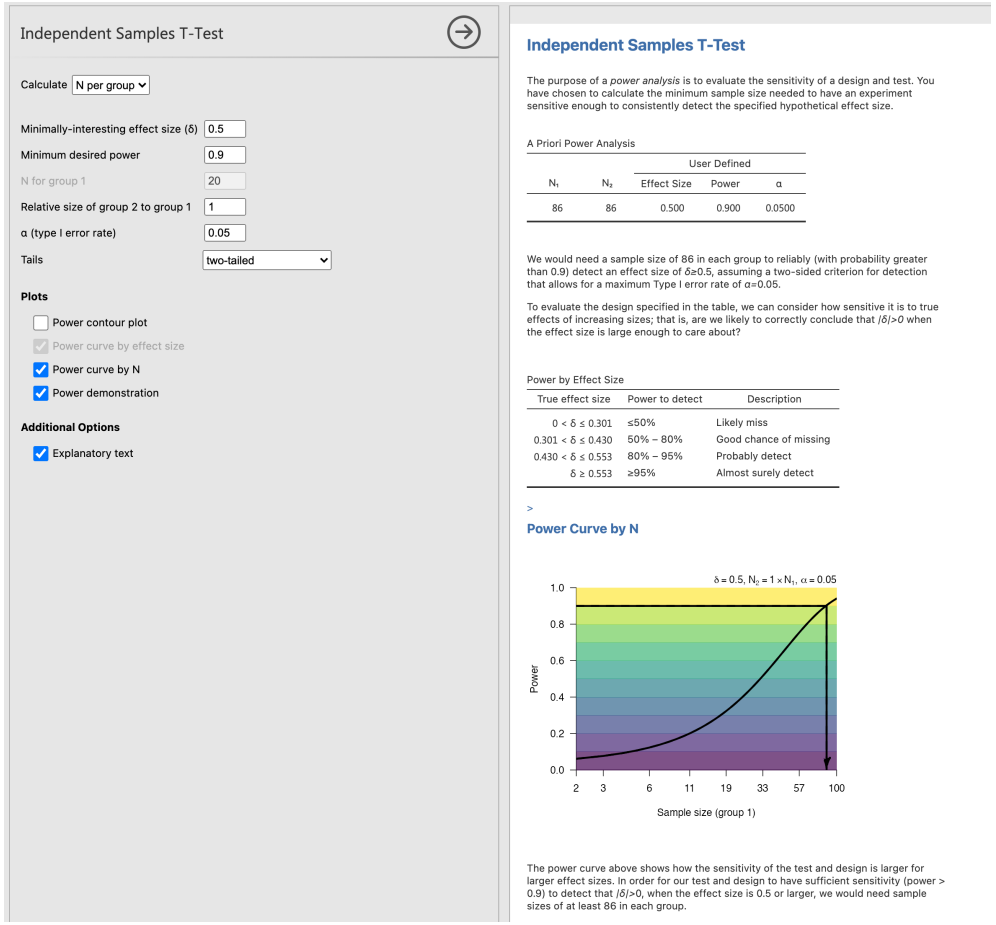
install the *jpower* module and try it

Example

Here is a screenshot from jamovi where a minimally interesting effect size is defined together with the minimum power to achieve

The defaults are a medium effect size and a power of .90 (can't never be wrong with that)

You can choose among different plots that are fully explained



https://www.jamovi.org

Power analysis
jamovi

Benefits

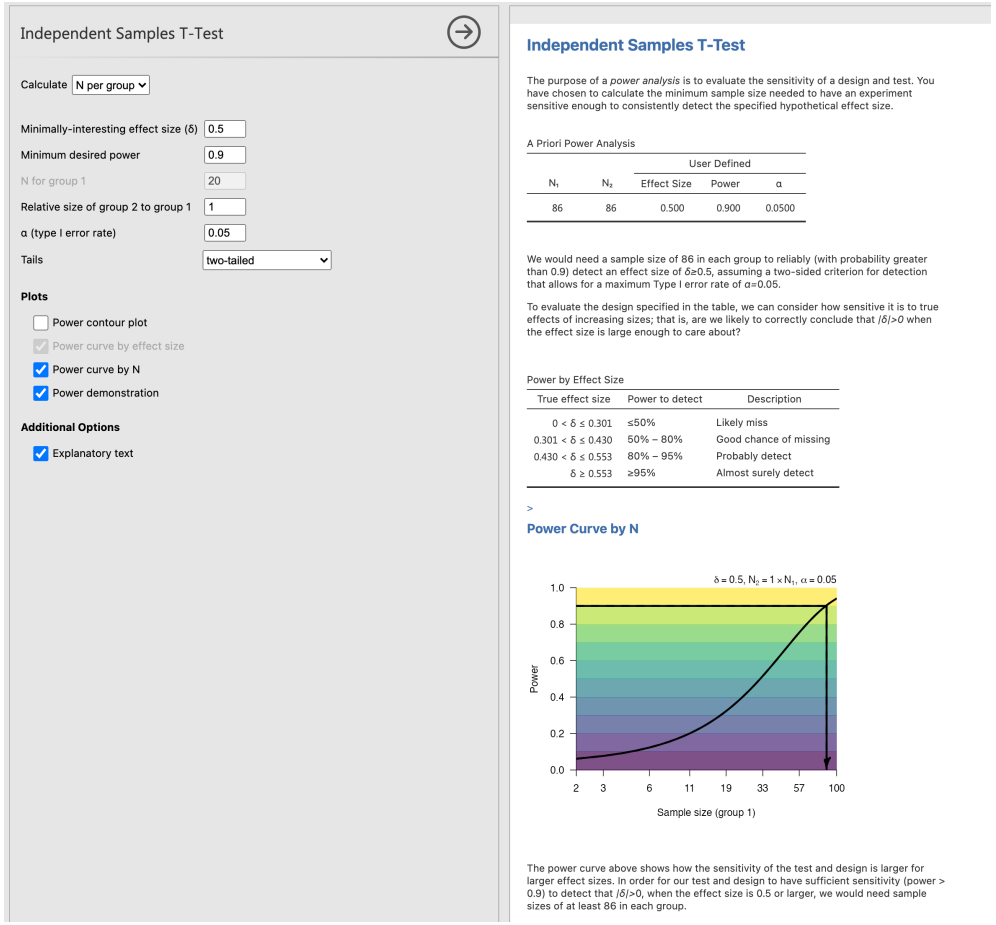
— enables you to better understand the essence of power analysis

— nice graphs

Drawbacks

— not commonly used

— can handle only *t*-tests



https://www.jamovi.org

Power analysis
G*Power

G*Power *is* the first instance to consult for power analysis

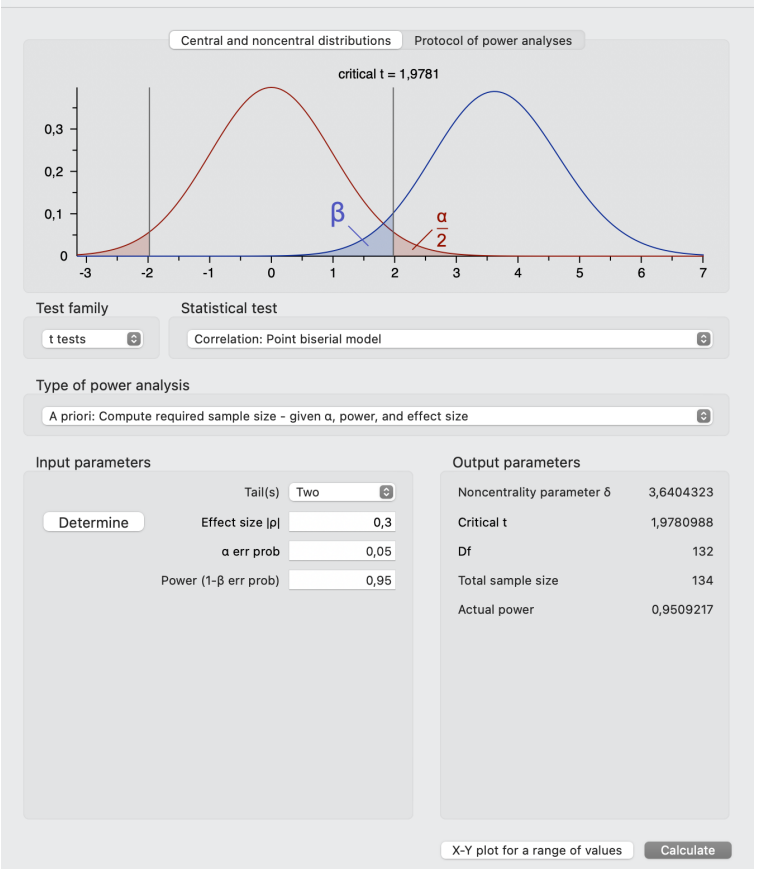
as it handles almost all common statistical tests (and it's free as well, download link below)

It also has a GUI …

yet, it is less intuitive, you need to know what you are doing

you have to be aware that you need to know a lot of things (e.g., what effect size f^2 means)

but in the end, it all boils down to have some idea what amount of variance an effect could or should explain (be it R^2 or η^2)



https://www.psychologie.hhu.de/arbeitsgruppen/allgemeine-psychologie-und-arbeitspsychologie/gpower

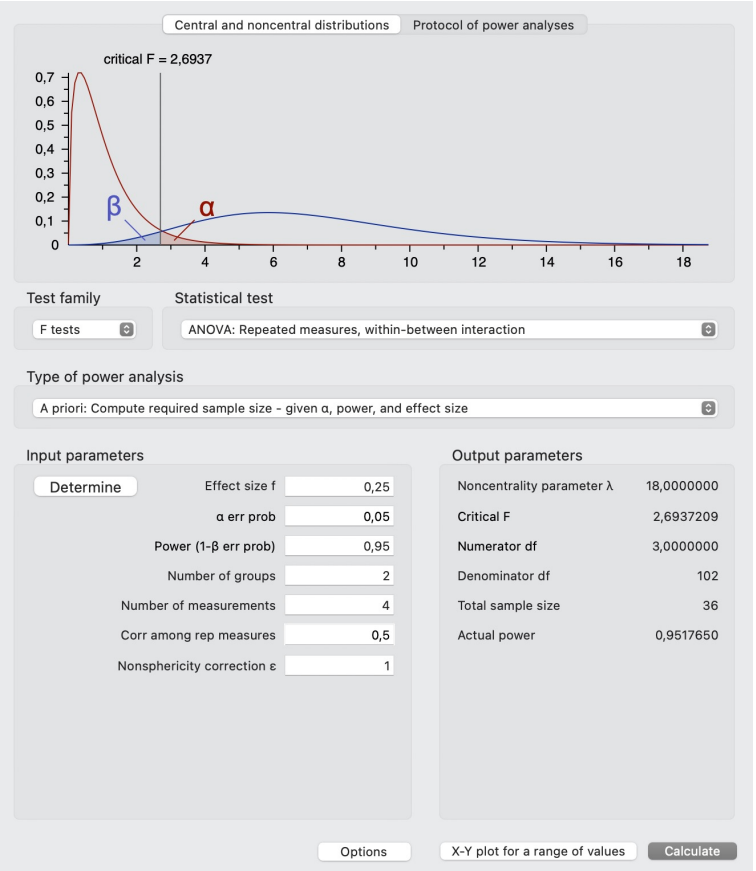
Power analysis
G*Power

Per default, G*Power comes with …

— a medium effect size (according to Cohen, 1988)
— a strong assumption of the power to achieve (.95)
— an option to determine the effect size in several ways

How to use G*Power as a novice

— try to navigate through the available tests, e.g., F tests
— if you have an *n*-back task and want to examine condition differences, use *ANOVA: Repeated measures, within factors*
— if you are interested in age differences in performance in the *n*-back task, use *ANOVA: Repeated measures, between factors*
— if you assume that age has different impact on performance across conditions, choose *ANOVA: Repeated measures, within-between interaction*



See video tutorials by Alexander Swan, e.g.,
https://www.youtube.com/watch?v=FelgUtL-8Sg | https://www.youtube.com/watch?v=51qGy5XRmy8

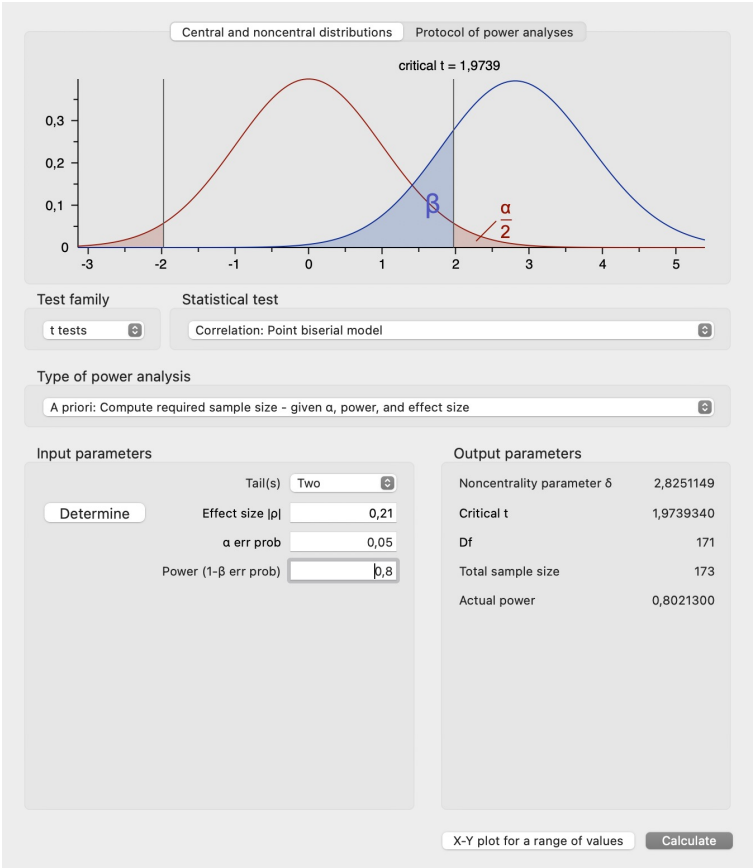
Power analysis
G*Power

Example

We want to determine the sample size needed to detect a correlation of r = .21 at α = .05 and 1– β = .80 (see above).

We enter the respective values as input parameters on the left and click *Calculate* – the output parameters are shown on the right.

Why does the sample size needed slightly differs from that given above? Apparently, the *pwr* package employed there uses a somewhat simplified algorithm compared to G*Power

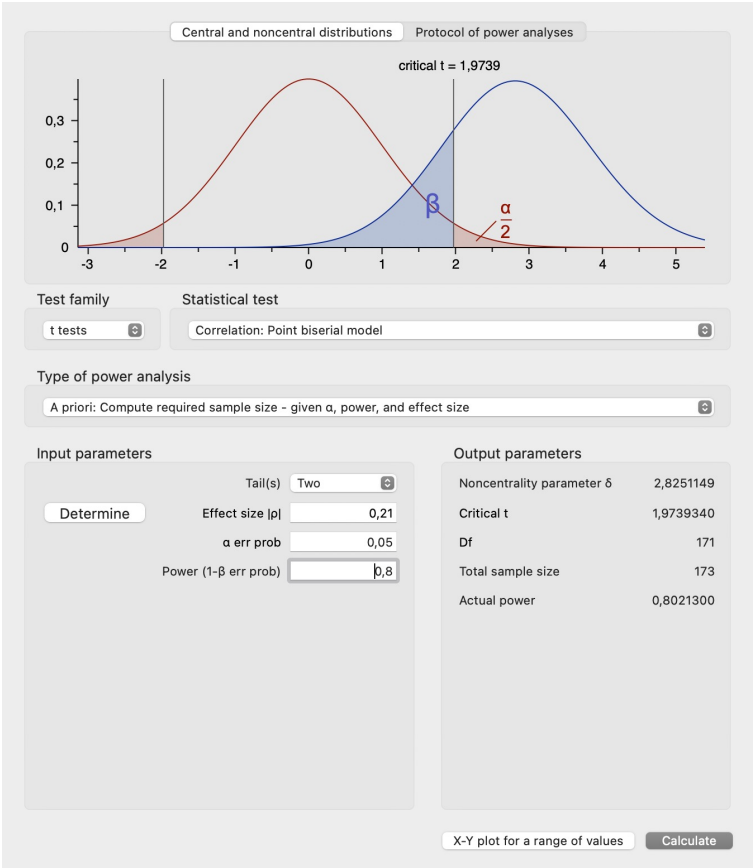


See video tutorials by Alexander Swan, e.g.,
https://www.youtube.com/watch?v=FelgUtL-8Sg | https://www.youtube.com/watch?v=51qGy5XRmy8

Power analysis
G*Power

Selected further calculation options

- **achieved power, given α, sample size and effect size:** useful when you explore an existing data set and find some interesting effect – what power did you actually have?
- **required effect size, given α, sample size and power:** useful when your sampling aim was to get as many participants in a given time – which effect size can you detect?



Benefits

- most commonly used software for power analysis
- extremely powerful with power analysis for a multitude of statistical tests

Drawbacks

- one might get lost in this multitude of options
- quite a bit of knowledge and prior information required

See video tutorials by Alexander Swan, e.g.,
https://www.youtube.com/watch?v=FelgUtL-8Sg | https://www.youtube.com/watch?v=51qGy5XRmy8

Power analysis

pwr

pwr is the tool to go for if you …

- are using R for your analyses anyway (although there are certainly further packages for power analysis)
- want to run simulations and create fancy plots
- want to share code without having others to install yet another software (although one should have installed G*Power anyway)
- don't care about GUIs

Options

quite the same as G*Power, but less statistical tests available

Example: *N* needed to detect *r* = .21

```
pwr.r.test(r = .21, sig.level = .05, power = .80)
```

Package 'pwr'

March 17, 2020

```
approximate correlation power
calculation (arctangh
transformation)

              n = 174.8439
              r = 0.21
      sig.level = 0.05
          power = 0.8
    alternative = two.sided
```

Power analysis

pwr

pwr is the tool to go for if you …

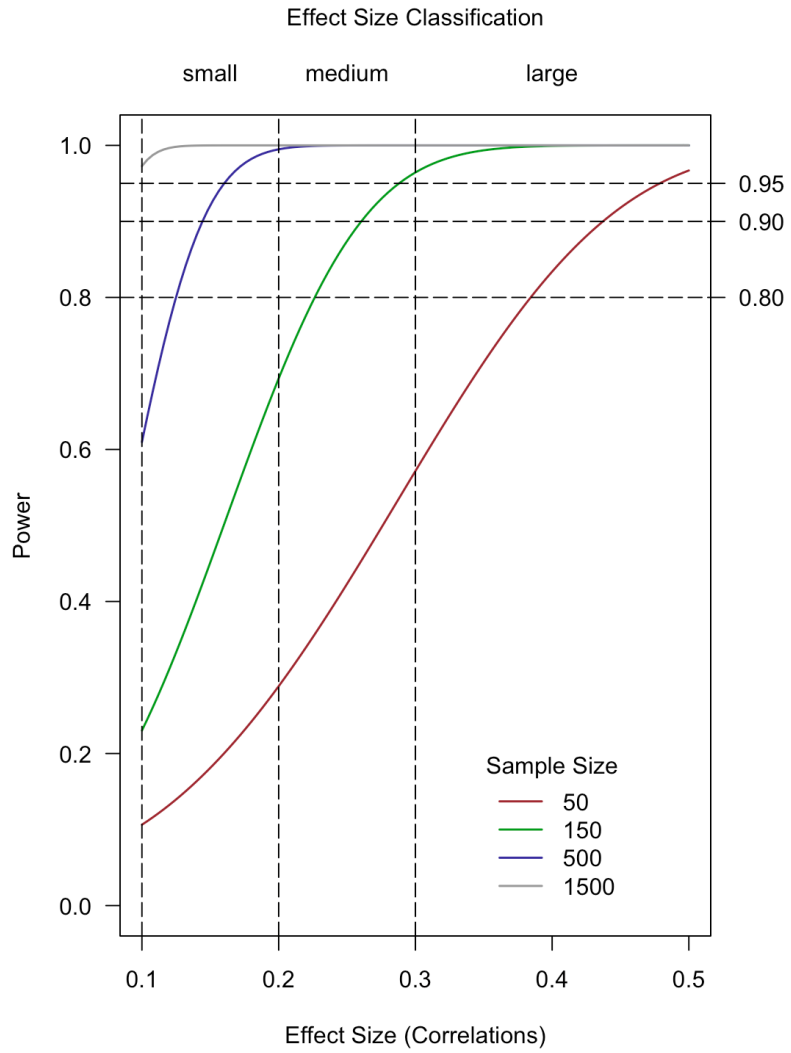
- are using R for your analyses anyway (although there are certainly further packages for power analysis)
- want to run simulations and create fancy plots
- want to share code without having others to install yet another software (although one should have installed G*Power anyway)
- don’t care about GUIs

Options

quite the same as G*Power, but less statistical tests available

Example: *N* needed to detect *r* = .21

```
pwr.r.test(r = .21, sig.level = .05, power = .80)
```



Further recommendations

Further recommendations
Further small sample disadvantages

Power is not the only reason to have large samples

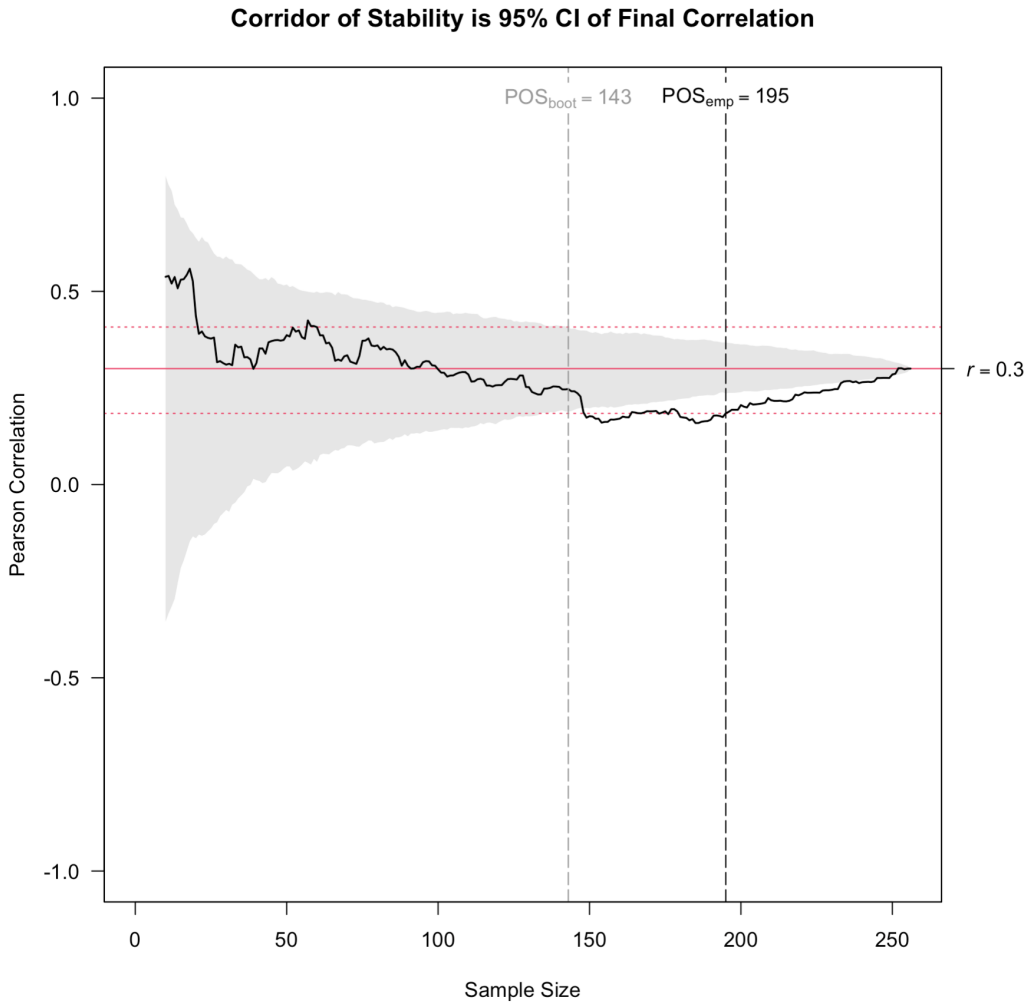
Unplanned exploratory analyses: You might have 80% power for detecting $r \geq .30$ (requiring $N \geq 84$), but then you want to run an exploratory hierarchical regression and test significance of R^2 increase (requiring $N \geq 100$)

Further recommendations
Further small sample disadvantages

Power is not the only reason to have large samples

Stability of effect size estimates: You might be interested in a stable estimate of an effect (e.g., for predictions). In small samples, you cannot be sure whether, say, a correlation is a stable estimate of the population correlation (Schönbrodt & Perugini, 2013)

The figure shows that a final r = .30 in N = 256 fluctuates considerably while the sample comes in. It stabilizes within the CI of the final correlation (= point of stability, POS) not before n = 195 have been sampled or – if sampled in a different order (grey area: evolution of correlation in bootstrapped sample order) – still beyond what your power analysis told you to be adequate to detect $r \geq .30$ at α = .05 and 1-β = .80 (a said, $N \geq 84$)



Schönbrodt & Perugini (2013). *J Res Pers, 47*, 609-612.

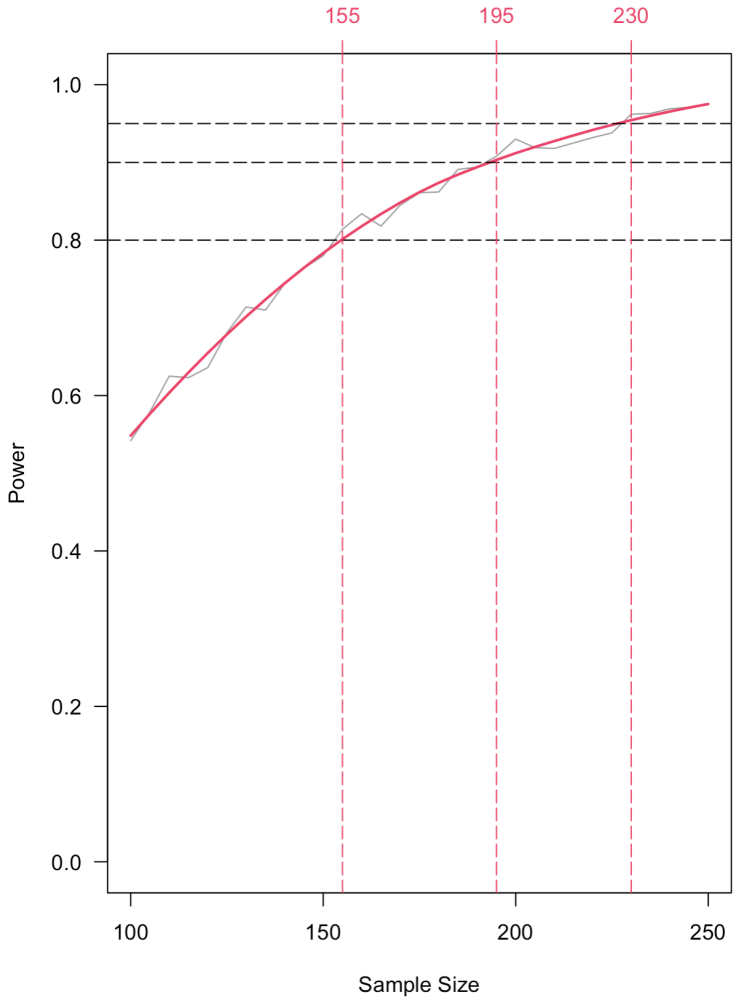
Further recommendations
Power determination via simulations

The problem and one solution

You may want to use a statistical procedure for which there is no agreed upon power analysis routine

An example for correlations would be Kendall's τ, where you assume τ = .20 and create a population where this correlation exists. You could think of some reasonable range of sample sizes *n* needed, draw 1000 samples for each *n* and plot the proportion of significant results in these samples against the respective sample size

Then you run a LOESS fit over the proportions and obtain the minimum *N* needed to yield a significant result at a power of 80, 90, and 95% (see figure on the right)



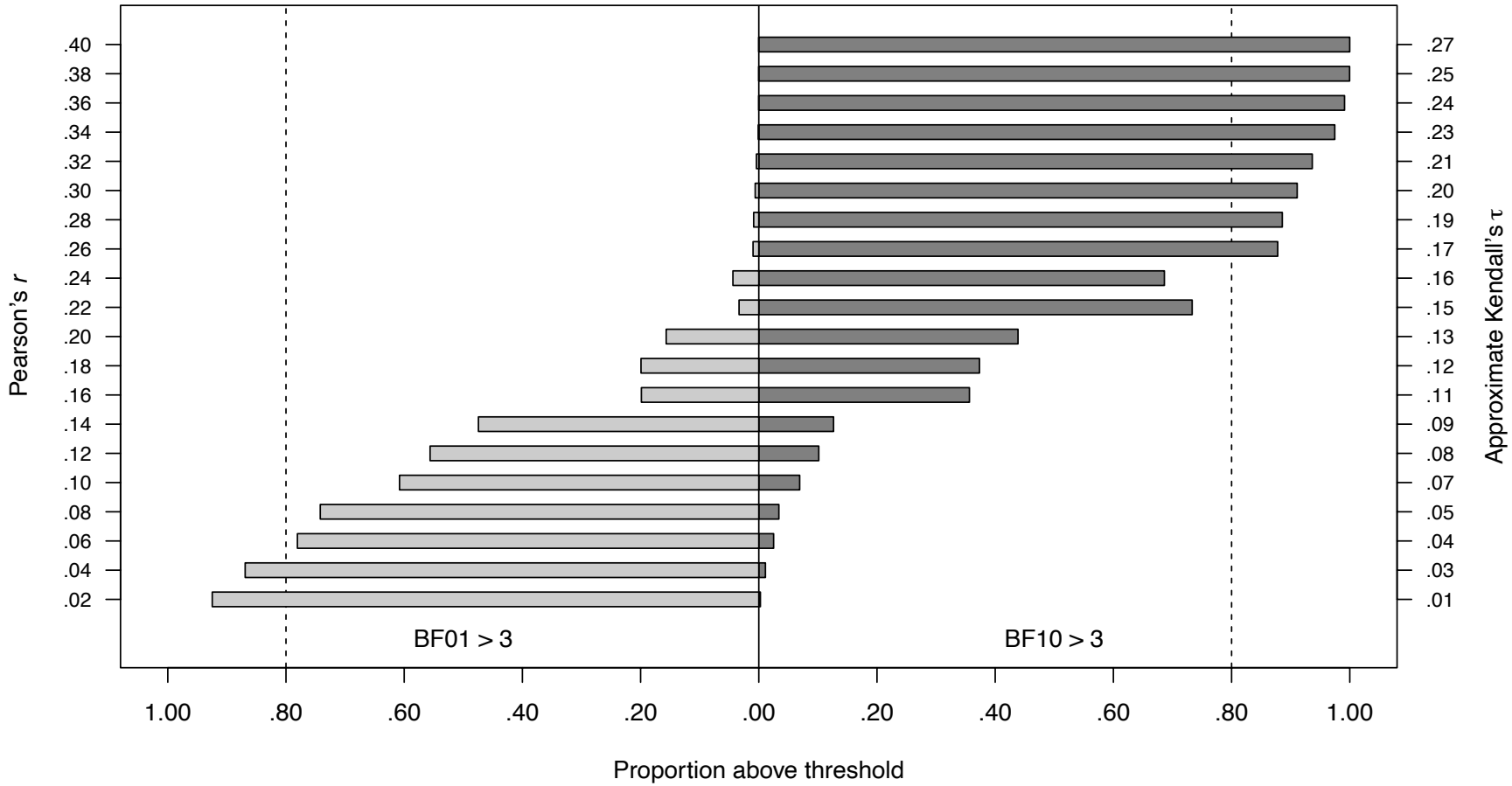
Further recommendations
Bayesian power analysis

Bayesian statistics …

… are relatively new, so usually there is no agreed upon power analysis routine (yet)

Given that you have a sample of *N* = 189, you are interested in the power you have to achieve at least moderate evidence for the *H1* (i.e., BF10 > 3) or the *H0* (i.e., BF01 > 3)

The right figure shows it …



Gärtner, Grass et al. (2020). *J Pers, 89*, 1113–1125.

Summary

Summary
Power analysis

If the literature tells you what effect size to expect …

This effect size is most likely inflated due to publication bias, so divide it by 2 and use that estimate for your power analysis, with an assumed power of at least 80% and a significance level that accounts for possible multiple testing

If you have no idea what effect size to expect …

Cohen's classification most likely will not reflect the typical effect sizes in your area of research. If there are no established guidelines (such as those of Gignac & Szodorai, 2016, for individual differences research), assume a correlation of r = .20 (or any derivative such as an explained variance of .04, see Fraley & Vazire, 2014). A small to medium effect is more likely than a large one!

The software you use for power calculations virtually makes no difference …

Yet, G*Power is more powerful than other software

If there is no power analysis software for your specific effect size, run simulations …

Gignac & Szodorai (2016), *Pers Ind Diff, 102*, 74-78. | Fraley & Vazire (2014). *PLoS ONE, 9*(10), e109019.

Thank you!

Exercises

1) What if the paper you want to follow up on does not report effect sizes?

Try to find out how to convert the reported information/statistics into an effect size (helpful resource for t-tests and ANOVAs: Lakens, 2013; helpful R package: *effectsize*)!

2) What to do with an empirical result that relies on a small sample?

In a paper, a sample of $N = 30$ was examined, split into two groups and a *t*-test was conducted. What is the effect size (in terms of Cohen's *d*) that can be detected with such a sample size at $\alpha = .05$ and $1\text{-}\beta = .80$? Is this effect size a reasonable one to assume before conducting an experiment?

How many participants would you need if you divide this effect size by two for power analysis?

3) What if you cannot afford more than a small sample?

What is the power to detect a correlation of $r = .30$ in a sample of $N = 30$ given that you have to correct for multiple testing because you need to perform 5 tests on the same hypothesis?

Given the result, is it worth the effort to run that study at all?