

***How to get it right I
(aka the signal and the noise):
Why you should think twice
before planning your next
study***

Marco Perugini
Milan, 16/12/2020

The problem

- Assume that, as scientists, we all want to get it right
 - What can we do to increase our chances?
 - a) Get it right \neq I am right
 - b) Get it right \neq Get it published
-

Outline

- **Replicability in Psychology (*sneak preview*)**
 - **Sample planning**
 - **Refresh of a few basic statistical concepts**
 - **Issues in Power analysis**
 - **Uncertainty**
 - **Sensitivity**
 - **Within vs. Between designs**
 - **A first tip for getting it right**
-

Replicability in Psychology

Why now?

- In the last decade, the issue of replicability has become central in Psychology (and Science)
- Many developments in research methodology
- Rapid changes in standards for research
- Rapid changes in standards for publishing
- What is the problem, what do you mean exactly by replicability, what can we do about it, why now?
- This will be the main issue of my Open Science lecture in February

Good things come to those who wait

Meanwhile...



Sneak preview

Why many effects are not replicated?

- A mix of different factors and possible explanations
- Two main factors
- **Low power** and **publication bias**
- Under these conditions, it is predictable that there will be many results in the literature that are difficult to replicate
- We will get back to this issue later

delayed sneak preview...

Power analysis

- You already know what is power and power analysis
 - We need first to have a sense in what context power analysis can be useful
 - ... and to double-refresh a few basic statistical concepts
 - ... and, finally, we will articulate a few specific issues linked to power analysis
-

Sample planning

Sample planning

- When you plan a study/research/intervention, you should think about the participants that you need

Some basic issues

- Representativeness
 - Generalizability
 - Robustness
 - Feasibility
 - Efficiency
-

Representativeness

- **Match (reduce gap) between what you will see and what you would like to say**
 - **What you will see:** data (behaviors, evaluations, physiological responses, etc.) from some participants. Who are these participants? Stratified sample? Specific sample? Random sample? Convenience sample?
 - **What you would like to say:** something about humans? students? working people? people with clinical problems?
 - The validity of your inference from the results derived in your sample to a certain “population”
 - Beware of the possible gap. Especially if you go for concrete applications in real life (e.g., interventions)
-

Generalizability

- **How much what you will say based on what you will see goes beyond the context in which you are saying it**
 - Study on sample of psychology students about prejudice. How much what you find can be generalized to workers in the supermarket? to retired people? to people living in a small village or a big city?
 - A form of stratified sample is desirable
 - Beware of the possible gap (especially for interventions)
-

Applicability to real world

- Be careful for applicability to real world
- Psychology has not developed yet a robust and established translational tradition of results

NATURE HUMAN BEHAVIOUR | VOL 4 | NOVEMBER 2020 | 1092-1094 | www.nature.com/nathumbehav

comment

Check for updates

Use caution when applying behavioural science to policy

Social and behavioural scientists have attempted to speak to the COVID-19 crisis. But is behavioural research on COVID-19 suitable for making policy decisions? We offer a taxonomy that lets our science advance in 'evidence readiness levels' to be suitable for policy. We caution practitioners to take extreme care translating our findings to applications.

Hans IJzerman, Neil A. Lewis Jr., Andrew K. Przybylski, Netta Weinstein, Lisa DeBruine, Stuart J. Ritchie, Simine Vazire, Patrick S. Forscher, Richard D. Morey, James D. Ivory and Farid Anvari

Social and Behavioural Science

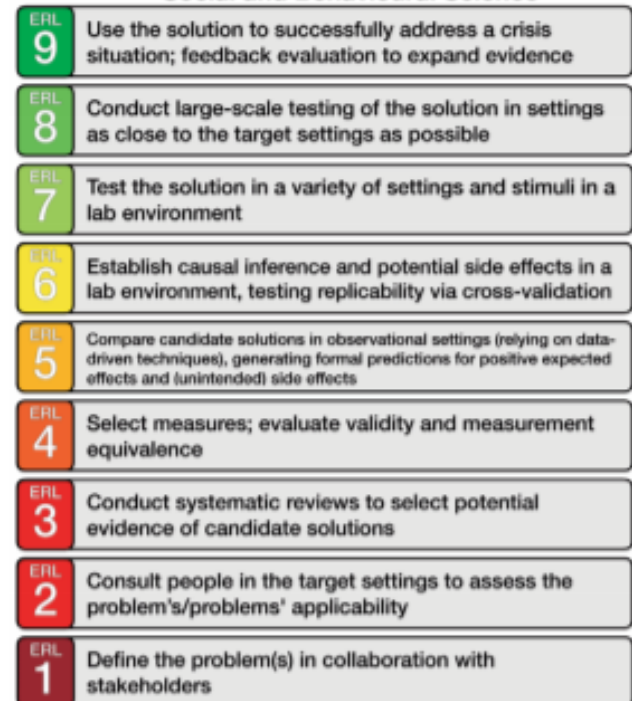


Fig. 2 | Proposed social and behavioural sciences evidence readiness levels.

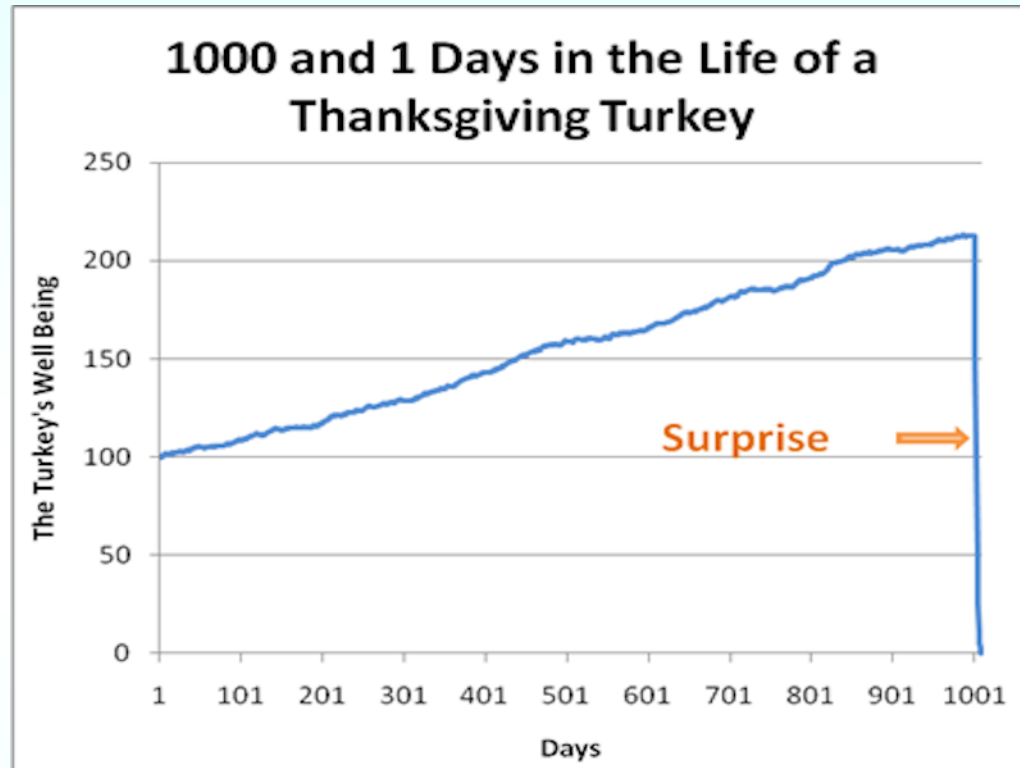
Robustness

- **How much what you will say based on what you will see will be robust (e.g., in future studies you or others will find similar results?)**
 - Everything else being equal, do you trust more results from a study with 50 Ss or from a study with 5000 Ss?
 - Attention to the uncertainty of the inference, both the one that you can estimate statistically (**known unknown**) and the one that you cannot estimate (**unknown unknown**)
-

The real problem (Unknown unknown)

- The world is uncertain
 - Knowledge is imperfect
 - We deal with “samples” rather than “population”
 - We try to make inferences from them
 - We try to predict what will happen based on what has happened and on the regularities that we are able to learn from that
-

Bertrand Russell's turkey



“Essentially, all models are wrong, but some are useful” (Box & Draper, 1987)

Feasibility

- **How much what you would like to know can be known with the resources that you have**
 - There are always logistical constraints (time, money, people, space)
 - Attention to the feasibility of what you would like to do
 - Ask questions to “Nature” that can be reasonably answered within your “budget”
-

Efficiency

- **The minimum (or optimal) effort needed to know what you would like to know**
 - Sometimes there are high costs involved in research
 - Sometimes you could be in the position to ask yourself what is the minimum data needed to answer in a reasonable way your question
 - Sometimes you might try to go for the optimal number
 - More data is always better than less data but the informational value of every additional data decreases over a certain point
 - Costs/benefits logic
-

Mindsets for sample planning

- **Accuracy:** collect as many participants as needed to have a certain level of accuracy in your parameter estimation
 - **Efficiency:** collect as few participants as needed to reach the conclusion that you want to reach
 - **Redundancy:** collect as many participants are needed to reach a reliable conclusion concerning what you want to reach
-

Some statistical approaches to sample size planning

- ***Accuracy***
AIPE (Maxwell, 2008): decide sample size based on a chosen level of Accuracy In Parameter Estimation
 - ***Efficiency***
Sequential designs:
Frequentist (Lakens, 2014): Start with a planned N and number of interim tests, add N if needed (but adjust alpha)
Bayesian (Schonbrodt et al., 2017, 2018): Start with a minimum N , add N until BF reaches a pre-defined threshold
 - ***Efficiency/Redundancy***
Heuristic: in different fields there are “magical” rules ($N \geq 20$ per cell, $N > 100$, ratio k/N). At best, approximate wise suggestion, at worse misleading
 - **Power analysis**
-

Some (advanced) references

Sample Size Planning for Statistical Power and Accuracy in Parameter Estimation

Scott E. Maxwell,¹ Ken Kelley,²
and Joseph R. Rausch³

¹Department of Psychology, University of Notre Dame, Notre Dame, Indiana 46556;
email: smaxwell@nd.edu

²Inquiry Methodology Program, Indiana University, Bloomington, Indiana 47405;
email: kkii@indiana.edu

³Department of Psychology, University of Minnesota, Minneapolis, Minnesota 55455;
email: rausch@umn.edu

European Journal of Social Psychology, Eur. J. Soc. Psychol. **44**, 701–710 (2014)

Published online in Wiley Online Library (wileyonlinelibrary.com) DOI: 10.1002/ejsp.2023

Special issue article: Methods and statistics in social psychology: Refinements and new developments

Performing high-powered studies efficiently with sequential analyses

DANIËL LAKENS*

Human Technology Interaction Group, Eindhoven University of Technology, Eindhoven, The Netherlands

Annu. Rev. Psychol. 2008. 59:537–63

Key Words

Psychon Bull Rev (2018) 25:128–142
DOI 10.3758/s13423-017-1230-y



CrossMark

Psychological Methods
2017, Vol. 22, No. 2, 322–339

© 2015 American Psychological Association
1082-989X/17/\$12.00 http://dx.doi.org/10.1037/met0000061

BRIEF REPORT

Bayes factor design analysis: Planning for compelling evidence

Felix D. Schönbrodt¹ · Eric-Jan Wagenmakers²

Sequential Hypothesis Testing With Bayes Factors: Efficiently Testing Mean Differences

Felix D. Schönbrodt
Ludwig-Maximilians-Universität München

Eric-Jan Wagenmakers
University of Amsterdam

Michael Zehetleitner
Ludwig-Maximilians-Universität München

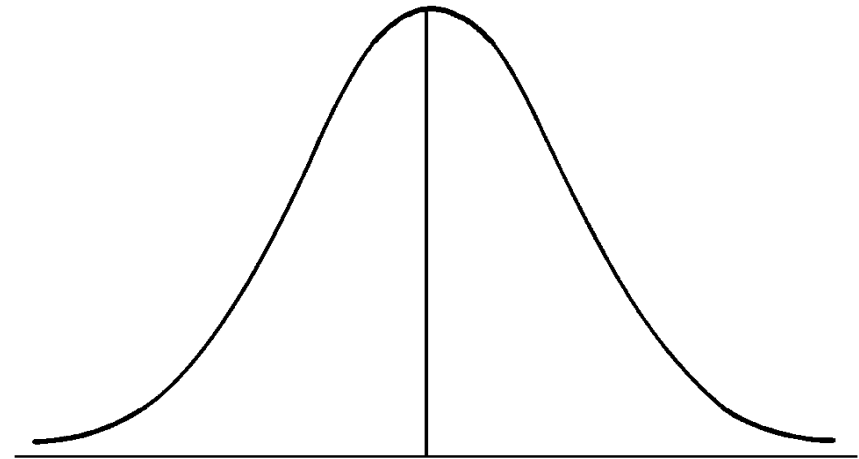
Marco Perugini
University of Milan–Bicocca

*A refresh of already fresh
basic statistical concepts*

Mean

- A single value that reflects the central point of a distribution
- If the distribution is normal, it is also the best simple way to summarize it

$$\bar{X} = \frac{\sum X_i}{N}$$

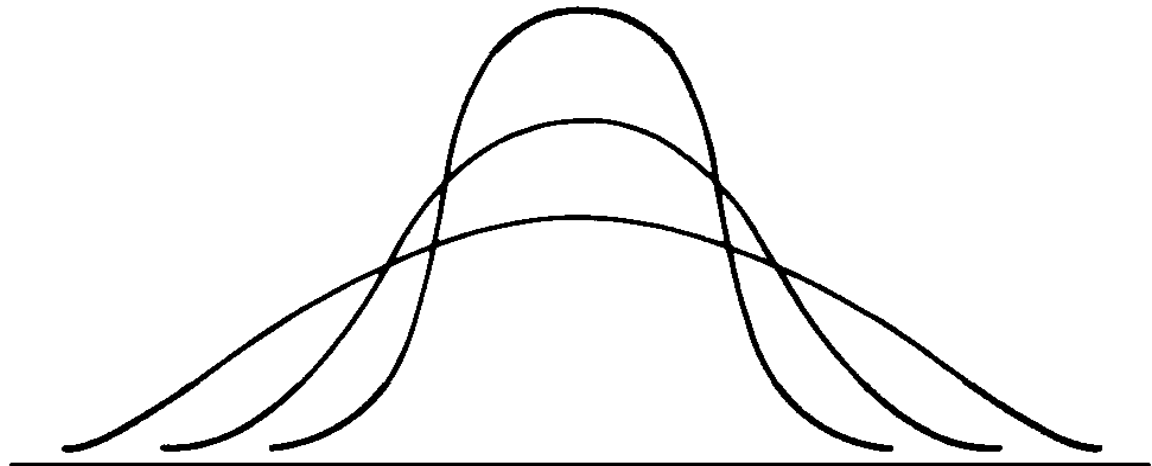


Variance and standard deviation

- Reflects the dispersion (variability) around the mean

$$s^2 = \frac{\sum (X_i - \bar{X})^2}{N} = \frac{\sum X^2}{N} - \bar{X}^2$$

$$s = \sqrt{s^2}$$



Standard error

- When we measure something, more data means less measurement error
- Exit polls are more accurate (less error) the more the sampled voters or polling stations
- We have a sample but would like to say something about the underlying population (or anyway something that generalizes beyond that sample)

Standard error and variance

- Standard error does not depend only from how big is a sample size but also from the variability (variance) of the study object
- If everyone answers in the same way, one needs to ask to only one person...
- If people have very different opinions, one need many of them to be able to say something about «what they think»...
- Standard error provides a link between sample and population

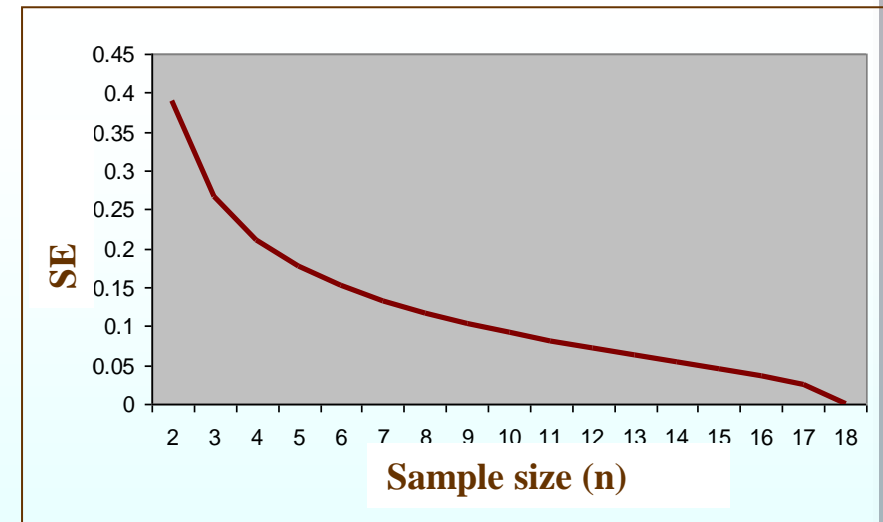
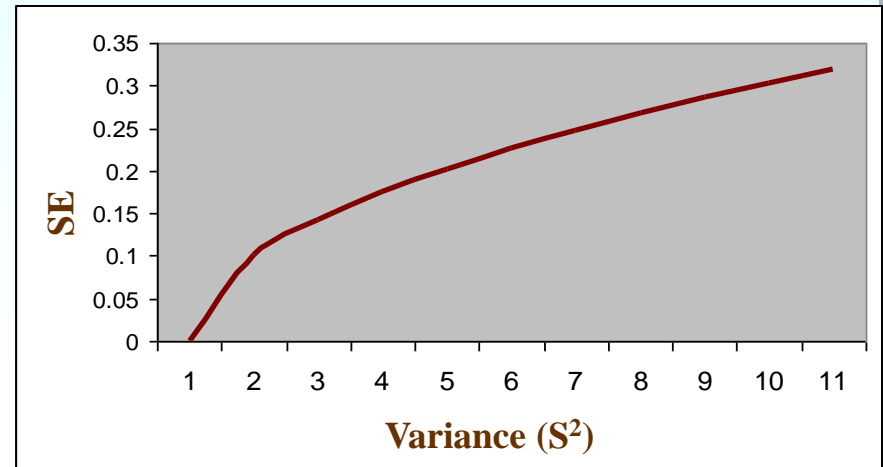
Standard error

- Error in estimating a population parameter (e.g., mean) from a sample

Goes up with increasing variance

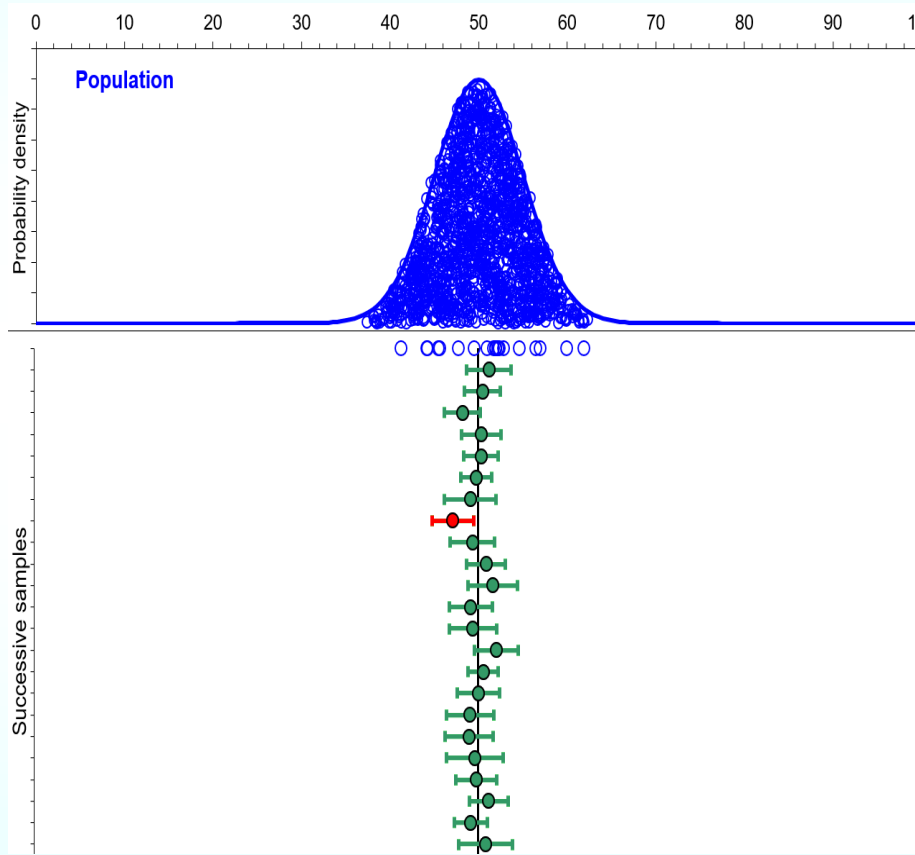
$$SE = \sqrt{\frac{S^2}{n}}$$

Goes down with increasing sample size



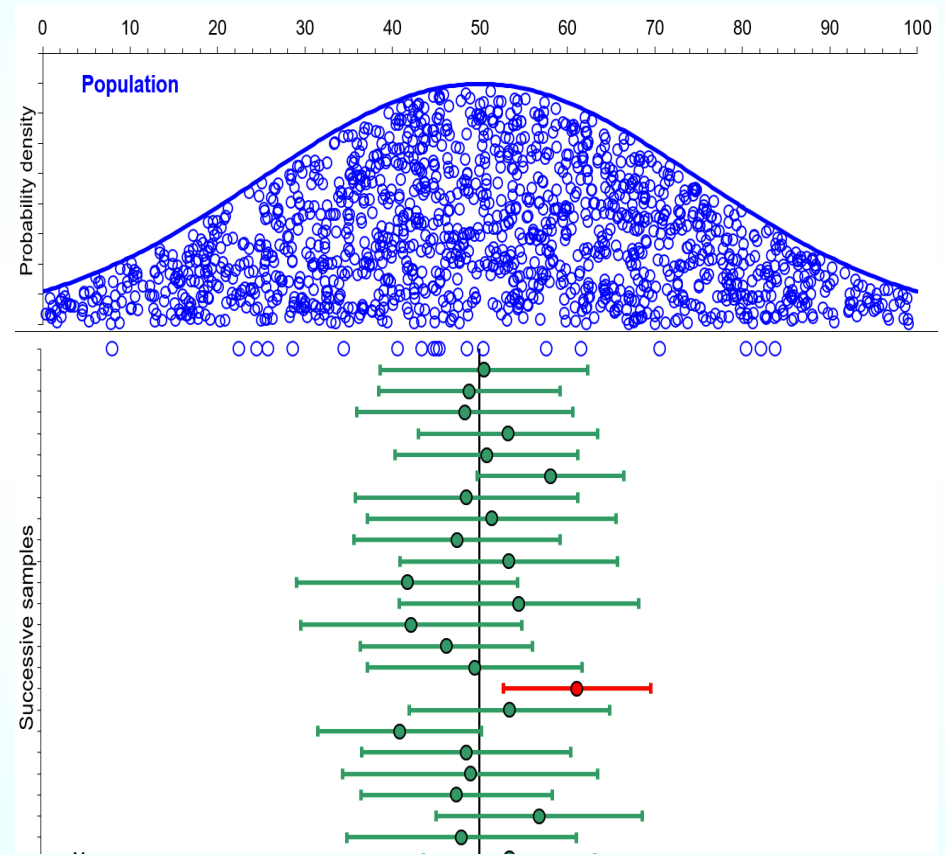
Parameter estimation: Error and variability

N=20



Small variability = small SE

N=20

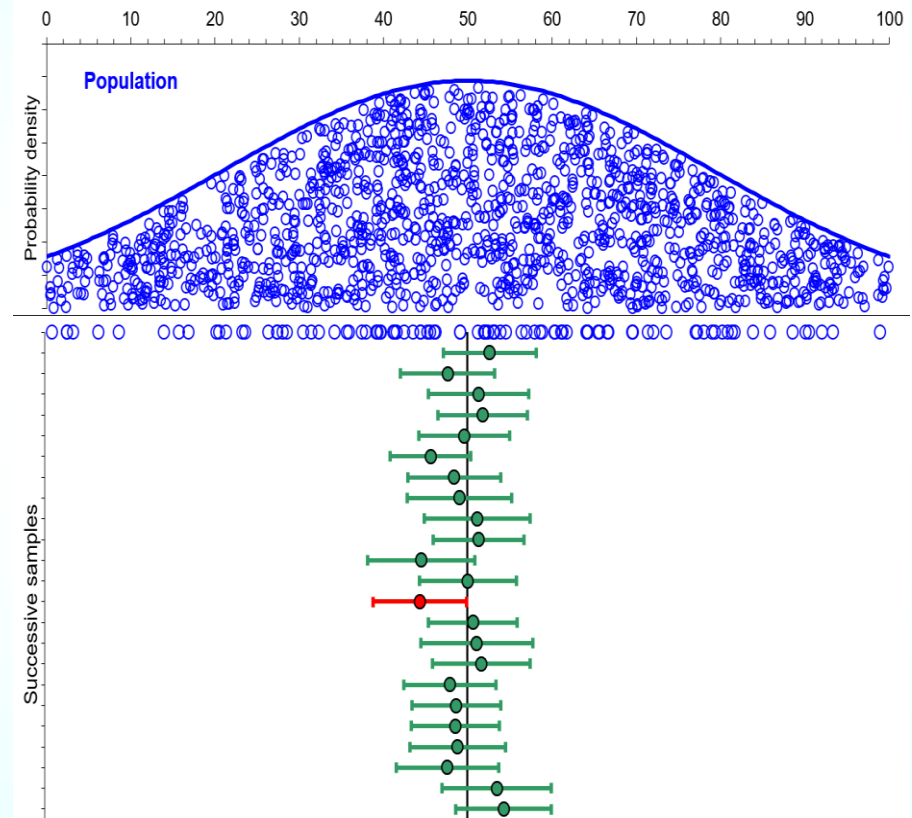
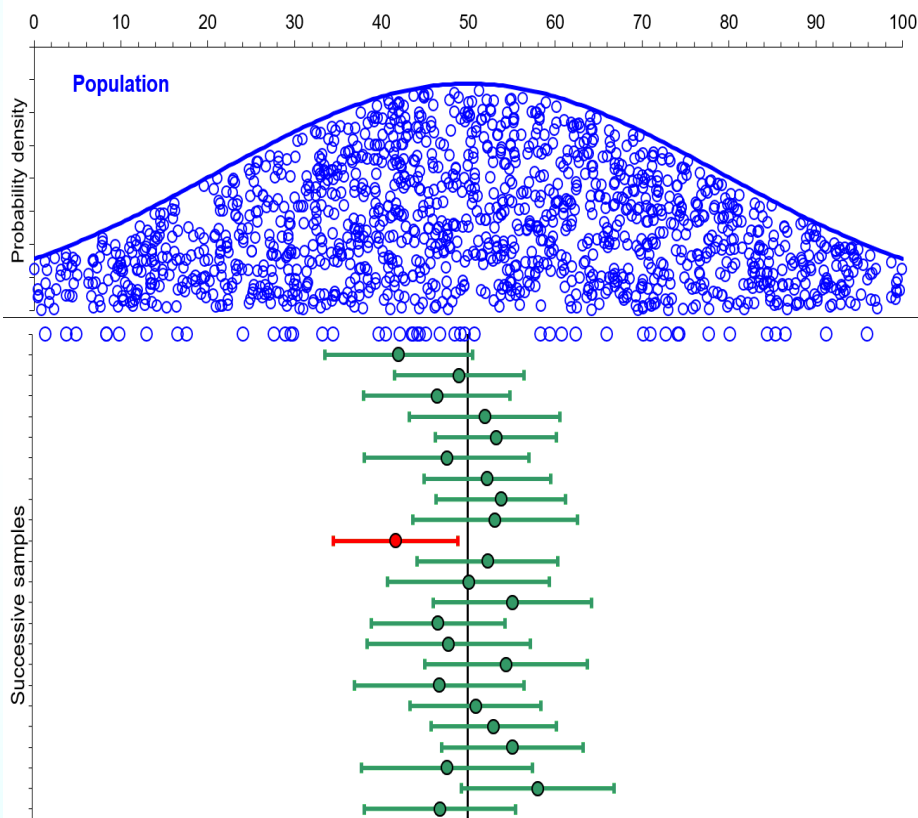


Large variability = large SE

Error and variability

N=50

N=100



Large variability = large SE

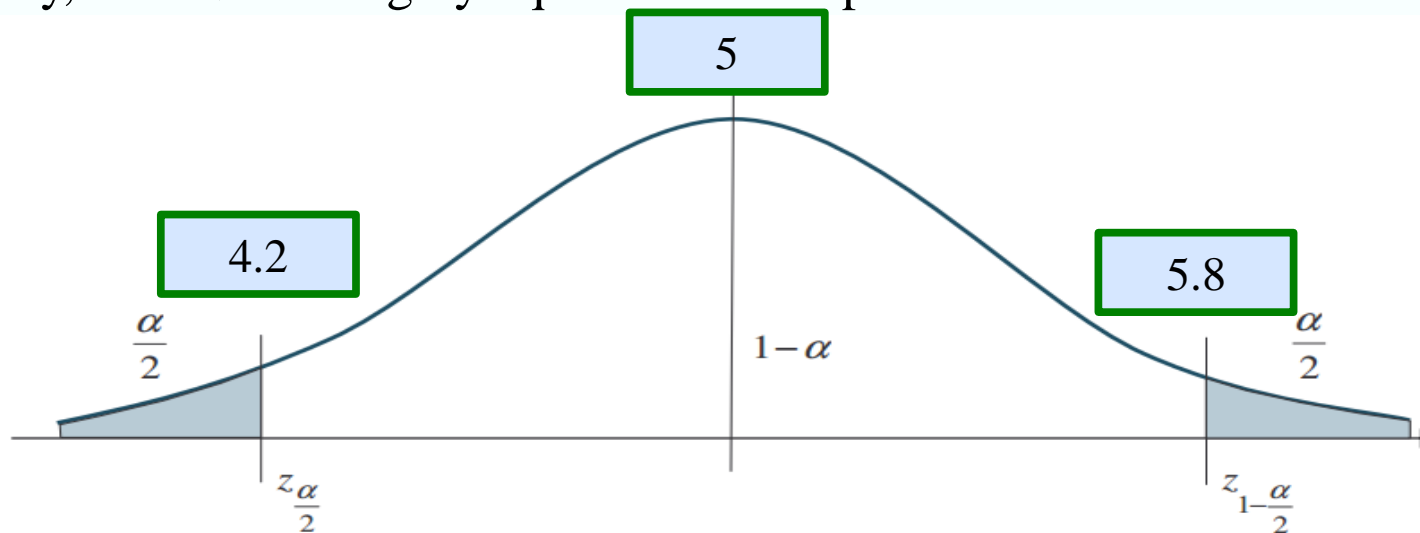
Small variability = small SE

Error and variability

- **Remember**: basically you have almost always results from samples and not from populations
- There is an error in inferring results from samples as if they apply to a population
- Greater variability means more errors

From SE to Confidence Interval (CI)

The sample estimate does not correspond to the population value.
 Confidence Interval provides a range of values that contain the population value with a certain likelihood (e.g., 95%), should the study be repeated many times
 To simplify, CI 95% is roughly equal to the sample mean +/- 2 SE



For example: $M = 5$; $DS = 4$ $N=100$

$$SE = \sqrt{\frac{4^2}{100}} = \frac{4}{\sqrt{100}} = 0.4$$

$$\text{Range: } 2 \times SE = 0.8$$

$$95\% \text{ CI} = [4.2, 5.8]$$

$$\mu \in \left(\bar{X}_n \pm t_{1-\frac{\alpha}{2}}^{(n-1)} \sqrt{\frac{s_n^2}{n}} \right)$$



Standard Error

The Confidence Interval (CI)

The CI reflects the concept of **accuracy** in estimating a parameter

Imagine this research scenario. We want to understand the efficacy of 2 ads for a product (e.g., snack). $N=100$

We computed the mean evaluation of the two ads

- A) $M = +3.10$; $DS = 15$, $p < .05$
- B) $M = +2.50$; $DS = 10$, $p < .05$

Which is the best ad? It is not obvious that it is A

- A) 95% CI= [0.16, 6.04]
- B) 95% CI= [0.54, 4.46]

A can be 6.04, but it can also be 0.16.

B is more accurate, so its possible values are less spread: it is very unlikely that its mean is lower than 0.54

Also correlations have confidence intervals

- Confidence intervals can be calculated for many statistical parameters
- CI for correlations (r) are bounded (-1, 1) and often asymmetrical

r=0, n=300

r=0, n=30

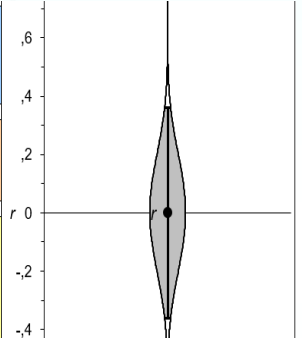
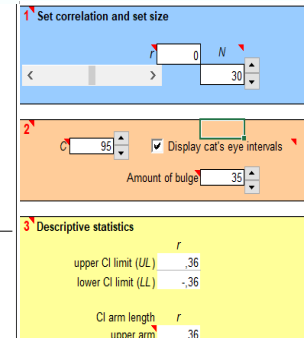
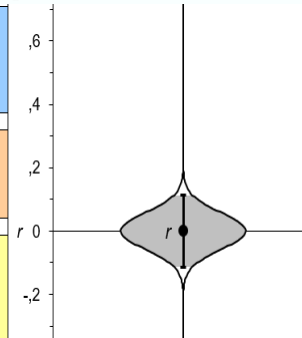
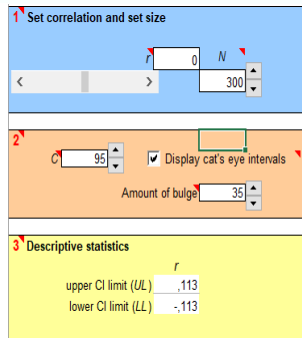
Define Fisher Transformation: $Z_r = \frac{\ln\left(\frac{1+r}{1-r}\right)}{2}$

Define: $L = Z_r - \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{N-3}}$

$U = Z_r + \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{N-3}}$

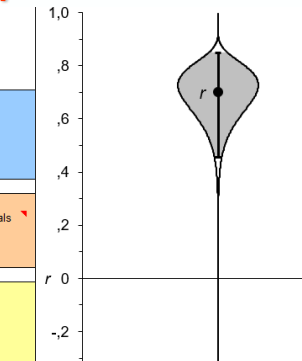
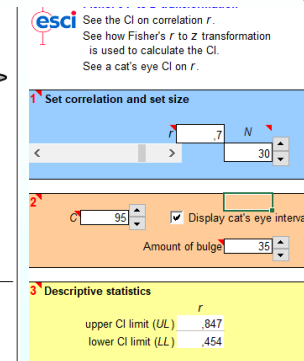
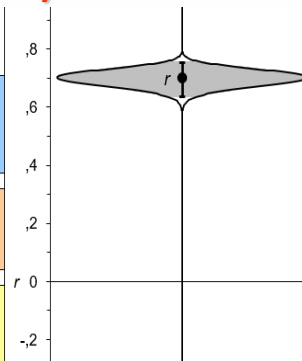
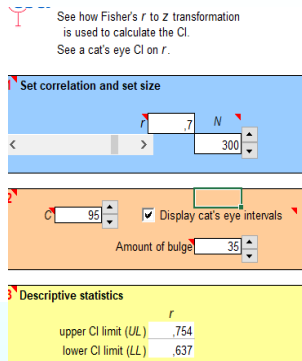
The 100(1-α)% confidence interval is defined as:

$$\left(\frac{e^{2L}-1}{e^{2L}+1}, \frac{e^{2U}-1}{e^{2U}+1} \right)$$



r=0.70, n=300

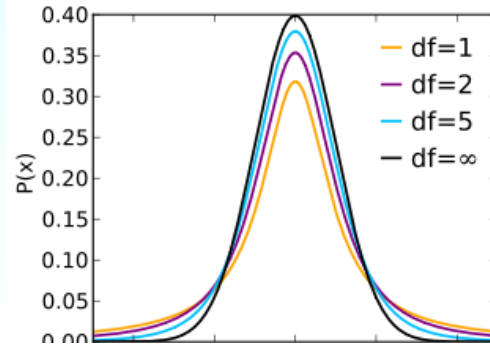
r=0.70, n=30



Hypothesis testing

- When we have data, we can estimate some parameters from them (e.g., mean, correlation)
- We saw that the estimate of this parameter can be more or less **accurate**
- But we can also make inferences from the estimated parameter
- If the parameter is different from a certain value (e.g., 0)
- If the parameter is different comparing certain groups (e.g., experimental vs. control, male vs. female)
- This is the realm of **hypothesis testing** (or statistical inferences from data)
- NHST: **H0** (e.g., parameter = 0) vs. **H1** (e.g., parameter \neq 0)

Fun fact about the t-test!



VOLUME VI MARCH, 1908 No. 1

BIOMETRIKA.

THE PROBABLE ERROR OF A MEAN.

By STUDENT.

Introduction.

ANY experiment may be regarded as forming an individual of a "population" of experiments which might be performed under the same conditions. A series of experiments is a sample drawn from this population.

Now any series of experiments is only of value in so far as it enables us to form a judgment as to the statistical constants of the population to which the experiments belong. In a great number of cases the question finally turns on the value



William Sealy Gosset



But literally H_0 is never true...

- Given an infinite sample size, two parameters (e.g., means) will always be significantly different unless they are exactly identical, or one parameter will always be different from zero unless it is exactly zero (cf. **standard error**)

$r = .01$ with $N=40000$ is significantly different from 0 with $p < .05$ ($p = .0456$)

- It is thus important to understand the **effect size** (even if significant, some effects can be of a trivial quantity)
- Different effect size estimators
- Most common: **Cohen's d** and **Pearson's r** (correlation coefficient)

$$\text{Cohen's } d = \frac{M_1 - M_2}{SD_{\text{pooled}}}$$

$$SD_{\text{pooled}} = \sqrt{\frac{(SD_1^2 + SD_2^2)}{2}}$$

$$r = \sqrt{\frac{\chi^2(1)}{N}}$$

$$r = \sqrt{\frac{t^2}{t^2 + df}}$$

$$r = \sqrt{\frac{F(1,-)}{F(1,-) + df_R}}$$

$$d = \frac{2r}{\sqrt{1-r^2}}$$

Effect size: examples

- A: ad product
- B: control group (irrelevant ad)
- VD: Product evaluation (from 0 to 10)

- A (n=60) : M= 6.50, SD=1.20
- B (n=60) : M= 5.50, SD=1.30
- $SD_{\text{pool}}=1.25$

If A: M= 7.50, DS=1.20; B: M= 5.50, DS=1.30
 $SD_{\text{pool}}=1.25, d = \frac{7.50-5.50}{1.25}=\mathbf{1.60} \quad r = \mathbf{0.62}$

- *Cohen's d* = $\frac{6.50-5.50}{1.25}=\mathbf{0.80} \quad r=\mathbf{0.37}$

If A: M= 6.50, DS=2.20; B: M= 5.50, DS=2.30
 $SD_{\text{pool}}=2.25, d = \frac{6.50-5.50}{2.25}=\mathbf{0.44} \quad r = \mathbf{0.22}$

Rough guidelines (ES should be understood within research context)

$r = .1, d = 0.2$ (small effect): the effect explains 1% of the total variance.

$r = .3, d = 0.5$ (medium effect): the effect explains 9% of the total variance.

$r = .5, d = 0.8$ (large effect): the effect explains 25% of the variance.

Other effect size indexes (from Ellis, 2010)

Table 1.1 *Common effect size indexes*

| Measures of group differences (the <i>d</i> family) | | Measures of association (the <i>r</i> family) | |
|---|---|---|--|
| (a) Groups compared on dichotomous outcomes | | (a) Correlation indexes | |
| RD | The risk difference in probabilities: the difference between the probability of an event or outcome occurring in two groups | <i>r</i> | The Pearson product moment correlation coefficient: used when both variables are measured on an interval or ratio (metric) scale |
| RR | The risk or rate ratio or relative risk: compares the probability of an event or outcome occurring in one group with the probability of it occurring in another | ρ (or r_s) | Spearman's rho or the rank correlation coefficient: used when both variables are measured on an ordinal or ranked (non-metric) scale |
| OR | The odds ratio: compares the odds of an event or outcome occurring in one group with the odds of it occurring in another | τ | Kendall's tau: like rho, used when both variables are measured on an ordinal or ranked scale; tau-b is used for square-shaped tables; tau-c is used for rectangular tables |
| (b) Groups compared on continuous outcomes | | | |
| <i>d</i> | Cohen's <i>d</i> : the uncorrected standardized mean difference between two groups based on the pooled standard deviation | r_{pb} | The point-biserial correlation coefficient: used when one variable (the predictor) is measured on a binary scale and the other variable is continuous |
| Δ | Glass's delta (or <i>d</i>): the uncorrected standardized mean difference between two groups based on the standard deviation of the control group | ϕ | The phi coefficient: used when variables and effects can be arranged in a 2x2 contingency table |
| <i>g</i> | Hedges' <i>g</i> : the corrected standardized mean difference between two groups based on the pooled, weighted standard deviation | <i>C</i> | Pearson's contingency coefficient: used when variables and effects can be arranged in a contingency table of any size |
| PS | Probability of superiority: the probability that a random value from one group will be greater than a random value drawn from another | <i>V</i> | Cramér's V: like C, V is an adjusted version of phi that can be used for tables of any size |
| | | λ | Goodman and Kruskal's lambda: used when both variables are measured on nominal (or categorical) scales |

(cont.)

Table 1.1 (cont.)

| Measures of group differences (the <i>d</i> family) | | Measures of association (the <i>r</i> family) | |
|---|--|---|--|
| | | (b) Proportion of variance indexes | |
| | | r^2 | The coefficient of determination: used in bivariate regression analysis |
| | | R^2 | R squared, or the (uncorrected) coefficient of multiple determination: commonly used in multiple regression analysis |
| | | $adjR^2$ | Adjusted R squared, or the coefficient of multiple determination adjusted for sample size and the number of predictor variables |
| | | <i>f</i> | Cohen's <i>f</i> : quantifies the dispersion of means in three or more groups; commonly used in ANOVA |
| | | f^2 | Cohen's <i>f</i> squared: an alternative to R^2 in multiple regression analysis and ΔR^2 in hierarchical regression analysis |
| | | η^2 | Eta squared or the (uncorrected) correlation ratio: commonly used in ANOVA |
| | | ε^2 | Epsilon squared: an unbiased alternative to η^2 |
| | | ω^2 | Omega squared: an unbiased alternative to η^2 |
| | | R^2_c | The squared canonical correlation coefficient: used for canonical correlation analysis |

General logic behind ES

$$\hat{\eta}^2 = \frac{SS_{\text{Effect}}}{SS_T},$$

$$\hat{\eta}_P^2 = \frac{SS_{\text{Effect}}}{SS_{\text{Effect}} + SS_{s/\text{Cells}}},$$

$$\hat{\omega}_P^2 = \frac{SS_{\text{Effect}} - df_{\text{Effect}} MS_{s/\text{Cells}}}{SS_{\text{Effect}} + (N - df_{\text{Effect}}) MS_{s/\text{Cells}}}$$

$$d = \frac{M_1 - M_2}{\text{pooled SD}}$$

$$r = \frac{\text{Covariance}(x,y)}{S.D.(x)S.D.(y)}$$

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

➤ Effect sizes go up when “**signal**” (*numerator*) increases relative to “**noise**” (*denominator*)

Effect size: useful tools

Read this: <https://doi.org/10.3389/fpsyg.2013.00863>

(Lakens, 2013) **frontiers in
PSYCHOLOGY**

REVIEW ARTICLE
published: 26 November 2013
doi: 10.3389/fpsyg.2013.00863



Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for *t*-tests and ANOVAs

Daniël Lakens*

Human Technology Interaction Group, Eindhoven University of Technology, Eindhoven, Netherlands

Use this: https://www.psychometrica.de/effect_size.html

(give a look also here <http://www.stat-help.com/spreadsheets.html>)

Check (or ask) your analysis output (SPSS, R) for effect sizes

Effect size can be calculated starting from different bits of information and can be transformed (e.g., from *r* to *d*)

Other readings and tools

Some bibliographic references:

- Fritz, C.O., Morris, P.E., & Richler, J.J. (2012). Effect size estimates: Current use, calculations, and interpretation. *Journal of Experimental Psychology:General*, 141, 2–18.
- Ellis (2010). *The essential guide to effect sizes*. Cambridge University Press.
- Cohen (1992). A power primer. *Psychological Bulletin*, 112, 155-159.
- Cohen (1994). The earth is round ($p < .05$). *American Psychologist*, 49, 997-1003.
- Cohen (1988). *Statistical power analysis for the behavioral sciences*. LEA

Some online calculators

- https://www.psychometrica.de/effect_size.html
- <https://sites.google.com/site/lakens2/effect-sizes>
- <https://www.campbellcollaboration.org/this-is-a-web-based-effect-size-calculator/explore/this-is-a-web-based-effect-size-calculator>
- <http://www.stat-help.com/spreadsheets.html>

Errors of inference

- Frequentist approach
- There are three types of errors
- NHST*: Type I error (False positives)
Type II error (False negatives)
- CI: Estimate error (imprecision)

NHST= Null Hypothesis Significance Testing
(H0 vs. H1)

Errors of inference in NHST

Decision outcomes from NHST (one parameter against a value or one parameter in two (or more) groups)

Conclusion of the significance test (SAMPLE)

Null is true
Null is false

Real World (POPULATION)
Null is true (H0 is correct) Null is false (H1 is correct)

| | |
|---|--|
| <p>Correct decision ($1-\alpha$)</p> | <p>Type II error (β)</p> |
| <p>Type I error (α)</p> | <p>Correct decision ($1-\beta$)</p> |

We analyze results in a sample but make an inference to population. We can make errors of inference

Errors of inference in NHST

- **Type I error:** *Erroneously rejecting the null hypothesis (False positive)*.

The result in the sample is significant ($p < .05$), so the null hypothesis is rejected, but the null hypothesis is actually true in the population.

- **Type II error:** *Erroneously accepting the null hypothesis (False negative)*. The result in the sample is not significant ($p > .05$), so the null hypothesis is not rejected, but it is actually false in the population.
-

How to control Type I errors?

- The Type I error rate (*False positive*) is controlled by the researcher.
- It is called the **alpha rate** and corresponds to the probability cut-off (p) that one uses in a significance test.
- Conventionally, researchers use an alpha rate (α) of .05. This means that the null hypothesis is rejected when a value such as the one found is likely to occur 5% of the time or less when the null hypothesis is true.
- The test can be two-tailed (more common) or one-tailed (directional)

One-tailed and two-tailed test

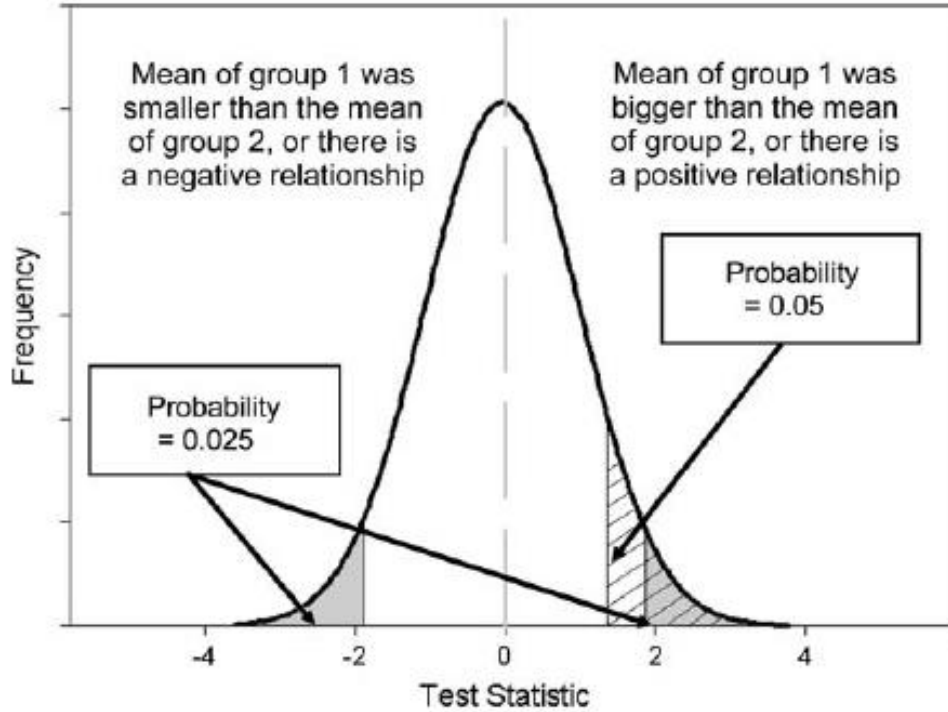


FIGURE 2.10
Diagram to show the difference between one- and two-tailed tests

How to control Type II errors?

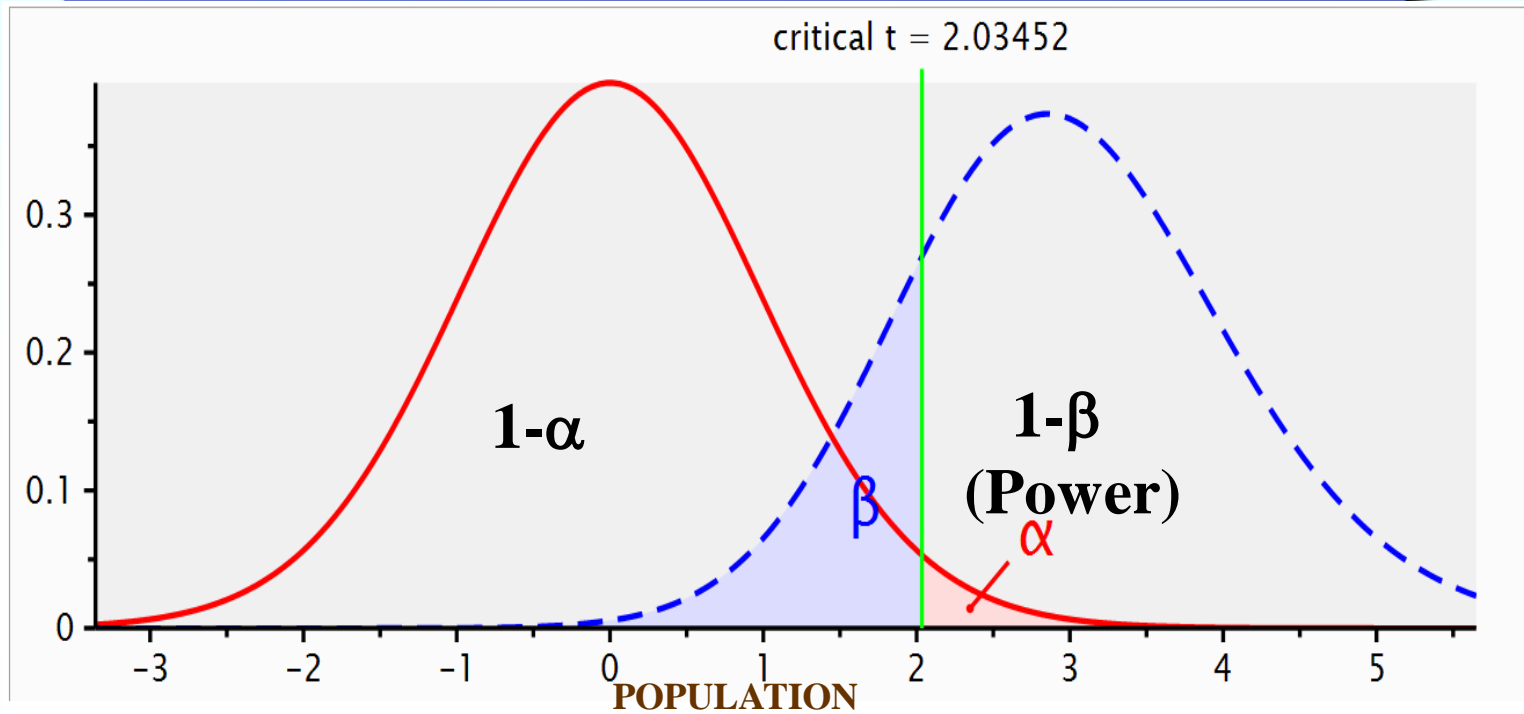
- The Type II error (*False negative*) can also be controlled by the experimenter.
- The Type II error rate is called **beta** (β) as a complement to alpha.
- How can the beta rate be controlled? The easiest way to control Type II errors is by increase the **statistical power** of a test.
- **Statistical power**= probability of finding an effect, if it exists
- **Power** = $1 - \beta$
- Conventionally a power of at least .80 ($\beta=.20$) is considered as acceptable

Power analysis

Power analysis

- Power analysis is a basic tool for planning studies
 - You already know it
 - We will quickly refresh the basic concepts and then articulate three specific issues linked to power analysis:
 - a) uncertainty of the estimates
 - b) sensitivity
 - b) within vs. between design
-

What is power?



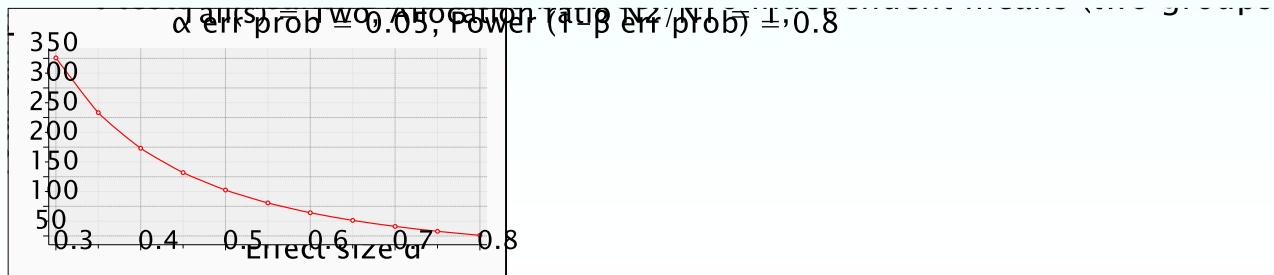
| | | Null is true (H0 is correct) | Null is false (H1 is correct) |
|--------|---------------|------------------------------------|-----------------------------------|
| SAMPLE | Null is true | Correct decision ($1-\alpha$) | Type II error (β) |
| | Null is false | Type I error (α) | Correct decision ($1-\beta$) |

The key determinants of power

- Power is determined by four elements
 - 1) Decision criterion (α)
 - 2) Sample size (n)
 - 3) Effect size (δ)
 - 4) Desired power ($1 - \beta$)
 - Fixing one of the elements one can derive the others
-

A simple example

- Fix $\alpha=.05$ and $(1-\beta)=.80$
- Plot sample size and effect size for a two sample t-test



What affects power?

- Power goes up with larger effect sizes and sample sizes, given a certain decision criterion (e.g., $\alpha=.05$)
- When effect sizes become larger? When the portion of variability (difference) ascribed to the effect of interest grows more than the general (non specific) variability

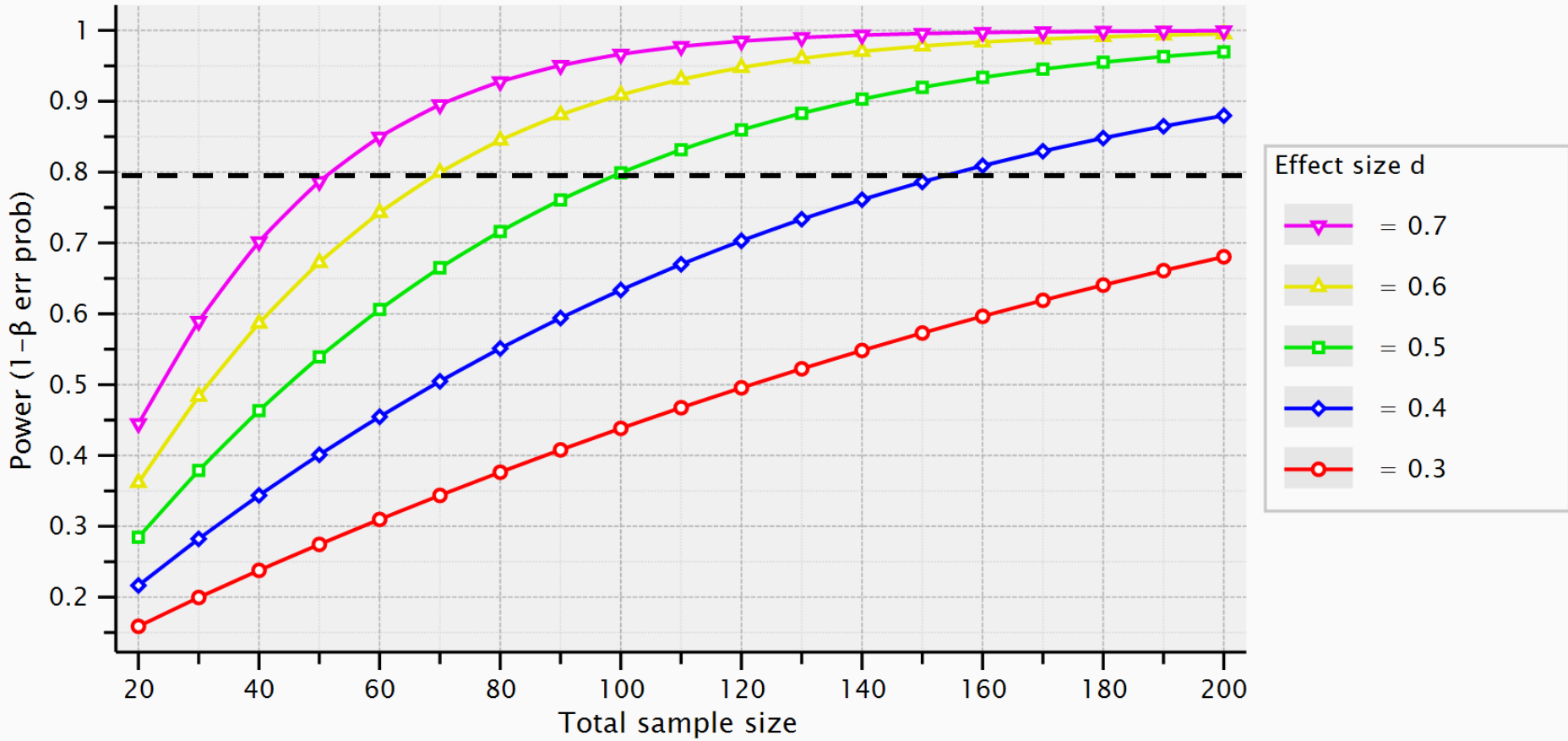
$$d = \frac{M_1 - M_2}{\text{pooled } SD}$$

$$\eta^2 = \frac{SS_{\text{Effect}}}{SS_T}$$

$$r(v, x) = \frac{\text{cov}(v, x)}{sd(v) * sd(x)}$$

Power as a function of ES and N

t tests – Means: Difference between two independent means (two groups)
Tail(s) = One, α err prob = 0.05, Allocation ratio $N_2/N_1 = 1$



How to increase power?

Power is affected by

- *Sample size*



- *Construct-related (i.e., SIGNAL) variance*



- *Construct-unrelated (i.e., NOISE) variance*



What is affected by power?

Higher power means

- *Less False Negatives*
- *Lower overall errors of inference (crucial error rates)*

Lower power means

- *with multiple outcomes and HARKing: body of conflicting evidence in the literature*
 - *with publication bias: presence of many false-positives in the literature*
-

Why power analysis to plan studies?

- Without logistical constraints (infinite resources and no costs), only accuracy in estimating parameters should matter (e.g., AIPE, Maxwell et al, 2008)
 - In an accuracy (precision) approach, one thing matters a lot: sample size, the bigger, the better (*ceteris paribus*)
 - The point is not whether some effect exists (or not) but how precise is our estimate of it
 - All effects exist given an infinite sample size (Cohen)
 - Increased accuracy means less inference errors (both Type I and Type II)
 - ***If you want to get it right, increase sample size***
-

Precision vs. Power

- They have different aims

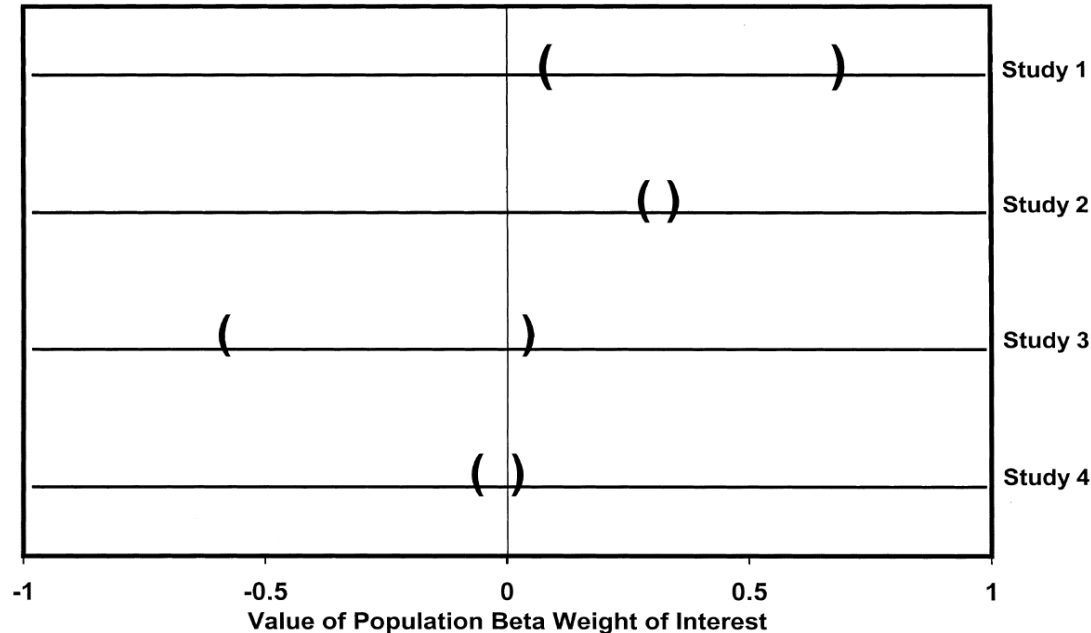
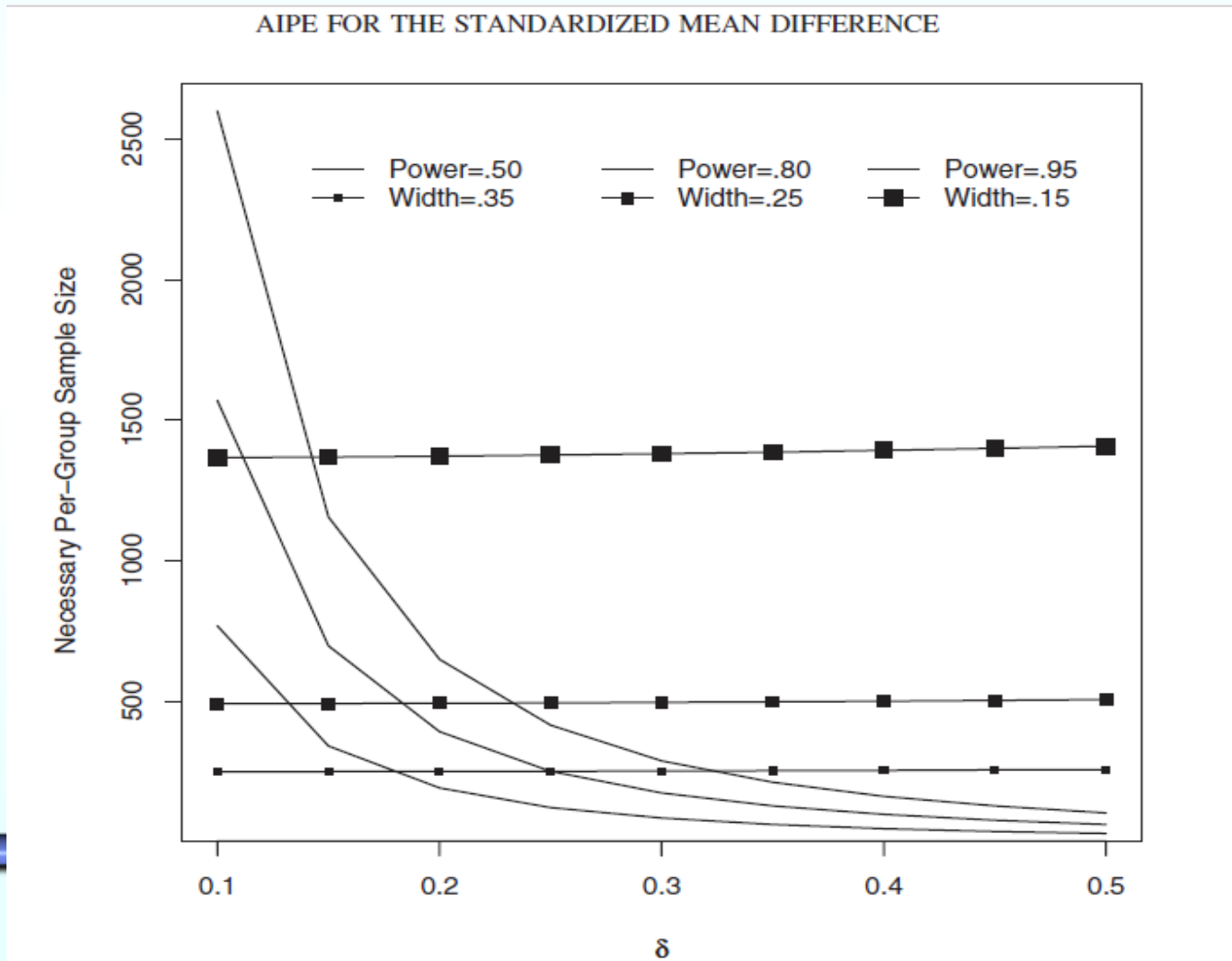


Figure 1. Illustration of possible scenarios in which planned sample size was considered a “success” or “failure” according to the accuracy in parameter estimation and the power analysis frameworks. Parentheses are used to indicate the width of the confidence interval.

- Precision is valuable no matter everything else

a MINOR practical problem...

- **Big** sample sizes are needed for precise estimates no matter the effect size



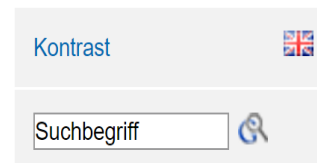
How to calculate power

- Different software and routines (e.g., in R)
- A free comprehensive package is G*Power

<http://www.gpower.hhu.de/>

G*Power: Statistical Power Analyses for Windows and Mac

G*Power is a tool to compute statistical power analyses for many different t tests, F tests, χ^2 tests, z tests and some exact tests. G*Power can also be used to compute effect sizes and to display graphically the results of power analyses.



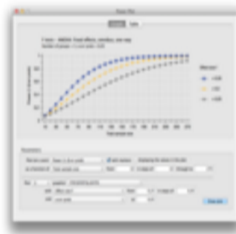
Screenshots (click to enlarge)



Main Window

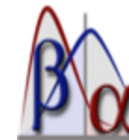


Main Window (Table)



Power Plot

Power Plot (Table)



Problems in power analysis

- **One main error:** post-hoc power (calculated after the results) is trivial and misleading. Sensitivity analysis is better

 INTERNATIONAL REVIEW
OF SOCIAL PSYCHOLOGY

Perugini, M., et al. (2018). A Practical Primer To Power Analysis for Simple Experimental Designs. *International Review of Social Psychology*, 31(1): 20, 1-23, DOI: <https://doi.org/10.5334/irsp.181>

- **Three issues:**

a) uncertainty


b) sensitivity

c) within vs. between design

RESEARCH ARTICLE

A Practical Primer To Power Analysis for Simple Experimental Designs

Marco Perugini, Marcello Gallucci and Giulio Costantini

 journal of cognition

Brybaert, M. 2019 How Many Participants Do We Have to Include in Properly Powered Experiments? A Tutorial of Power Analysis with Reference Tables. *Journal of Cognition*, 2(1): 16, pp. 1-38. DOI: <https://doi.org/10.5334/joc.72>

REVIEW ARTICLE

How Many Participants Do We Have to Include in Properly Powered Experiments? A Tutorial of Power Analysis with Reference Tables

Marc Brybaert

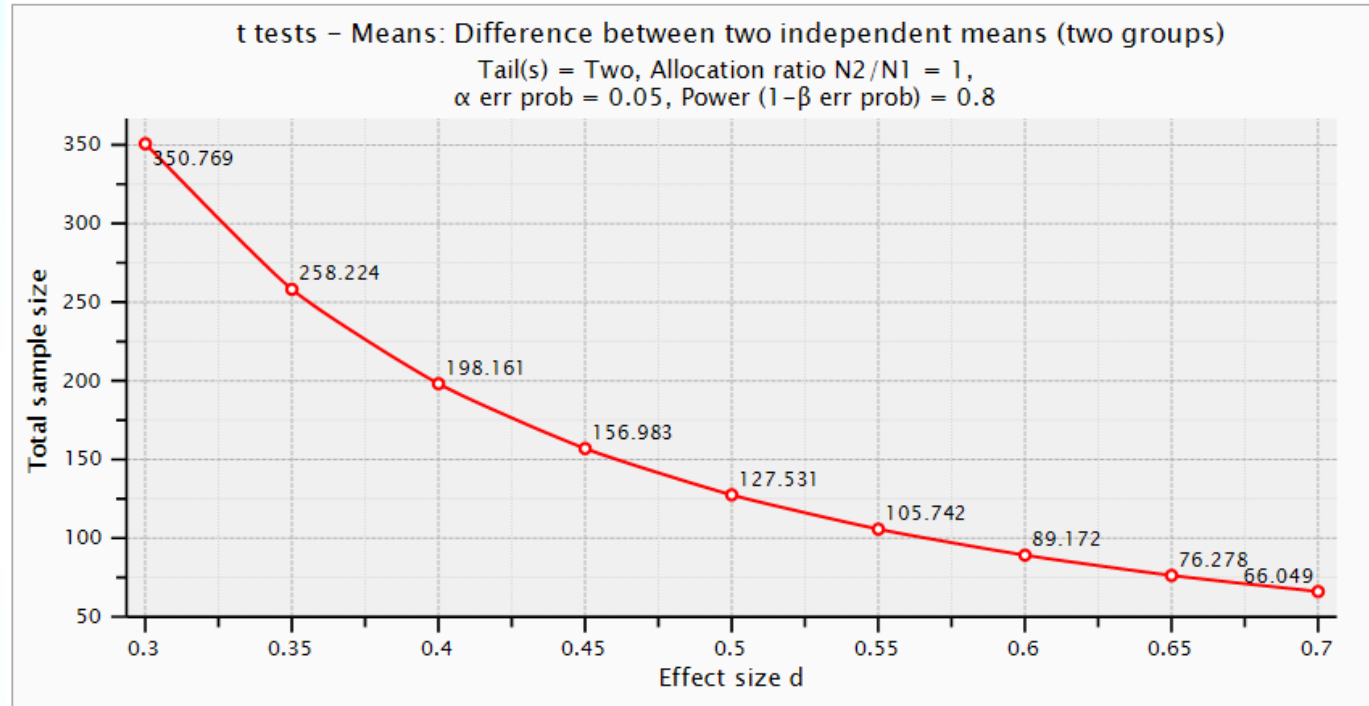
Department of Experimental Psychology, Ghent University, BE
marc.brybaert@ugent.be

a) Uncertainty

- One key element of power analysis for planning studies is the Effect Size (ES)
 - We can use only an **estimate** of ES (sample) but need the **unknown** expected ES (population). If we knew it, we wouldn't need to run the study...
 - At best we guess it from a meta-analysis (or previous studies), at worst based on a hunch or even arbitrarily set. **Uncertainty** of the estimate
 - What happens if the ES estimate is incorrect?
-

Uncertainty of ES

Graph Table



Plot Parameters

Plot (on y axis) with markers and displaying the values in the plot Show digits

as a function of from in steps of through to

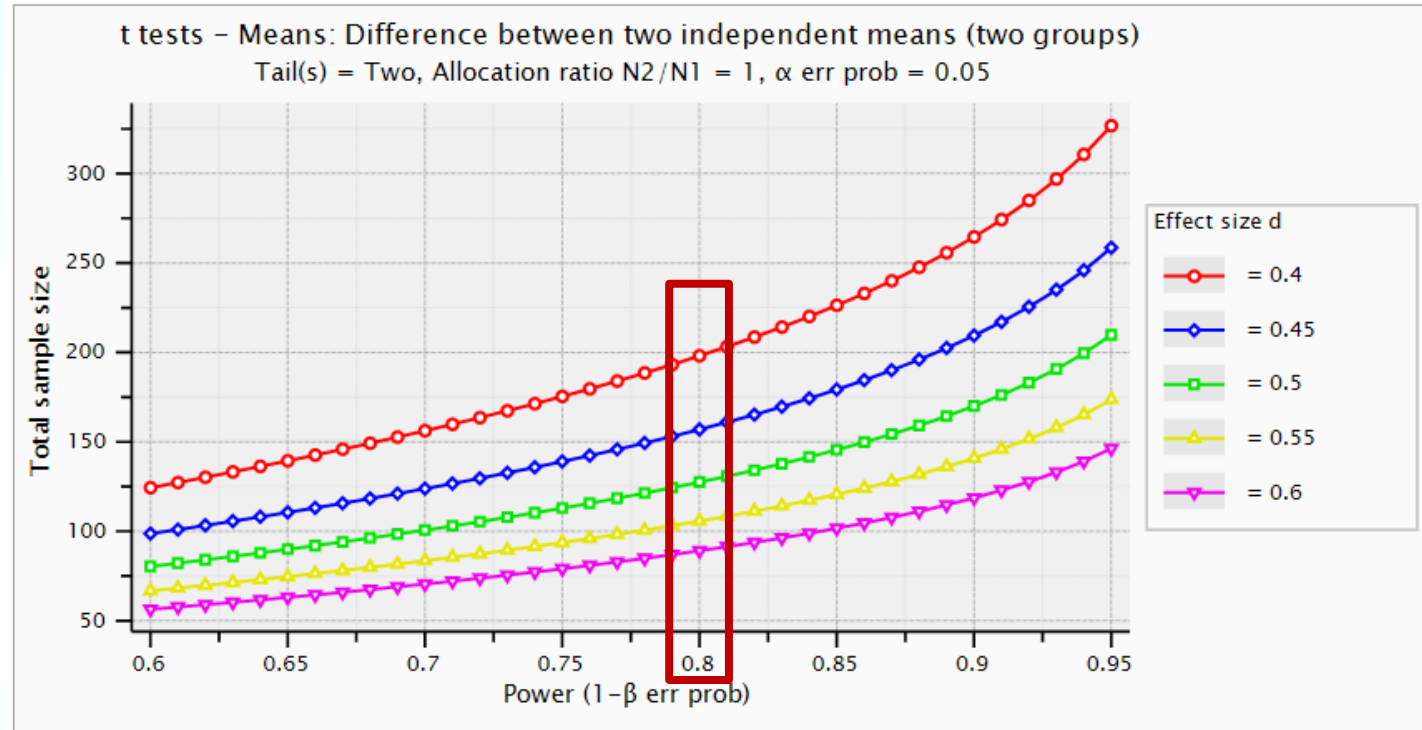
Plot graph(s)

with at

and at

Uncertainty of ES

Graph Table



Plot Parameters

Plot (on y axis) Total sample size with markers and displaying the values in the plot

as a function of Power (1-β err prob) from 0.6 in steps of 0.01 through to 0.95

Plot 5 graph(s) interpolating points

with Effect size d from 0.4 in steps of 0.05

and α err prob at 0.05

Draw plot

Uncertainty of ES

Graph Table

t tests - Means: Difference between two independent means (two groups)
Tail(s) = Two, Allocation ratio N2/N1 = 1, α err prob = 0.05

| # | Power (1- β err prob) | Effect size d = 0.4 Total sample size | Effect size d = 0.45 Total sample size | Effect size d = 0.5 Total sample size | Effect size d = 0.55 Total sample size | Effect size d = 0.6 Total sample size |
|----|-----------------------------|--|---|--|---|--|
| 16 | 0.750000 | 175.449 | 139.039 | 112.997 | 93.7315 | 79.0806 |
| 17 | 0.760000 | 179.664 | 142.369 | 115.695 | 95.9607 | 80.9535 |
| 18 | 0.770000 | 184.029 | 145.818 | 118.488 | 98.2689 | 82.8927 |
| 19 | 0.780000 | 188.556 | 149.395 | 121.385 | 100.663 | 84.9042 |
| 20 | 0.790000 | 193.261 | 153.112 | 124.396 | 103.151 | 86.9946 |
| 21 | 0.800000 | 198.161 | 156.983 | 127.531 | 105.742 | 89.1716 |
| 22 | 0.810000 | 203.275 | 161.024 | 130.804 | 108.446 | 91.4438 |
| 23 | 0.820000 | 208.626 | 165.252 | 134.228 | 111.276 | 93.8216 |
| 24 | 0.830000 | 214.241 | 169.689 | 137.822 | 114.246 | 96.3167 |
| 25 | 0.840000 | 220.153 | 174.359 | 141.605 | 117.372 | 98.9433 |
| 26 | 0.850000 | 226.307 | 179.203 | 145.601 | 120.675 | 101.718 |

Plot Parameters

Plot (on y axis) Total sample size with markers and displaying the values in the plot

as a function of Power (1- β err prob) from 0.6 in steps of 0.01 through to 0.95

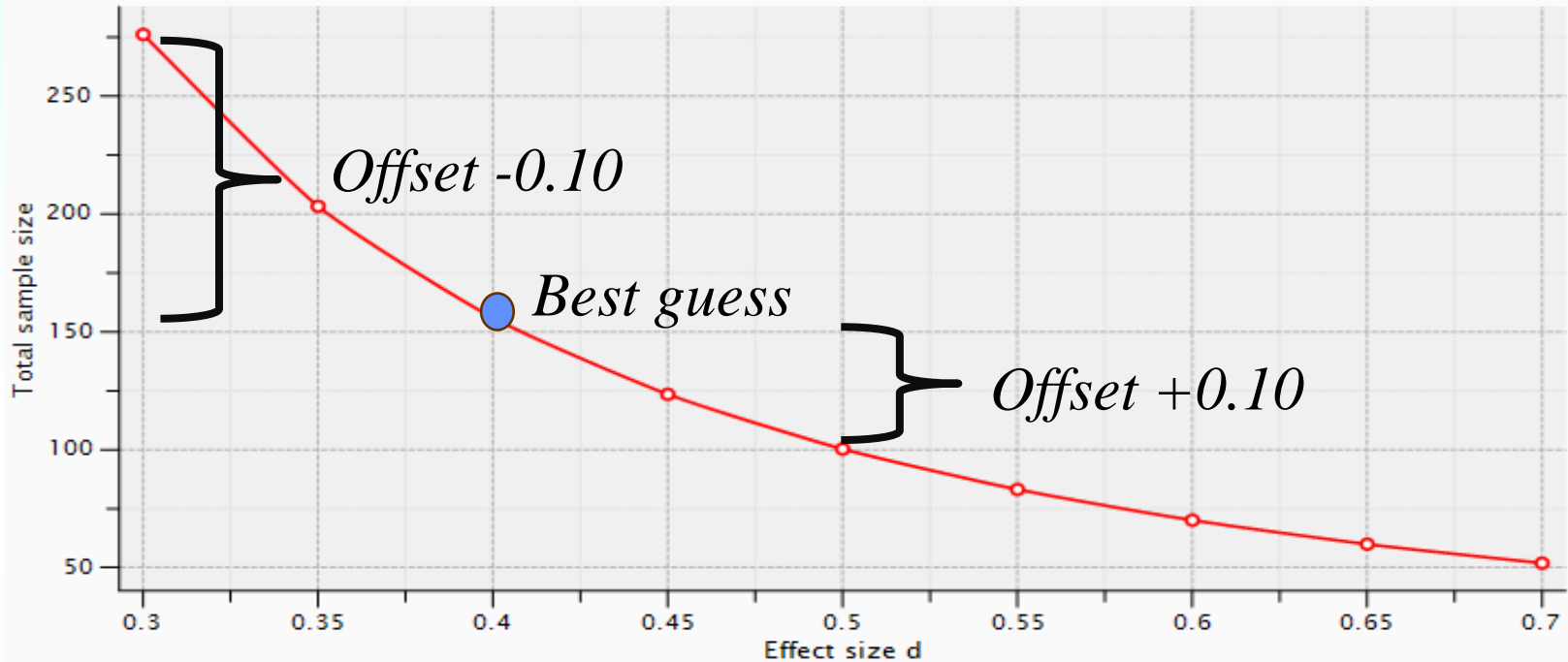
Plot 5 graph(s) interpolating points

with Effect size d from 0.4 in steps of 0.05

and α err prob at 0.05 Draw plot

Asymmetry of ES errors

t tests - Means: Difference between two independent means (two groups)
Tail(s) = One, Allocation ratio $N2/N1 = 1$,
 α err prob = 0.05, Power ($1-\beta$ err prob) = 0.8



Plot Parameters

Plot (on y axis) with markers and displaying the values in the plot

is a function of from in steps of through to

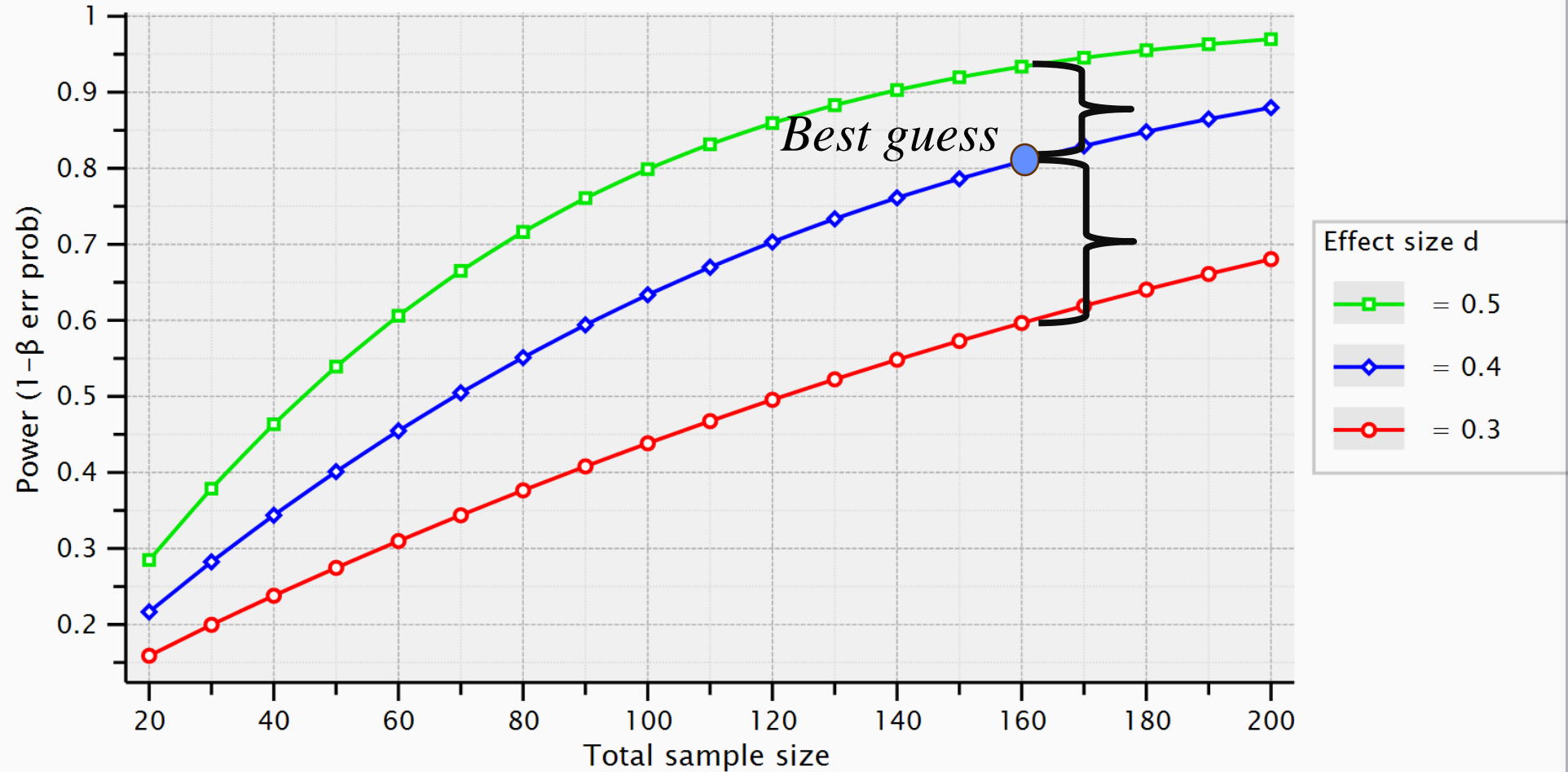
Plot graph(s)

with at

and at

Asymmetry of ES errors

t tests – Means: Difference between two independent means (two groups)
Tail(s) = One, α err prob = 0.05, Allocation ratio $N_2/N_1 = 1$



What to do then?

- Power depends on estimated ES (we don't know the “true” ES)
- ES over-estimation is more common (*optimistic bias*) and more influential than under-estimation (*asymmetric effect*)
- Should consider different scenarios rather than a single value
- Could consider minimum effect of interest (SESOI, Lakens, 2014)
- Could consider sensitivity analysis
- Could consider safeguarding yourself against “optimistic” ES estimates

Equivalence Testing for Psychological Research: A Tutorial



Daniël Lakens , Anne M. Scheel , and Peder M. Isager
Human-Technology Interaction Group, Eindhoven University of Technology

Advances in Methods and Practices in Psychological Science
2018, Vol. 1(2) 259–269
© The Author(s) 2018

Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/2515245918770963
www.psychologicalscience.org/AMPPS

Safeguard Power as a Protection Against Imprecise Power Estimates

Marco Perugini, Marcello Gallucci, and Giulio Costantini
University of Milan-Bicocca, Italy



Perspectives on Psychological Science
2014, Vol. 9(3) 319–352
© The Author(s) 2014
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/1745691614528519
pps.sagepub.com

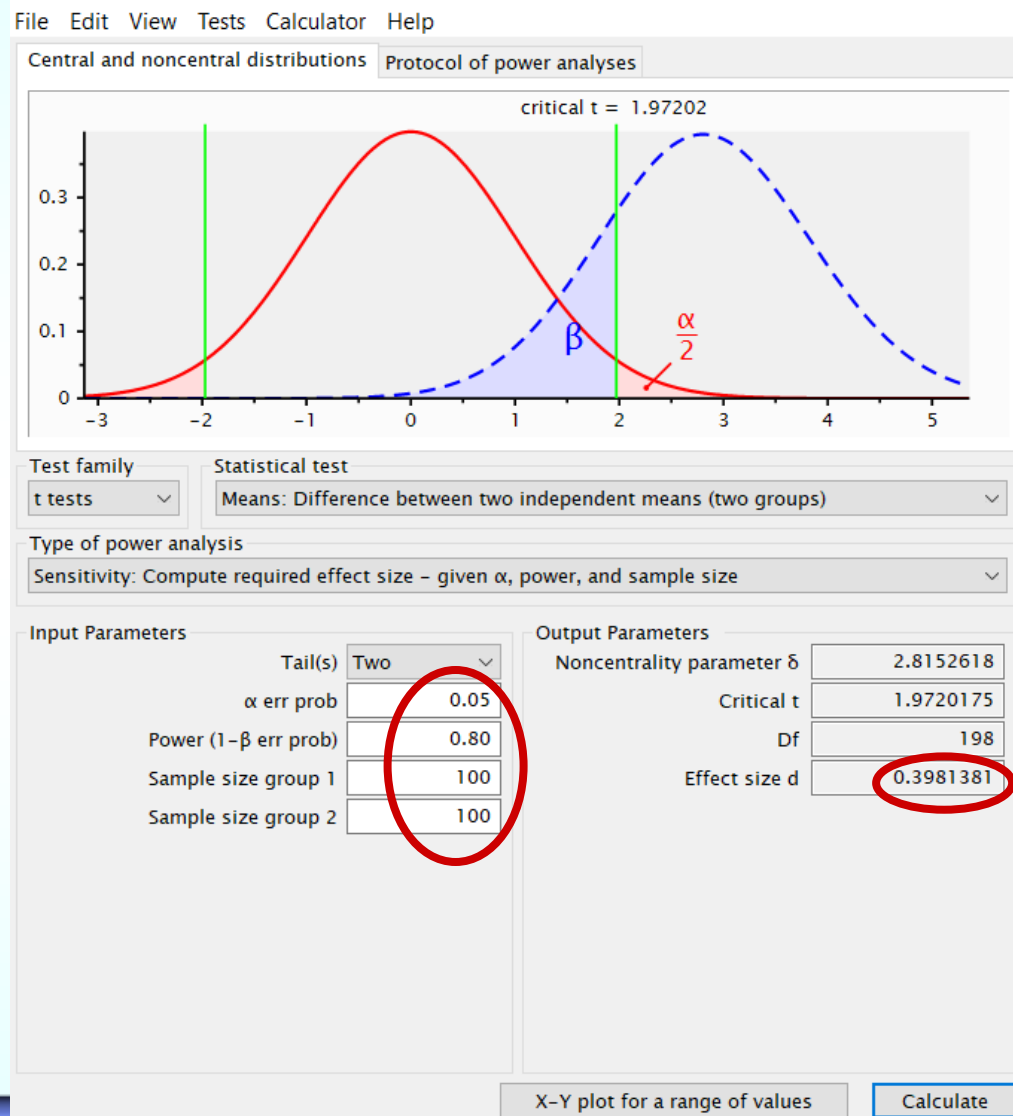
b) Sensitivity: Starting from N

“Sometimes”
resources are fixed

You know that you can
collect a certain N

The question becomes
what ES can be found
with sufficient power

Sensitivity analysis



Sensitivity plot: N by ES

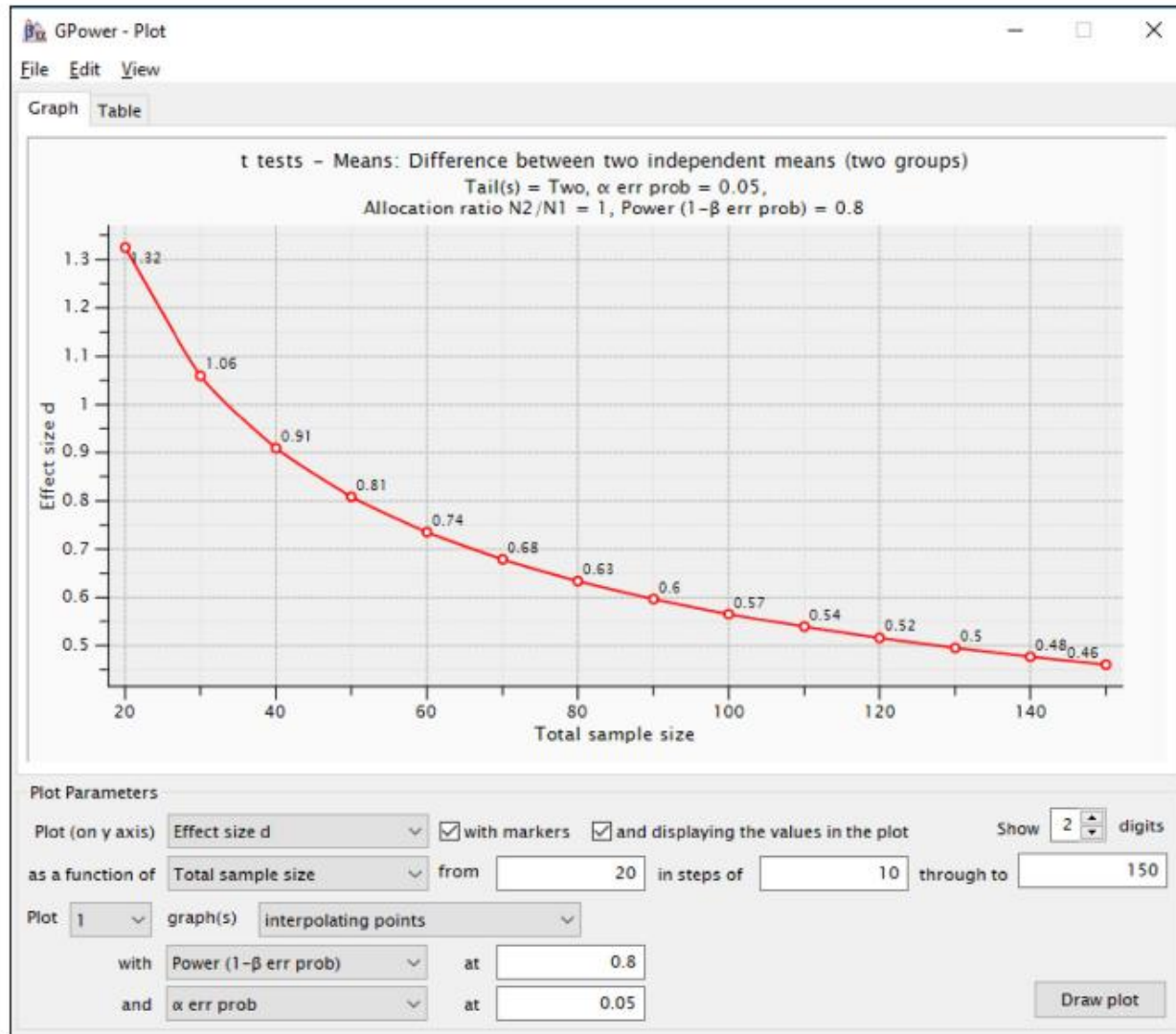


Figure 2: Sensitivity Plot of G*Power calculating the power of a two independent samples t-test: Lowest detectable effect size as a function of required N.

Sensitivity plot: N by Power

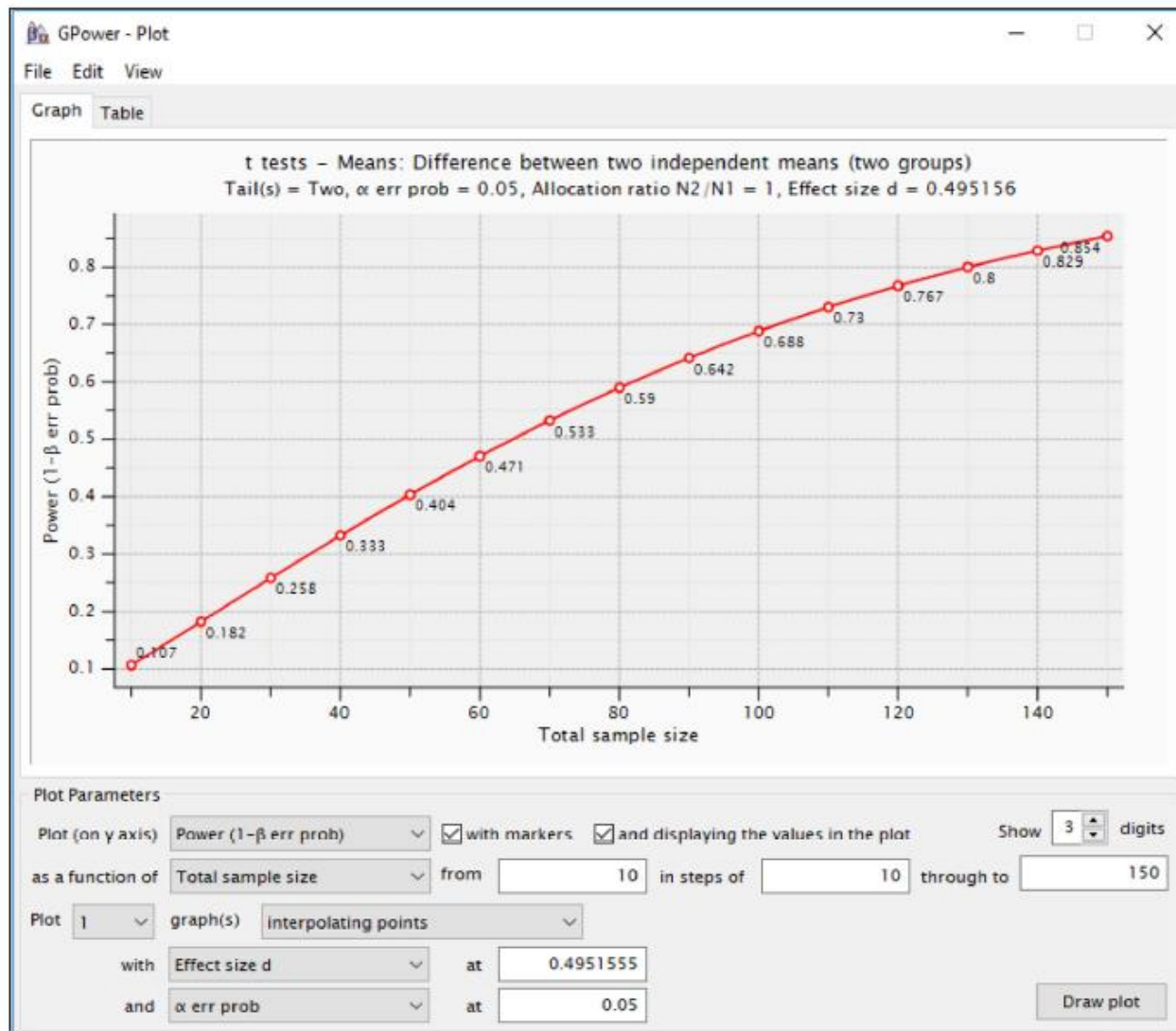
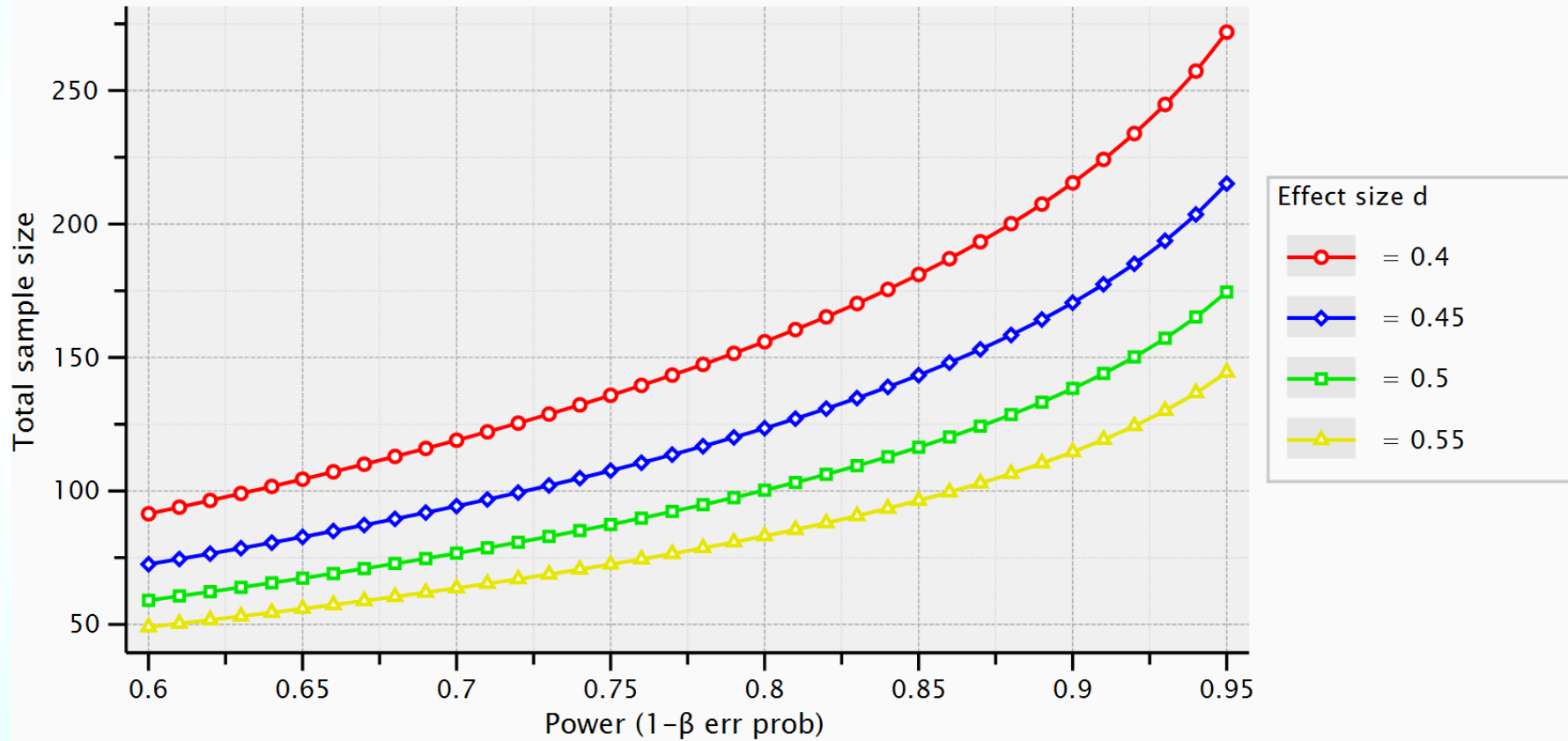


Figure 3: Sensitivity Plot of G*Power calculating the power of a two independent samples t-test: Power as a function of required N for fixed effect size.

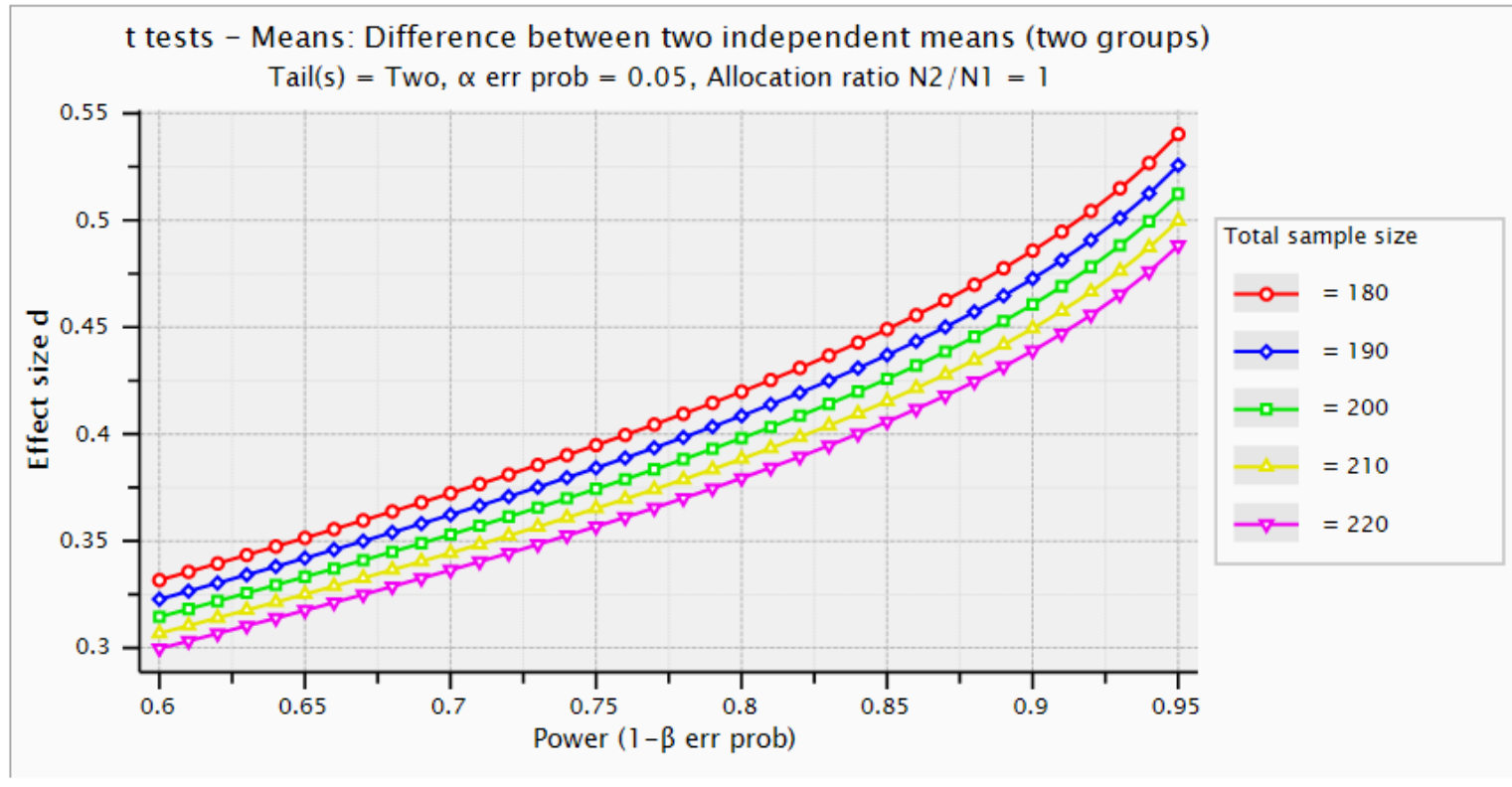
Inspecting scenarios around ES

t tests – Means: Difference between two independent means (two groups)
Tail(s) = One, Allocation ratio $N_2/N_1 = 1$, α err prob = 0.05



Inspecting scenarios around N

Graph Table



Plot Parameters

Plot (on y axis) with markers and displaying the values in the plot

as a function of from in steps of through to

Plot graph(s)

with from in steps of

and at

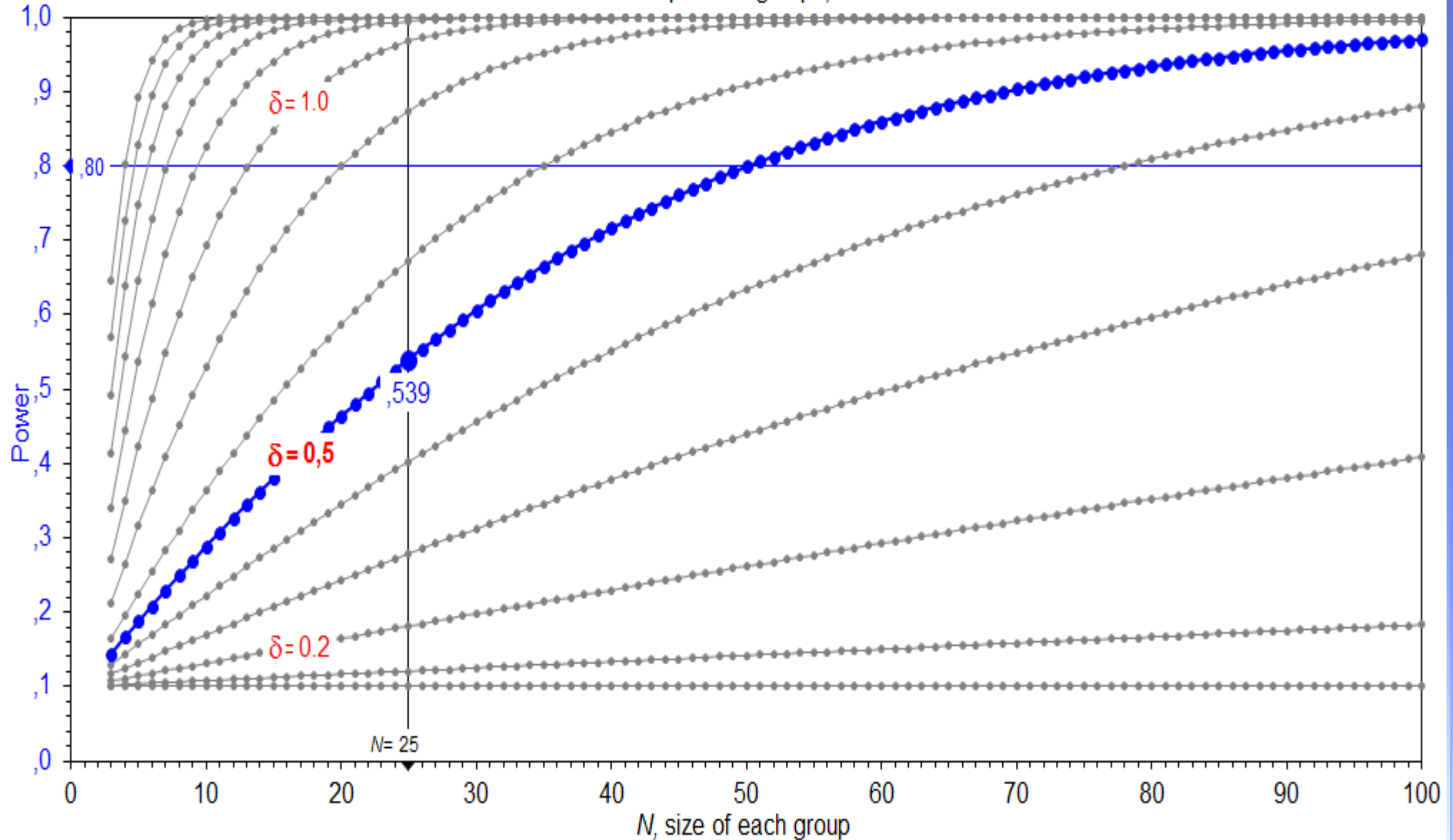
Draw plot

c) Within vs. Between designs

- Everything else being equal, within studies are more powerful than between studies
 - Example with a simple two groups/two measures design
 - Example with 2 x 2 design
-

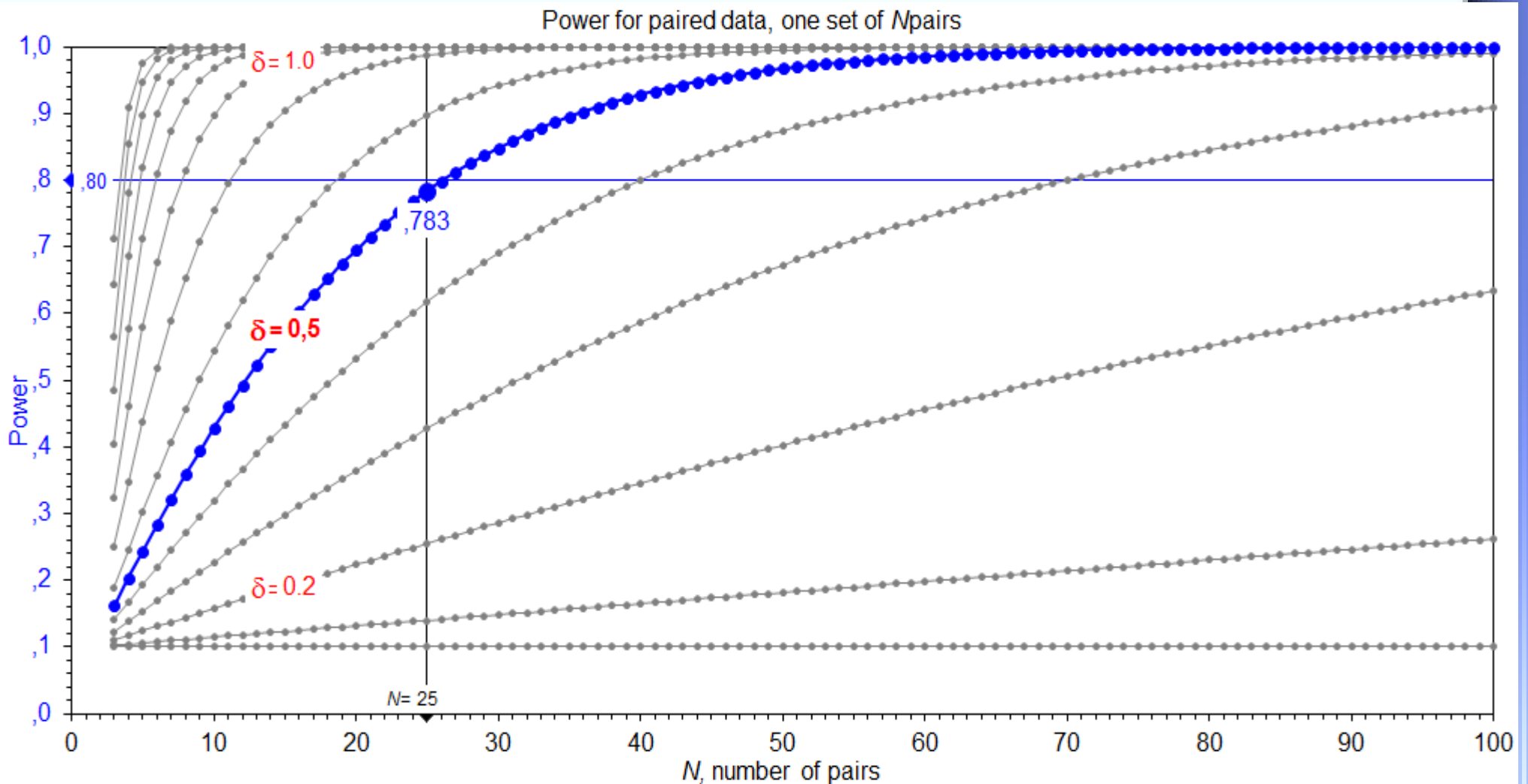
Power Between Ss

Power for two independent groups, each size N



Power Within Ss

- Power for within Ss studies is greater (*ceteris paribus*) but depends also on r (e.g., $r = .50$) between DVs



ANOVA Within and Mixed

Web app: GLIMMPSE (<https://glimmpse.samplesizeshop.org>)
but check also <https://samplesizeshop.org/>

2 x 2 Mixed ANOVA

◆ GLIMMPSE

General Linear Mixed Model Power and Sample Size

Design a Study

Welcome to GLIMMPSE. The GLIMMPSE software calculates power and sample size for study designs with normally distributed outcomes. Select one of the options below to begin a power or sample size calculation.

New Study

Start a new design.

Upload

You have previously used GLIMMPSE and wish to work on a saved design.

Solve for sample size

◆ GLIMMPSE

General Linear Mixed Model Power and Sample Size

2 x 2 Mixed ANOVA: Study title

Please pick a concise title for the study:

2 x 2 Mixed ANOVA

Solve for

Please indicate whether you would like to solve for power or total sample size.

If you have a rough idea of the number of research participants you will be able to recruit, then solve for power.

If you have few restrictions on recruitment then you may wish to solve for sample size.

Power

Sample Size

Target power

Progress  Help  Sa

Please choose one or more power values, for which you wish to calculate minimum sample size.

All target power values must be between 0 and 1, exclusive.



Target Power

remove

0.8





0.9



Define test and alpha

es : Statistical tests

Progress  Help  Sa

Please choose one or more statistical tests. If you are unsure which to pick, we recommend the Hotelling Lawley Trace test due to its equivalence to a mixed model test.

- Hotelling Lawley Trace
- Pillai-Bartlett Trace
- Wilks Likelihood Ratio
- Box Corrected
- Geisser-Greenhouse Corrected
- Huynh-Feldt Corrected
- Uncorrected



: Type I error rates

Progress  Help  Sa

A Type I error occurs when a scientist declares a difference when none is present in the population. The Type I error rate is the probability of that kind of error, a false positive, and is often referred to as α (alpha). A Type I error rate can range from 0 to 1. Although the most commonly used value is 0.05, we recommend 0.01.



Type I Error Rate

remove

0.05



Define Within factor

What is the name of the dimension you will be measuring?

The text entered in the "Dimension" text box indicates the dimension over which measures were taken (e.g. time, days, locations, etc.). The choice of "Type" indicates whether the repeated measures are numeric (e.g. time) or categorical (e.g. arm, leg, hand)

Dimension: Repeated measures: Number of measurements of time?

What type of data is time?



You must have between 2 and 10 repeats (inclusive)

Spacing

If the repeated measures are numeric, the spacing values must be unique nonnegative integers, in ascending order.

Measurement #1 at

Measurement #2 at

| Repeated Measure Dimension | Type | Measurements | Edit | Remove |
|----------------------------|---------|--------------|---|---|
| time | Numeric | ["1", "2"] |  |  |

Define Between factor

Clustering

Progress Help Save

An independent sampling unit provides one or more observations such that observations from one unit are statistically independent from any other distinct unit while observations from the same unit may be correlated.

In a clustered design, the independent sampling unit is a cluster, such as a community, school, or classroom. Observations within a cluster are correlated. The labels for observations within a cluster must be exchangeable. For example, child "ID" within classroom can be reassigned arbitrarily. In contrast, observations across time cannot be reassigned and should not be considered clustered observations. The common correlation between any pair of cluster members is termed the intraclass correlation or intracluster correlation.

To include clustering in the study, click "Add Clustering" and follow the prompts.

You may specify up to 10 levels of clustering.

[Add Clustering](#)**SKIP**

: Fixed predictors

Progress Help Save

Each independent sampling unit has one or more observations which are statistically independent from observations from any other unit.

GLIMMSE allows you to define fixed predictors which divide the independent sampling unit into groups. One common example of a fixed predictor is treatment, with values placebo and drug, for which the independent sampling unit is randomized to a placebo group or a drug group. Another is gender, with values male or female.

If the design has no fixed predictors, do not define any here.

[Define Fixed Predictor](#)

Define Between factor

Fixed predictors

Please name the predictor:

Fixed predictors

What type of data is Condition?

Fixed predictors

Please name at least two groups:

Groups:


- Control
- Experimental

Each independent sampling unit has one or more observations which are statistically independent from observations from any other unit.

GLIMMPSE allows you to define fixed predictors which divide the independent sampling unit into groups. One common example of a fixed predictor is treatment, with values placebo and drug, for which the independent sampling unit is randomized to a placebo group or a drug group. Another is gender, with values male or female.

If the design has no fixed predictors, do not define any here.



Fixed Predictors

| Name | Type | Units | Groups | Remove | Edit |
|-----------|---------|-------|-------------------------------|-------------------------------------|---|
| Condition | NOMINAL | | ["Control", "Experimental"] | <input checked="" type="checkbox"/> |  |



Select key hypothesis for power analysis

Hypothesis choice

Progress  Help  Save

Each power or sample size calculation is based on selecting a specific study hypothesis. The options below show the hypotheses which are available for the current study design. Specify the hypothesis that represents your scientific question.

GLIMMPSE chooses sensible contrast matrices based on cell means coding. Should you wish to define your own contrast matrices, pick the highest order interaction and choose from the advanced options in the hypothesis components.

Select a hypothesis from the list.



| | Effects Available for Consideration | Nature of Variation |
|----------------------------------|-------------------------------------|---------------------|
| <input checked="" type="radio"/> | Condition x time: Interaction | Between x Within |
| <input type="radio"/> | time: Main Effect | Within |
| <input type="radio"/> | Condition: Main Effect | Between |
| <input type="radio"/> | Grand Mean | Between |

Test hypothesis

Hypothesis

Progress 

What type of contrast do you wish among the means defined by your groups and repeated measures?

All mean differences zero

A parameter is a characteristic of a population. The parameters of interest are differences between groups at individual repeated measures.

The null hypothesis is that all pairwise differences between groups are the same among all pairs of repeated measures.

Show Advanced Options

Theta 0

Progress  Help  Sa

A hypothesis compares parameters to a constant, the contrast comparison constant, θ_0 . This is almost always zero. If you choose a value other than zero, be sure that you understand that the hypothesis you define is scientifically meaningful. Also note that the description and interpretation of your hypothesis given when choosing your contrasts will be affected.



$$\begin{bmatrix} \wedge \\ 0 \end{bmatrix}$$

Group size ratios

Progress 

For equal group sizes, input a "1" in the block next to each group. This is the default study design.

For unequal group sizes, specify the ratio of the group sizes. For example, consider a design with an active drug group and a placebo group. If twice as many study participants receive the placebo, a value of "2" would be selected for the placebo group, and a value of "1" would be selected for the active drug group.

Group size ratios

| | | |
|-----------|--------------|---|
| Condition | Control | 1 |
| | Experimental | 1 |

Expected means under key hypothesis

Marginal means

Progress (

The table below shows the mean values for outcome **Performance** within each group in the study. Each group is represented by a row in the table, and each repeated measure dimension is represented by a column.

Enter the mean values you expect to observe for outcome **Performance** within each group. The table should contain at least one value that is non-zero. Also, at least two groups should have means which differ by a scientifically meaningful amount.

Expected mean values, per group, for *Performance*

| | | time | |
|-----------|--------------|------|---|
| | | 1 | 2 |
| Condition | Control | 5 | 5 |
| | Experimental | 5 | 6 |

Set blank values to

value


Scale factors (different scenarios) and SD

Scale factor for the marginal means

Progress 

In power analysis, it is not possible to know the exact values of means before the experiment is observed. Scale factors allow you to consider alternative values for the means by scaling the values entered on the previous screen. For example, entering the scale factors 0.5, 1, and 2 would compute power for the mean values divided by 2, the mean values as entered, and the mean values multiplied by 2.

Enter a scale factor:



Scale Factor

remove

1




: Variability across outcomes

Enter the standard deviation you expect to observe for each outcome.

| Outcome | Standard Deviation |
|-------------|--------------------|
| Performance | 1 |

: Repeated measure standard deviation ratios

Progress 

Define the ratios of standard deviations for time. One of your values should be 1 and the others should represent the ratio of that value to that value:

For example, if you believe that the standard deviation doubles at each time, enter the values 1, 2, 4, 8... etc.

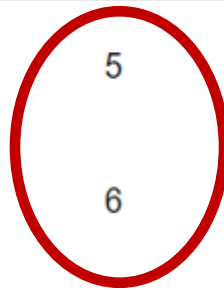
| time | Standard Deviation Ratio |
|------|--------------------------|
| 1 | 1 |
| 2 | 1 |

Expected Means (EM) vs. Effect Size (ES)

- How to relate EM and ES?
- Unless you have a sense of the strength of the effect in raw metrics, you can find useful to standardize values

| | | time | |
|-----------|--------------|------|---|
| | | 1 | 2 |
| Condition | Control | 5 | 5 |
| | Experimental | 5 | 6 |

as if
d=1.00
here



See also

http://shiny.ieis.tue.nl/anova_power/

and

<https://psyarxiv.com/baxsf/>

Enter the standard deviation you expect to observe for each outcome.

| Outcome | Standard Deviation |
|-------------|--------------------|
| Performance | 1 |

Repeated measures correlations and scale factors

Repeated measure correlation

Progress

For a given research participant, responses vary across outcomes and across repeated measurements. The amount of variability can dramatically impact power and sample size.

Define the **time** correlation matrix, by entering correlations you expect to observe among the chosen spacing values of **time**:

Unstructured

time

1 2

$$\begin{bmatrix} 1 & 0,5 \\ 0,5 & 1 \end{bmatrix}$$

(each off-diagonal correlation must be between -1 and 1, exclusive)

: Scale factor variance

Progress

Changes in variability can dramatically affect power and sample size results. It is not possible to know the variability until the experiment is observed. Scale factors allow you to consider alternative values for variability by scaling the calculated covariance matrix. For example, entering the scale factors 0.5, 1, and 2 would compute power for the covariance matrix divided by 2, the covariance matrix as entered, and the covariance matrix multiplied by 2.

You may add up to 10 scale factors.

Choose a number greater than zero



Scale Factor

remove

1



Finally, the calculation...

Calculate

Progress  Help 

Calculate

Download result

Results Matrices Design

Design



Hypothesis



Design Dimensions



Parameters



Optional Specifications



...and the results!

Calculate

Download result

Results Matrices Design

| Power | Total Sample Size | Target Power | Means Scale Factor | Variability Scale Factor | Test | Power Method | Type I Error Rate |
|-------|-------------------|--------------|--------------------|--------------------------|------------------------|--------------|-------------------|
| 0.807 | 34 | 0.8 | 1 | 1 | Hotelling Lawley Trace | conditional | 0.05 |
| 0.912 | 46 | 0.9 | 1 | 1 | Hotelling Lawley Trace | conditional | 0.05 |

Suppose expected correlation is lower

time

| | | |
|---|------|------|
| | 1 | 2 |
| 1 | 1 | 0,25 |
| 2 | 0,25 | 1 |

(each off-diagonal correlation must

Calculate

Download result

Results Matrices Design

| Power | Total Sample Size | Target Power | Means Scale Factor | Variability Scale Factor | Test | Power Method | Type I Error Rate |
|-------|-------------------|--------------|--------------------|--------------------------|------------------------|--------------|-------------------|
| 0.807 | 50 | 0.8 | 1 | 1 | Hotelling Lawley Trace | conditional | 0.05 |
| 0.904 | 66 | 0.9 | 1 | 1 | Hotelling Lawley Trace | conditional | 0.05 |

Suppose no correlation

time

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

(coefficient diagonal correlation matrix)

Calculate

Download result

Results Matrices Design

| Power | Total Sample Size | Target Power | Means Scale Factor | Variability Scale Factor | Test | Power Method | Type I Error Rate |
|-------|-------------------|--------------|--------------------|--------------------------|------------------------|--------------|-------------------|
| 0.808 | 66 | 0.8 | 1 | 1 | Hotelling Lawley Trace | conditional | 0.05 |
| 0.906 | 88 | 0.9 | 1 | 1 | Hotelling Lawley Trace | conditional | 0.05 |

Recap Examples Mixed ANOVA

- 1) The design was a 2 x 2 Mixed ANOVA
- 2) We varied the expected correlations

Required N for power at .80

- $r=.00$, $N= 66$
- $r=.25$, $N= 50$
- $r=.50$, $N= 34$





Required N goes down as the correlation between DVs of the Within factor goes up

Suppose instead a 2 x 2 Between Ss

The design you've described, means that every level of **Group** occurs at every level of **Condition**. This concept applies to every combination of fixed predictors.

Define Fixed Predictor

Fixed Predictors

| Name | Type | Units | Groups | Remove | Edit |
|-----------|---------|-------|---|---|---|
| Condition | NOMINAL | | ["Control", "Experimental"] |  |  |
| Group | NOMINAL | | ["No previous experience", "Previous experience"] |  |  |

Effects Available for Consideration

Nature of Variation

Condition x Group: Interaction Between x Between

Expected mean values, per group, for *Performance*

| Condition, Group | Mean Value |
|--------------------------------------|------------|
| Control, No previous experience | 5 |
| Control, Previous experience | 5 |
| Experimental, No previous experience | 5 |
| Experimental, Previous experience | 6 |

Calculate

Download result

Results Matrices Design

| Power | Total Sample Size | Target Power | Means Scale Factor | Variability Scale Factor | Test | Power Method | Type I Error Rate |
|-------|-------------------|--------------|--------------------|--------------------------|------------------------|--------------|-------------------|
| 0.801 | 128 | 0.8 | 1 | 1 | Hotelling Lawley Trace | conditional | 0.05 |
| 0.903 | 172 | 0.9 | 1 | 1 | Hotelling Lawley Trace | conditional | 0.05 |

Suppose instead a 2 x 2 Within Ss (r=.25)

The design you have described, means that every level of **Variable B** is measured at every level of **Variable A**. This concept applies to every combination of repeated measures.

Expected mean values, per group, for Performance

Define Repeated Measure

| Repeated Measure Dimension | Type | Measurements | Edit | Remove |
|----------------------------|-------------|--------------|------|--------|
| Variable A | Categorical | ["1", "2"] | | |
| Variable B | Categorical | ["1", "2"] | | |

Effects Available for Consideration

| Effects Available for Consideration | Nature of Variation |
|---|---------------------|
| <input checked="" type="radio"/> Variable A x Variable B: Interaction | Within x Within |

Variable A, Variable B

| | 1,1 | 1,2 | 2,1 | 2,2 |
|------------|-----|-----|-----|-----|
| Variable A | 5 | 5 | 5 | 6 |

| | Variable A | | Variable B | |
|------------|------------|------|------------|---|
| | 1 | 2 | 1 | 2 |
| Variable A | 1 | 0,25 | 0,25 | 1 |

Calculate

Download result

| Results | | Matrices | Design | | | | |
|---------|-------------------|--------------|--------------------|--------------------------|------------------------|--------------|-------------------|
| Power | Total Sample Size | Target Power | Means Scale Factor | Variability Scale Factor | Test | Power Method | Type I Error Rate |
| 0.807 | 20 | 0.8 | 1 | 1 | Hotelling Lawley Trace | conditional | 0.05 |
| 0.904 | 26 | 0.9 | 1 | 1 | Hotelling Lawley Trace | conditional | 0.05 |

2 x 2 Within Ss with $r=.50$ and $r=.0$

| Variable A | | Variable B | |
|------------|-----|------------|-----|
| 1 | 2 | 1 | 2 |
| 1 | 0,5 | 1 | 0,5 |
| 0,5 | 1 | 0,5 | 1 |

Calculate

Download result

Results Matrices Design

| Power | Total Sample Size | Target Power | Means Scale Factor | Variability Scale Factor | Test | Power Method | Type I Error Rate |
|-------|-------------------|--------------|--------------------|--------------------------|------------------------|--------------|-------------------|
| 0.803 | 10 | 0.8 | 1 | 1 | Hotelling Lawley Trace | conditional | 0.05 |
| 0.911 | 13 | 0.9 | 1 | 1 | Hotelling Lawley Trace | conditional | 0.05 |

| Variable A | | Variable B | |
|------------|---|------------|---|
| 1 | 2 | 1 | 2 |
| 1 | 0 | 1 | 0 |
| 0 | 1 | 0 | 1 |

Results Matrices Design

| Power | Total Sample Size | Target Power | Means Scale Factor | Variability Scale Factor | Test | Power Method | Type I Error Rate |
|-------|-------------------|--------------|--------------------|--------------------------|------------------------|--------------|-------------------|
| 0.808 | 34 | 0.8 | 1 | 1 | Hotelling Lawley Trace | conditional | 0.05 |
| 0.900 | 44 | 0.9 | 1 | 1 | Hotelling Lawley Trace | conditional | 0.05 |

Power Comparison

- Three 2 x 2 ANOVA designs (Mixed, Between, Within)
- In each design the same pattern of expected means
- Always $SD=1$
- Always powered for interaction effect
- Required N for power at .80
 - Between** = 128
 - Mixed (r=.00)** = 64
 - Mixed (r=.25)** = 50
 - Mixed (r=.50)** = 34
 - Within (r=.00)** = 34
 - Within (r=.25)** = 20
 - Within (r=.50)** = 10
- You can draw your own conclusion...

| | A1 | A2 |
|----|----|----|
| B1 | 5 | 5 |
| B2 | 5 | 6 |

How to increase power?

- **Increase sample size** (also multi-lab collaborations)
- Use blocking or repeated measures (**within**) design BUT sometimes can be inappropriate
- Administer stronger treatments (e.g., experimental manipulation) BUT be wary of possible reduced ecological validity
- Avoid restrictions of range for dependent variables
- Standardize experimental procedures
- Increase reliability of measures
- Use more homogenous subject samples BUT increased risks to generalizability of results
- Meta-analytic mindset

Increasing power without increasing sample size

$$SE = \sqrt{\frac{S^2}{n}}$$

Increasing Statistical Power Without Increasing Sample Size

Gary H. McClelland
University of Colorado at Boulder
August 2000 • *American Psychologist* 963

Increasing the Power of Your Study by Increasing the Effect Size

TOM MEYVIS
STIJN M. J. VAN OSSELAER
[Journal of Consumer Research](#), Feb2018, Vol. 44 Issue 5, p1157-1173.

- Standard errors depend on N and SD (smaller SD means smaller SE)
- SE can be reduced with more reliable measures (more trials, more items), more precise experimental designs, less Ss variability (e.g., also within Ss designs)
- Plan your design as simple and as clean as possible

$$SE = \sqrt{\frac{S^2}{n}}$$

Distinguish conceptually between unnecessary (“added noise”) and necessary (“natural”) variance

Improve your design. Optimize it. Think carefully about it. Few extra hours spent on this can be worth hundreds of extra participants (and avoid frustrations...)

Reduce the **noise**! Increase the **signal**!

Summing up Power Analysis

- Power analysis is one important way to efficiently plan a study
 - Try to power your study adequately
 - A main problem is to best guess a predicted ES
 - Beware of the uncertainty of ES estimates and the asymmetric impact of ES estimate errors
 - Wise to consider uncertainty in the ES estimate (e.g., by running different scenarios)
 - Think in terms of range of values rather than a specific value
-

What does it really mean to have enough power?



The Crest-tailed Mulgara is a species of marsupial that was recently rediscovered living in an area where it had been presumed extinct for about 100 years (Credit: Reece Pedler)



Power as fuel in the tank



- Have enough fuel to find what you are looking for (hoping that it is there) in a place at a distance that you hope have guessed reasonably well

Some readings for some advanced issues

Contrast, regression, moderation, and mediation effects

- Perugini, M., Gallucci, M., & Costantini, G. (2018). A Practical Primer To Power Analysis for Simple Experimental Designs. *International Review of Social Psychology*, 31(1).

Within and Mixed ANOVA

- Guo, Y., Logan, H. L., Glueck, D. H., & Muller, K. E. (2013). Selecting a sample size for studies with repeated measures. *BMC medical research methodology*, 13(1), 100
- Web app: GLIMMPSE (<https://glimmpse.samplesizeshop.org>)

Mixed/Multilevel Models

- Judd, C. M., Westfall, J., & Kenny, D. A. (2016). Experiments with more than one random factor: Designs, analytic models, and statistical power. *Annual Review of Psychology*. Web app: https://jakewestfall.shinyapps.io/two_factor_power/
- See also Brysbaert, M., & Stevens, M. (2018). Power analysis and effect size in mixed effects models: a tutorial. *Journal of Cognition*, 1(1).
- Kelcey, B., Xie, Y., Spybrook, J., & Dong, N. (2020). Power and sample size determination for multilevel mediation in three-level cluster-randomized trials. *Multivariate Behavioral Research* <https://www.causalevaluation.org/power-analysis.html>

Simulation based power analysis

- Gelman, A., Hill, J. (2006) *Data analysis using regression and multilevel/hierarchical models*. Cambridge: Cambridge University Press.

Advanced models and exemplary R code

- Liu, X. S. (2014). *Statistical Power Analysis for the Social and Behavioral Sciences: Basic and Advanced Techniques*. New York: Routledge.

First tip for getting it right

(many more will come

to those who wait)

Back to the problem

- As scientists, we all want to get something right
 - If we get it right, it is replicable and will be replicated
 - But what does it mean “to get it right”?
 - So, what can we do to increase our chances?
 - Some pointers (today only the first episode)
-

1st pointer: Power

- Design your study with adequate power (*probability of finding an effect if it does exist*)
 - Underpowered studies produce conflicting evidence and false negatives **but also false positives** (Maxwell, 2004; Ioannidis, 2005)
 - Direct effect on **False Negatives** but also indirect effect on **False Positives**
(**False Discovery Rate / True False Positives**)
-

Why many effects are not replicated?

- A mix of different factors and possible explanations
 - Two main factors
 - a) Low power and b) Publication bias**
 - Under these conditions, it is predictable that the literature will contain many false positives (results that seems significant but are not) and artificially boosted effect sizes
 - Hence effects will be difficult to replicate
-

1.a Low power

Is a real problem for ψ ?

META-RESEARCH ARTICLE

Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature

Denes Szucs^{1*}, John P. A. Ioannidis²

¹ Department of Psychology, University of Cambridge, Cambridge, United Kingdom, ² Meta-Research Innovation Center at Stanford (METRICS) and Department of Medicine, Department of Health Research and Policy, and Department of Statistics, Stanford University, Stanford, California, United States of America

We have empirically assessed the distribution of published effect sizes and estimated power by analyzing 26,841 statistical records from 3,801 cognitive neuroscience and psychology papers published recently. The reported median effect size was $D = 0.93$ (interquartile range: 0.64–1.46) for nominally statistically significant results and $D = 0.24$ (0.11–0.42) for nonsignificant results. Median power to detect small, medium, and large effects was 0.12, 0.44, and 0.73, reflecting no improvement through the past half-century. This is so because sample sizes have remained small. Assuming similar true effect sizes in both disci-

Researchers' Intuitions About Power in Psychological Research



Marjan Bakker¹, Chris H. J. Hartgerink¹, Jelte M. Wicherts¹, and Han L. J. van der Maas²

¹Department of Methodology and Statistics, Tilburg School of Social and Behavioral Sciences, Tilburg University, and ²Department of Psychology, Psychological Methods, University of Amsterdam

1990; Maxwell, 2004). Specifically, given the typical effect sizes (ESs) and sample sizes reported in the psychological literature, the statistical power of a typical two-group between-subjects design has been estimated to be less than .50 (Cohen, 1990) or even .35 (Bakker et al., 2012). These low power estimates appear to con-

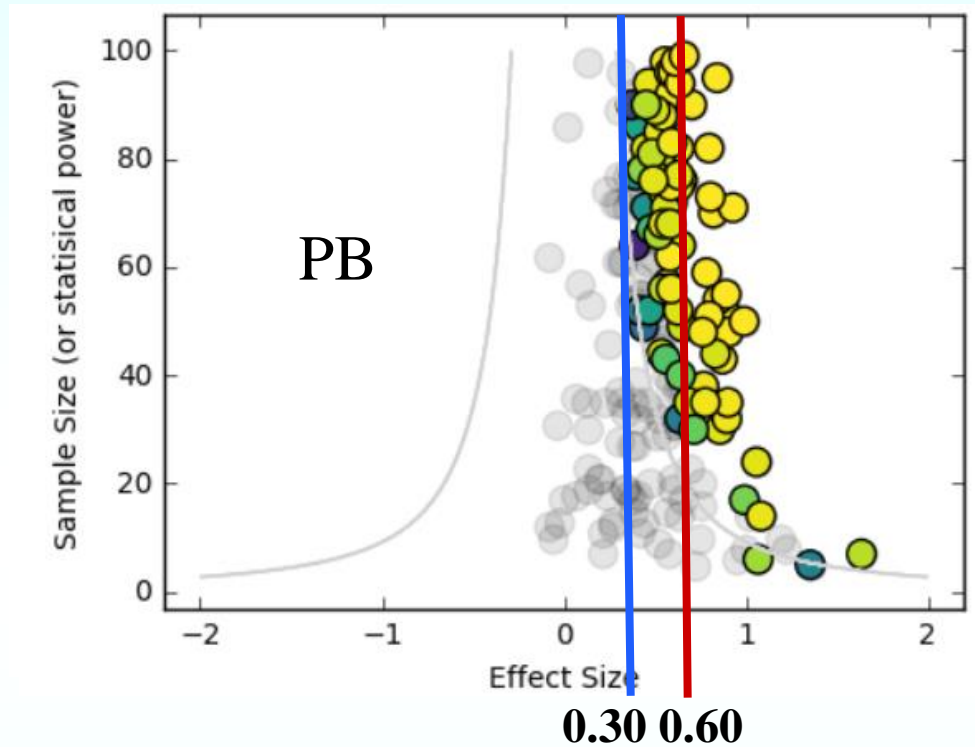
Psychological Science
2016, Vol. 27(8) 1069–1077
© The Author(s) 2016
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0956797616647519
pss.sagepub.com
SAGE

Yes!

1.b Publication bias

- Tendency to publish mainly significant results (and to submit for publication mainly studies with significant results)
 - There are sometimes understandable reasons (unclear evidence, contradictory support, pilot studies, tentative paradigms, etc.)
 - But often is a by-product of confirmation/positivity biases and insufficient culture of cumulative knowledge in a scientific field
-

Publication bias



The ES will be overestimated. How much depends on the extent of PB and on the prevalence of small samples.

A reader will think that Cohen's $d=0.60$ but in fact is $d=0.30$

Publication bias, Effect Sizes, underpowered studies

ES: Cohen's $d=0.60$ (vs. $d=0.30$)

N for power:

80%

90%

72 Ss (vs. 278)

98 Ss (vs. 382)

Suppose we run a study with 98 Ss.

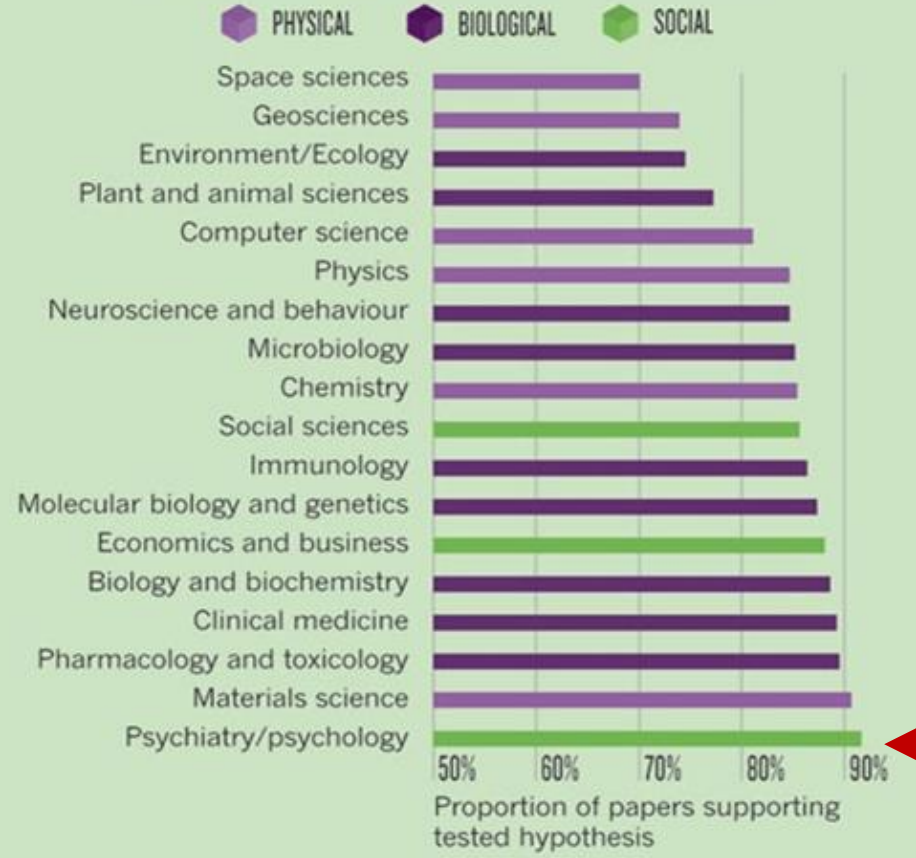
Expected power is 0.90 but **real power will be 0.43**

Vicious cycle: PB leads to overestimated ES leading to underpowered studies leading to non replicated effects, **even assuming that the effects are true and the researchers do not “cheat”**

Is there publication bias in science?

ACCENTUATE THE POSITIVE

A literature analysis across disciplines reveals a tendency to publish only 'positive' studies — those that support the tested hypothesis. Psychiatry and psychology are the worst offenders.



YES



(Fanelli, 2010)

Publication bias, Effect Sizes, sample sizes

Without publication bias, there should be **no** relation ($r=0$)

September 2014 | Volume 9 | Issue 9 | e105825

OPEN ACCESS Freely available online

PLOS ONE

Publication Bias in Psychology: A Diagnosis Based on the Correlation between Effect Size and Sample Size

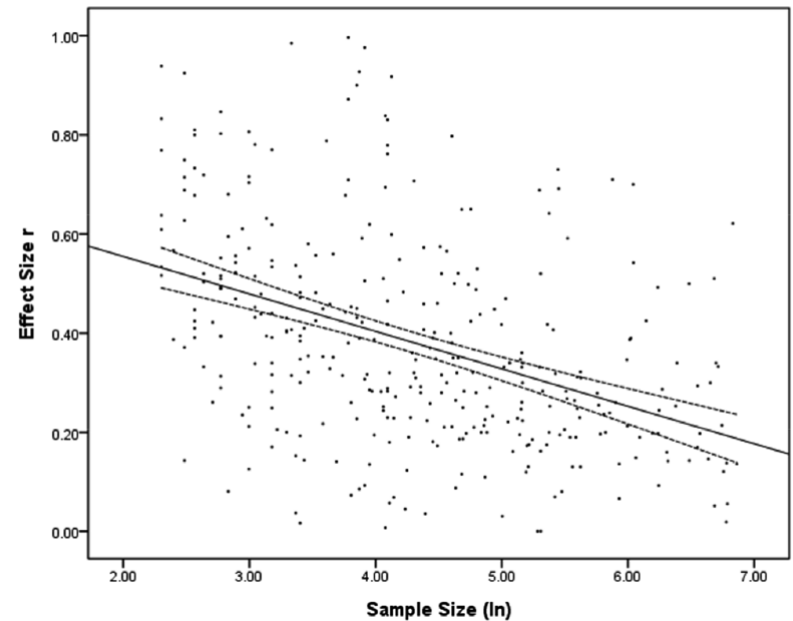
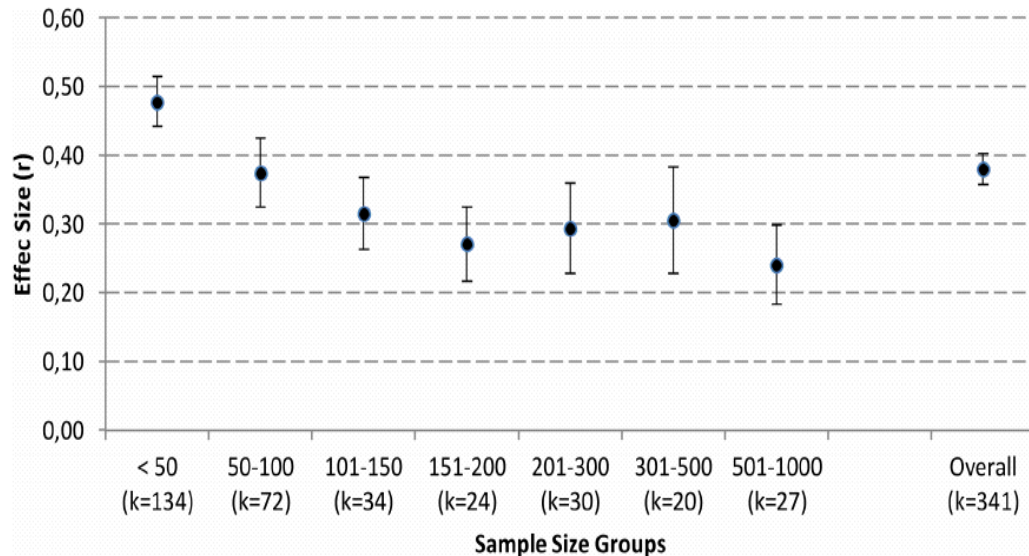
YES

Anton Kühberger^{1,2*}, Astrid Fritz³, Thomas Scherndl¹

¹ Department of Psychology, University of Salzburg, Salzburg, Austria, ² Centre for Cognitive Neuroscience, University of Salzburg, Salzburg, Austria, ³ Österreichisches

Methods: We investigate whether effect size is independent from sample size in psychological research. We randomly sampled 1,000 psychological articles from all areas of psychological research. We extracted p values, effect sizes, and sample sizes of all empirical papers, and calculated the correlation between effect size and sample size, and investigated the

$r = .54!$



Publication bias, Effect Sizes, sample sizes

Without publication bias, there should be no relation

Neuroinform (2012) 10:67–80
DOI 10.1007/s12021-011-9125-y

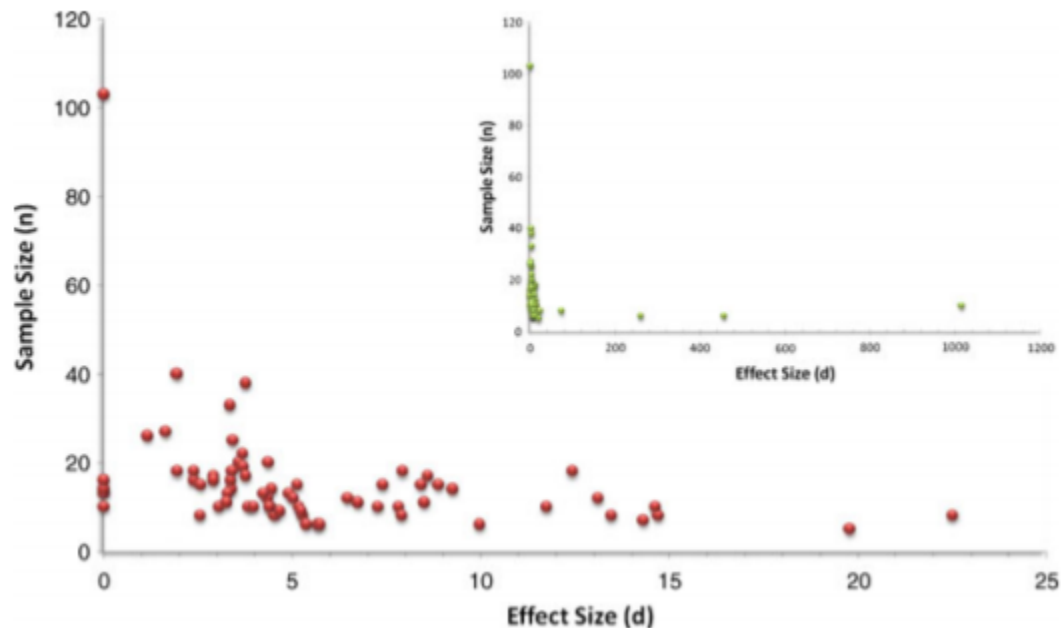
ORIGINAL ARTICLE

Publication Bias in Neuroimaging Research: Implications for Meta-Analyses

YES

Robin G. Jennings · John D. Van Horn

Fig. 3 Funnel plot of Cohen's d by sample size for studies without extreme values ($n=70$). While a 'large' Cohen's d value is usually $d > 0.8$, most of our values are between 1 and 25, with funnel plot asymmetry due to the heavy right-tail evident here. **Inset:** Funnel plot of Cohen's d by sample size for each study ($n=74$), showing the four extreme outlier values



Conclusions

- Increase sample size (trials/items) if you want to get it right
 - Decrease unnecessary variation
 - Decrease **noise** and increase **signal** in the study
 - To get it right means to reduce False positives (Type I error), False negatives (Type II error) and to have reasonably precise estimates
-

and remember...

