UNIVERSITA' DEGLI STUDI DI MILANO BICOCCA

Sample planning

Sample planning

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

Some basic issues

- **Representativeness** (inference from sample to population)
- **Generalizability** (inference from sample to population)
- **Robustness** (inference that you can "trust")
- **Feasibility** (something that you can do)
- **Efficiency** (most value for money, "spend" as little as possible)

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 <u>few</u> participants as needed to reach the conclusion that you want to reach

- **Redundancy**: collect as many participants are needed to reach the conclusion that you want to reach with not too much error (trade-off between accuracy and efficiency)

Some statistical approaches to sample size planning

- **AIPE** (Maxwell, 2008): decide sample size based on a chosen level of Accuracy In Parameter Estimation
- **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
- **Heuristic**: in different fields there are “magical” rules ($N \geq 20$ per cell, $N>100$, ratio k/N). At best, approximate wise suggestions, at worse misleading
- **Power analysis**: Today and tomorrow

Some references

Sample Size Planning for Statistical Power and Accuracy in Parameter Estimation

Scott E. Maxwell,[1] Ken Kelley,[2] and Joseph R. Rausch[3]

[1]Department of Psychology, University of Notre Dame, Notre Dame, Indiana 46556; email: smaxwell@nd.edu
[2]Inquiry Methodology Program, Indiana University, Bloomington, Indiana 47405; email: kkiii@indiana.edu
[3]Department of Psychology, University of Minnesota, Minneapolis, Minnesota 55455; email: rausch@umn.edu

Annu. Rev. Psychol. 2008. 59:537–63

Key Words

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

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

CrossMark

BRIEF REPORT

Bayes factor design analysis: Planning for compelling evidence

Felix D. Schönbrodt[1] · Eric-Jan Wagenmakers[2]

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

1082-989X/17/$12.00 http://dx.doi.org/10.1037/met0000061

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

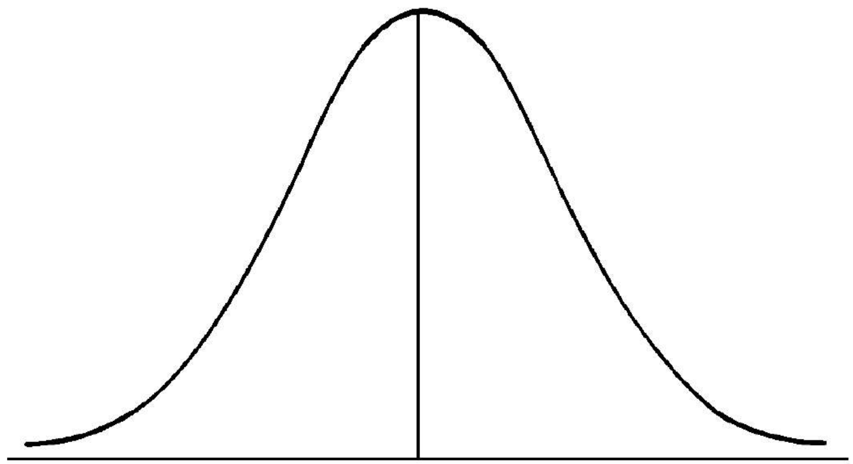
Marco Perugini
University of Milan–Bicocca

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

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

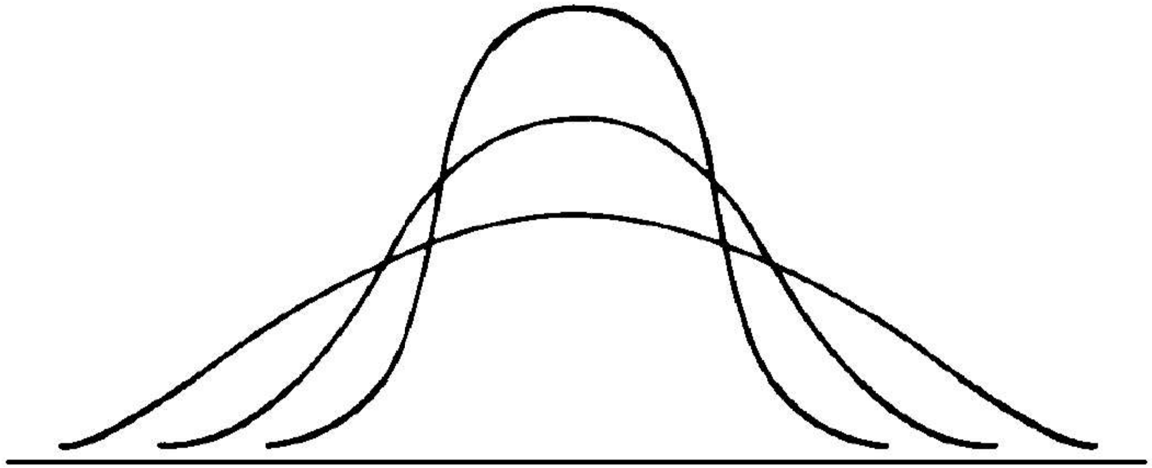


Variance and standard deviation

- Reflects the dispersion (variability) around the mean

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

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



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)

UNIVERSITA' DEGLI STUDI DI MILANO BICOCCA

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

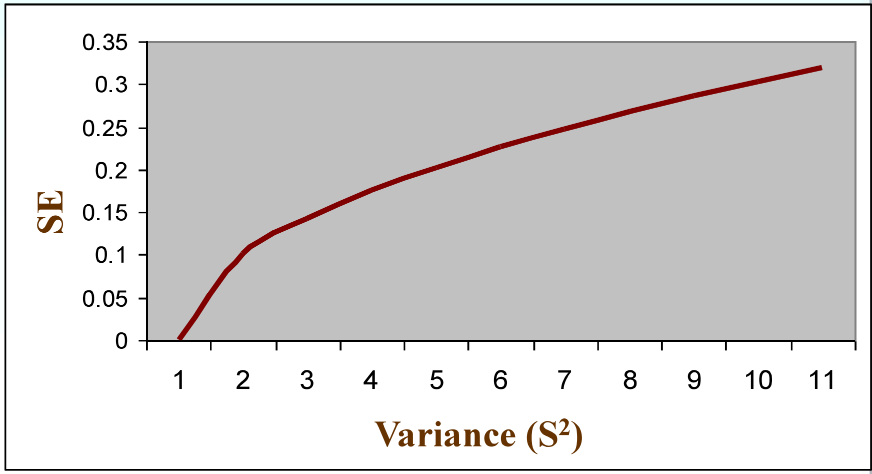
Standard error

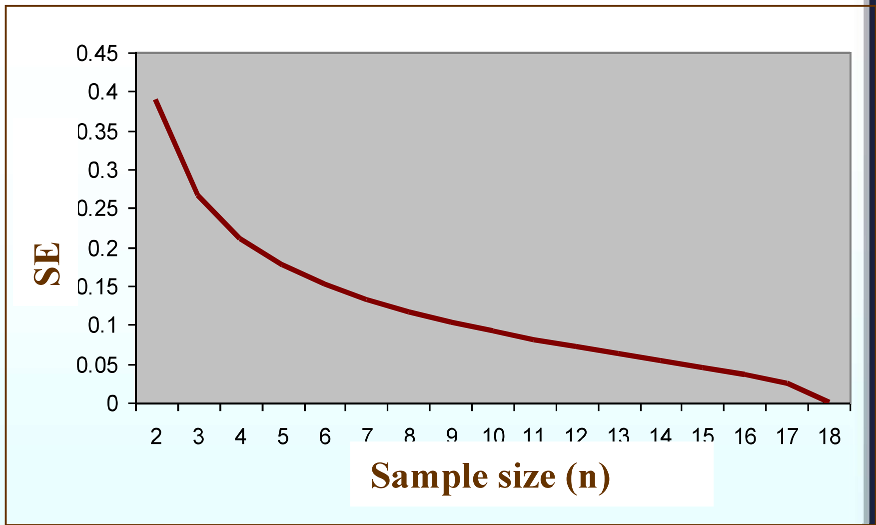
- Error in estimating a population parameter (e.g., mean) from a sample

Goes up with increasing variance

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

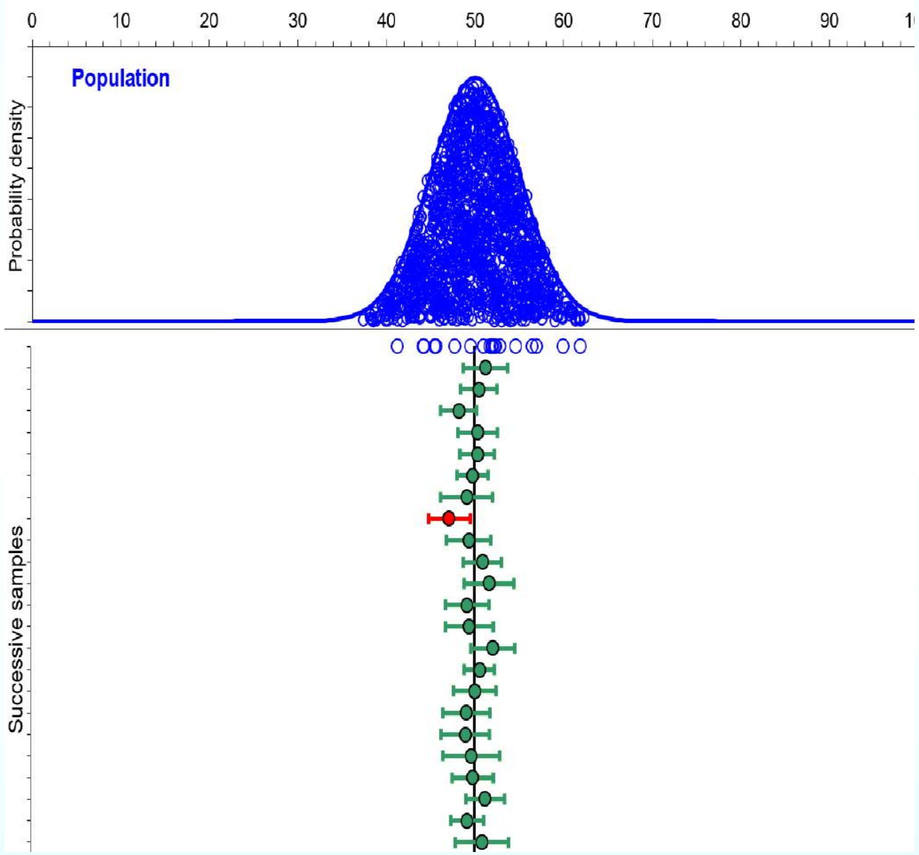
Goes down with increasing sample size





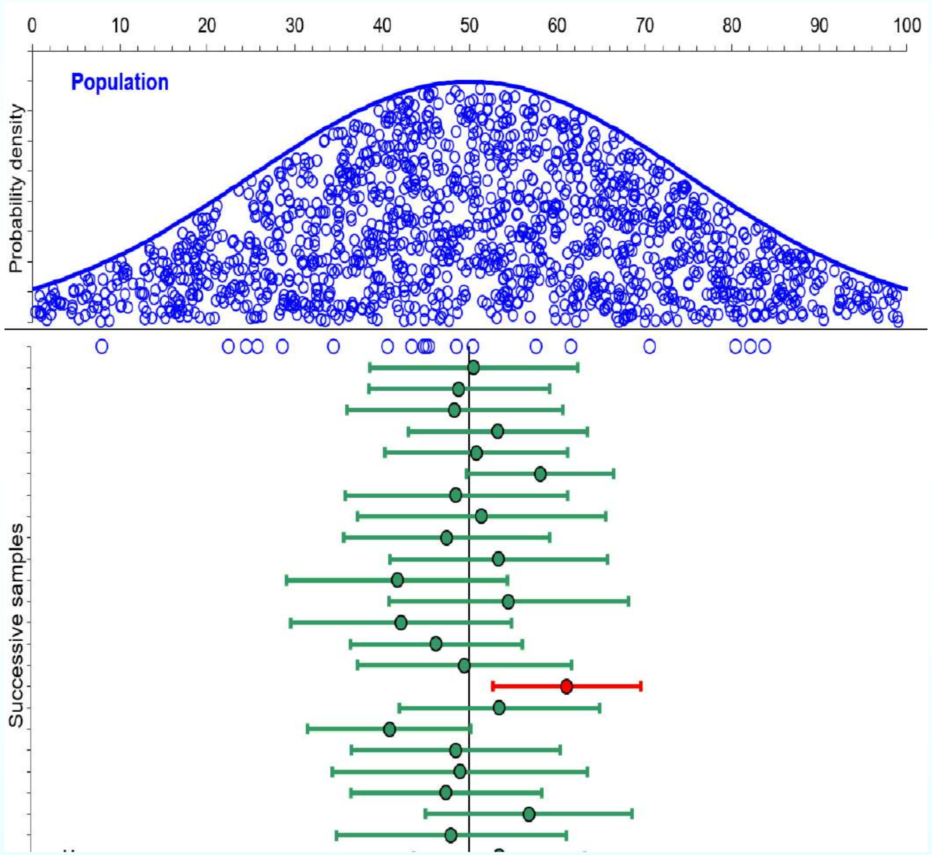
Parameter estimation: Error and variability

N=20



Small variability = small SE

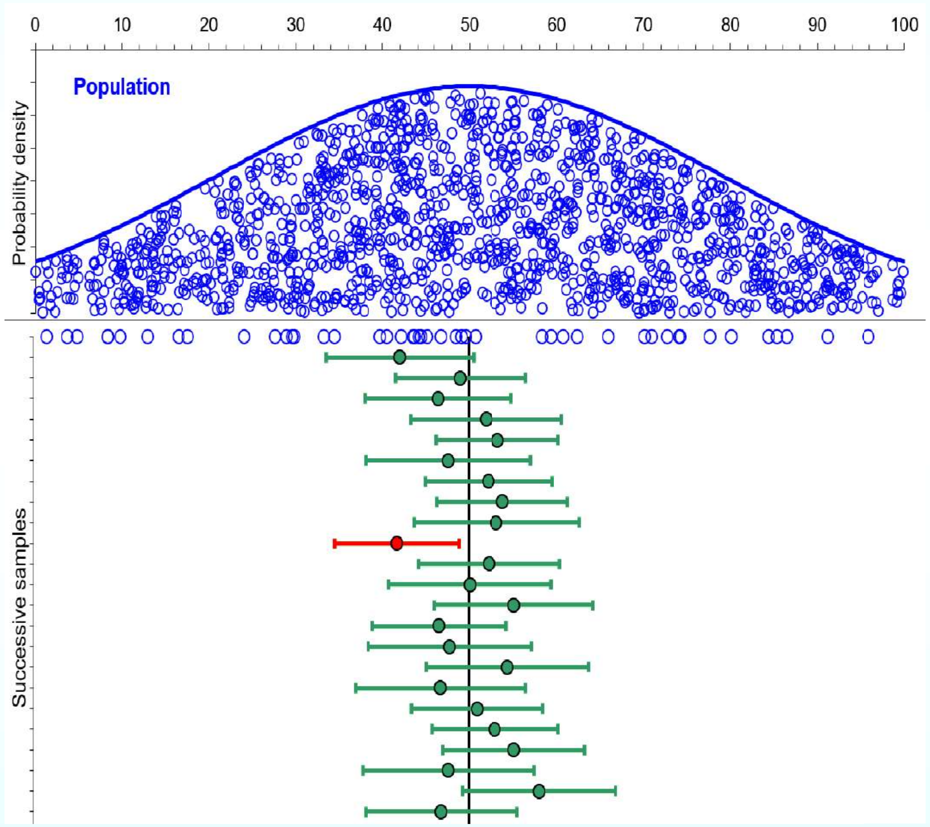
N=20



Large variability = large SE

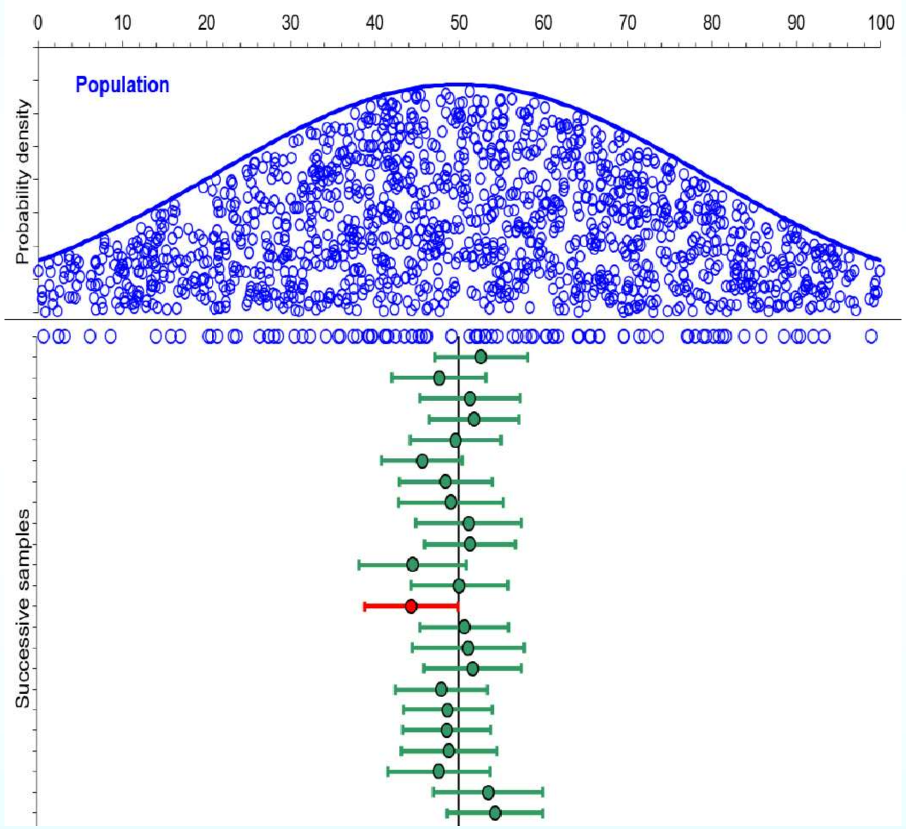
Error and variability

N=50



Large variability = large SE

N=100



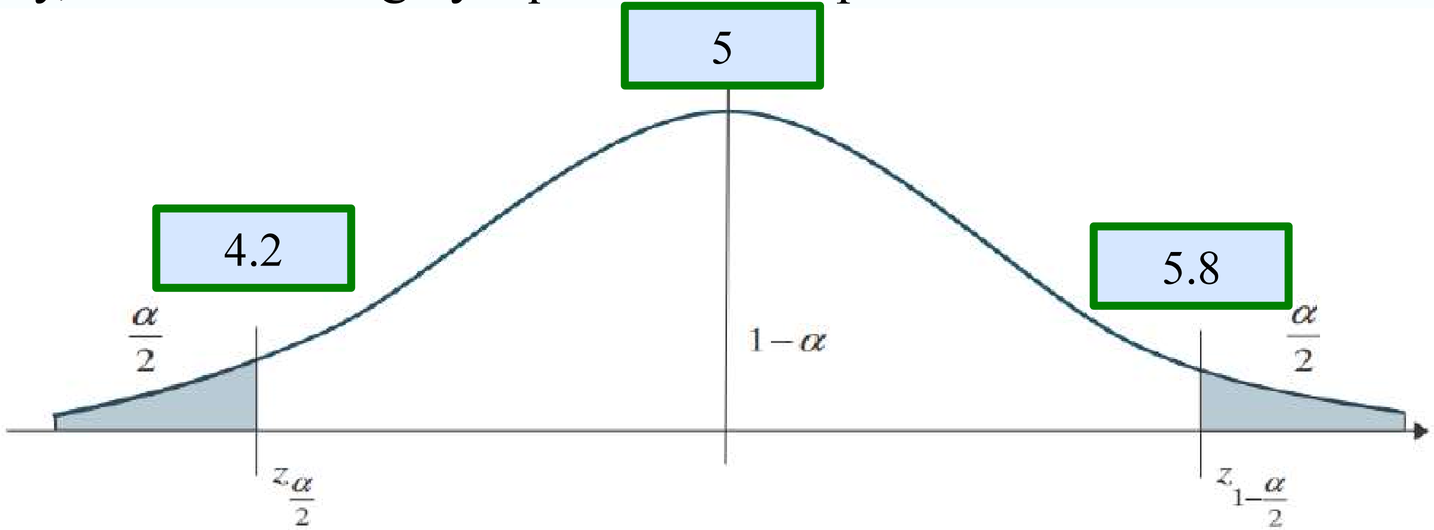
Small variability = small SE

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 = $\sqrt{\frac{4^2}{100}}$ o $\frac{4}{\sqrt{100}}$ =0.4
Range: 2 x SE = 0.8
95% CI = [4.2, 5.8]

$$\mu \in \left(\overline{X}_n \pm t^{(n-1)}_{1-\frac{\alpha}{2}} \boxed{\sqrt{\frac{\overline{s}_n^2}{n}}} \right)$$

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

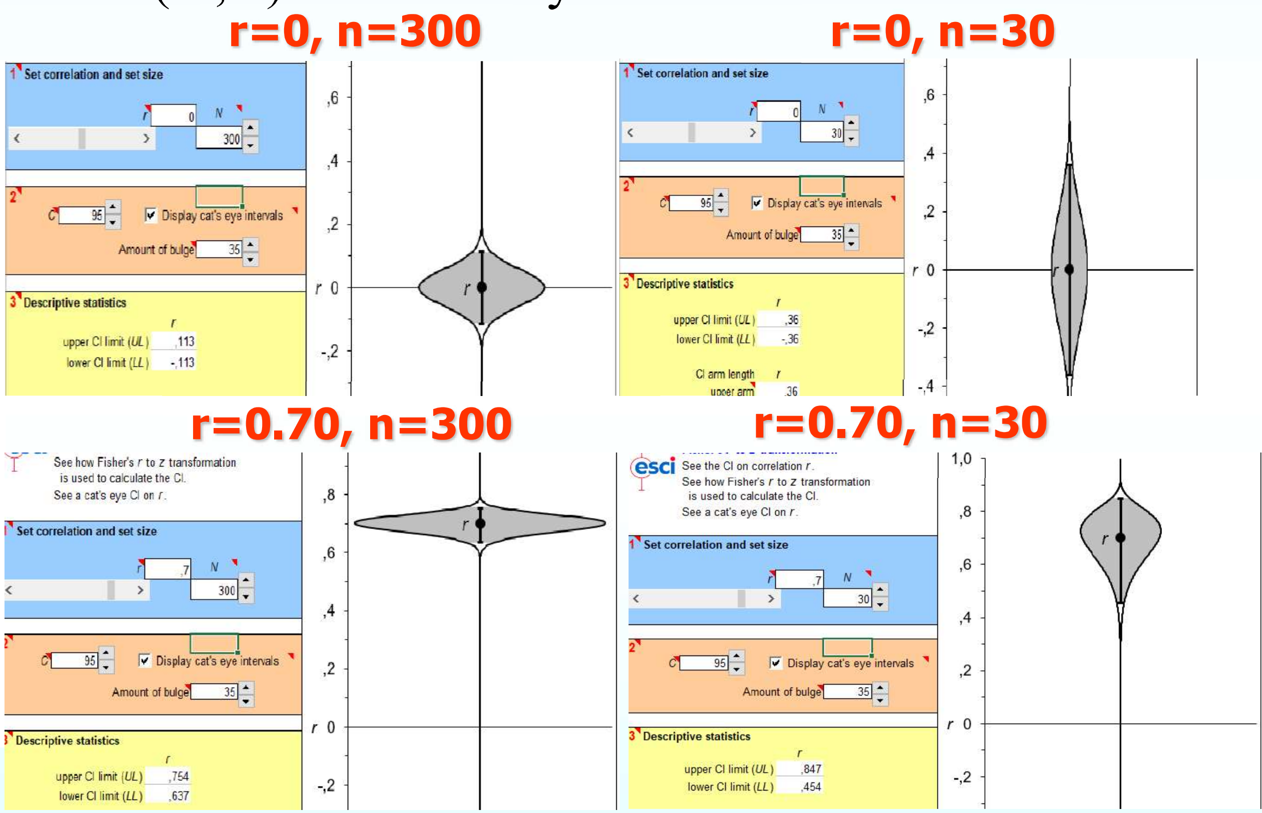
Define Fisher Transformation: $z_r = \frac{\ln\left(\frac{1+r}{1-r}\right)}{2}$

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:

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



Hypothesis testing I

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 between certain groups (e.g., experimental vs. control, male vs. female)

- This is the realm of **hypothesis testing** (or statistical inferences from data)

Statistical inference on a parameter

- After computing a parameter, we wish to exclude the possibility that the observed values are not simply due to chance (e.g, their difference from 0; their difference between groups)

Statistical Inference

- Null-Hypothesis Significance Testing (NHST)

Testing hypotheses on a parameter

- Suppose I find in a sample (N=100) that A (alcohol consumption) and B (aggressive behavior) are correlated r=.253. I want to say that there is a relationship between consuming alcohol and being aggressive.

- You do not believe it, and you think **it was a fluke**, but in reality there is no relationship between the two variables (r=0): one can be aggressive without drinking alcohol or drink alcohol without becoming more aggressive

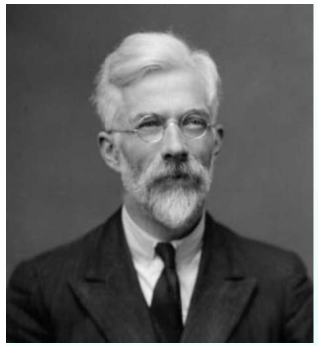
- I want to show you that it was not a fluke. I use a Sherlock Holmes-like approach: *if this result is by chance, how likely would I be to find it?* Depending on the results, you could become less skeptical

II

THE PRINCIPLES OF EXPERIMENTATION, ILLUSTRATED BY A PSYCHO-PHYSICAL EXPERIMENT

5. Statement of Experiment

A LADY declares that by tasting a cup of tea made with milk she can discriminate whether the milk or the tea infusion was first added to the cup. We will consider the problem of designing an experiment by means of which this assertion can be tested. For this purpose

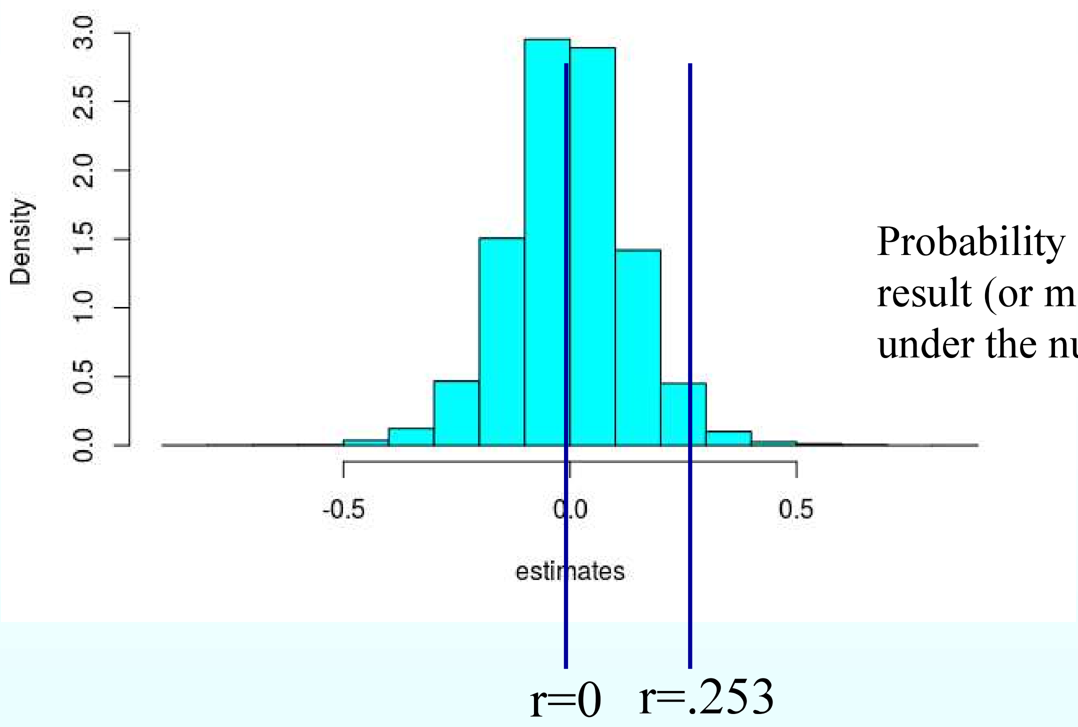


Null-hypothesis: the what-if scenario

- Let assume you are right (it was a fluke), and the actual relation between hours of studying and grade is null
- This means that in the **population,** r=0. Let’s call this population value the null-hypothesis population
- What if I could repeat my study hundreds of time, drawing each time an equivalent sample from the **null-hypothesis population**.
- Every time I would have a new estimate of my r in any given sample

Null-hypothesis: the what-if scenario, should r be 0

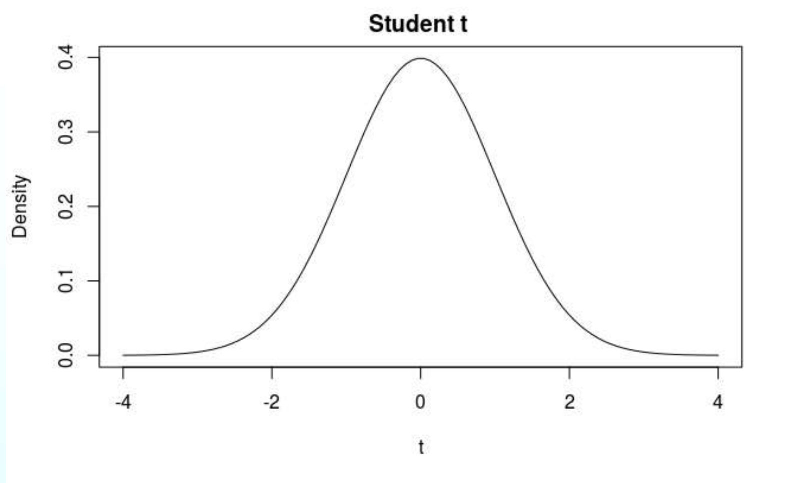
- We could ask how likely is to obtain the result, r=.253, under the null-hypothesis scenario (that is, if r=0 in the population) by extracting thousands of samples and looking at the distribution of the results



Probability of observing our result (or more extreme) under the null-hypothesis

Null-hypothesis: a priori distribution

- Under the given circumstances (the assumptions of the test) we know a-priori what will be the distribution of all possible estimates under the null-hypothesis
- We know that if we keep resampling from a **normal** distribution where r=0, we can use a Student-t distribution to estimate the probability of values

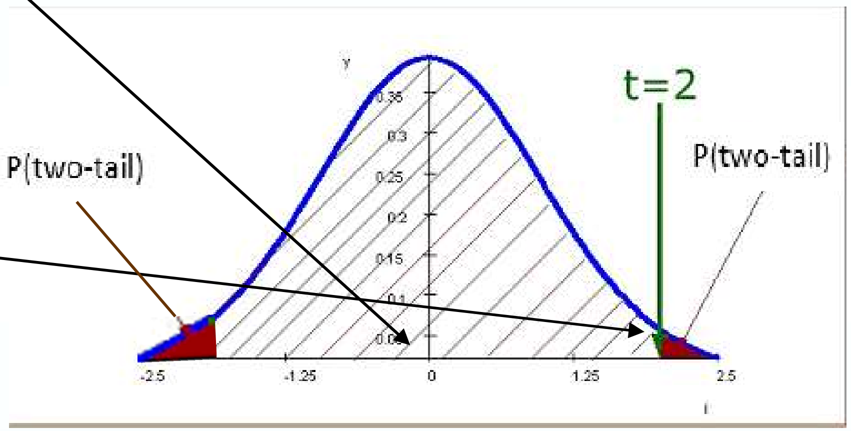


t-test

- Thus we can compare our result with the expected distribution under the null hypothesis

  - Null hypothesis: r=0

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



- This is a t-test: **p is the probability of obtaining our observed result (or larger) if the null hypothesis is true. Here t=2.589, and p=.011. What does it mean?**

p-value: How small is small?

- Conventional cut-off values are **0.05** (5% of error), and **0.01** (1% of error). Newer approaches suggest **0.005** in some circumstances (see after).

- These values are almost always used in social science research. It means that if you reject the null hypothesis if and only if the probability (***p***) of your value is less than 0.05 (or 0.01), you will be mistaken at most 5% (or 1%) of the times.

- Although these values represent a widespread standard, they are only a convention. In Physics they have the 5-sigma convention

2-sigma: 95.5 percent

3-sigma: 99.73 percent

4-sigma: 99.993 percent

5-sigma: 99.99994 percent

So that means that purely statistical fluctuations will give you a result way out in the 5-sigma range 0.00006 percent of the time.

When physicists announce that they have a 5-sigma result, that means that there's a 1 in 3.5 million chance that it was the result of a statistical fluctuation over the spectrum of experiments they performed. Particle physicists working on the CMS and ATLAS experiments are

How to interpret a p value

Yeahhhhh!! I have a p=.003, which is statistically significant (<.05)!!
This should be interpreted considering that:

a) p is the probability that the results are due to chance, the probability that the null hypothesis (H0) is true.

b) p is the probability that the results are not due to chance, the probability that the null hypothesis (H0) is false.

c) p is the probability of observing results as extreme (or more) as observed, if the null hypothesis (H0) is true.

d) p is the probability that the results would be replicated if the experiment was conducted a second time.

e) None of these.

Recap of terms

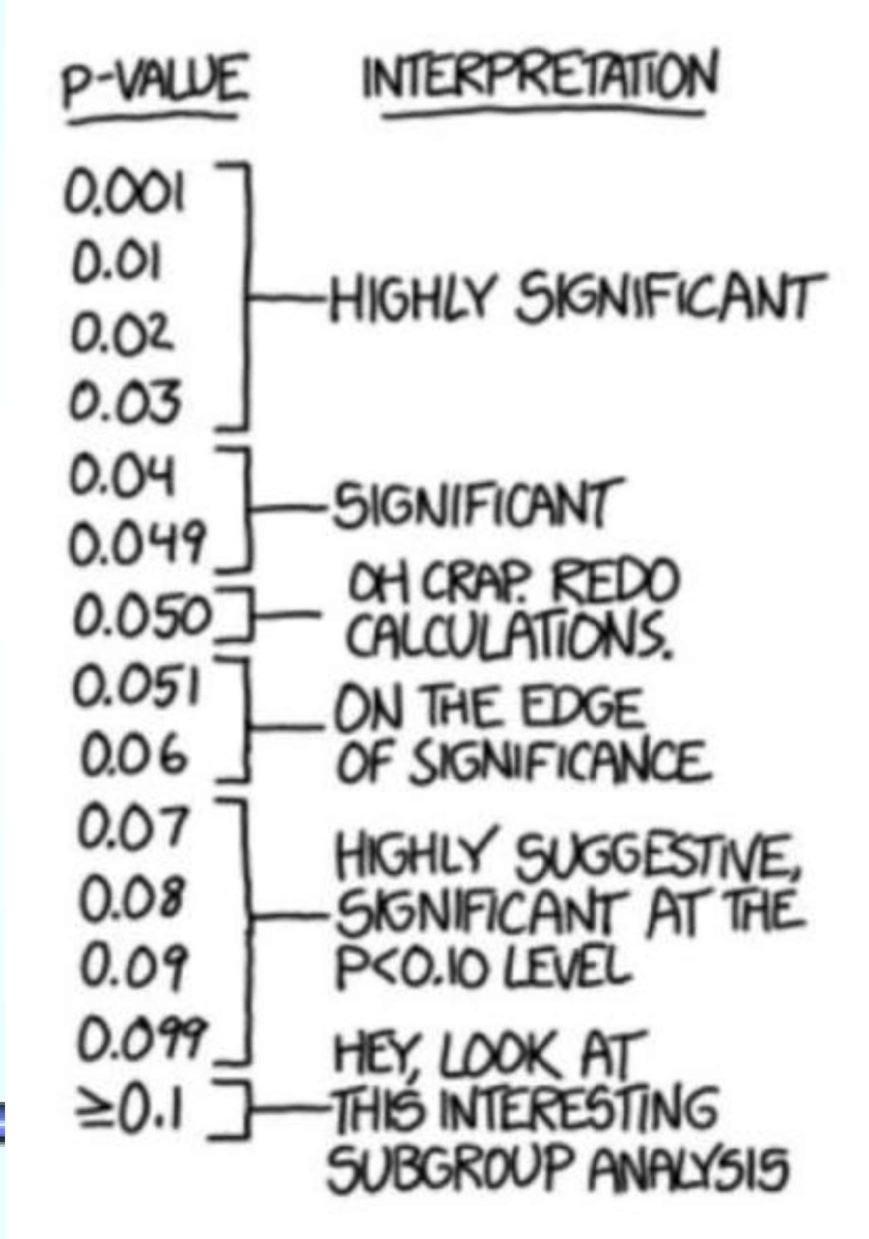
- The hypothesis which describes the effect of chance, is called the **null hypothesis** (**H0**)
- The probability of obtain our result (or even more extreme) if the null hypothesis (**H0**) is true is called **p**
- The cut-off (0.05 or 0.01) which we use to reject or not the null hypothesis is called **critical alpha**
- The operation which leads us to the decision is called **test of significance**
- If we reject the null hypothesis (**H1**), we say that our result is **statistically significant at level p (or below p)**
- If we do not reject the null hypothesis, we say that our result is **not significant**

Interpretation

Statistically significant does not mean scientifically significant, interesting or even important!

- Statistically significant means that we can exclude (with an error of **p**) that our result is equivalent to a purely random effect
- Statistically significance can often be considered a necessary condition for a result to be of interest, but it is not sufficient
- The importance of a result depends on its practical and theoretical relevance, its **strength** and direction

Alternative Interpretation



Subgroup analysis

https://xkcd.com/882/

Some suggested recent readings

Lakens, D. (2021). The practical alternative to the p value is the correctly used p value. *Perspectives on psychological science*, *16*(3), 639-648.

Wasserstein, R-L., & Lazar, N.A (2016) The ASA's statement on p-values: Context, process, and purpose. *The American Statistician*, *70*, 129-133.

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (Eds.). (2019). Statistical inference in the 21st century: A world beyond $p < 0.05$ [Special issue]. *The American Statistician, 73, 1-19.*

PSYCHOLOGICAL SCIENCE

The Practical Alternative to the *p* Value Is the Correctly Used *p* Value

Perspectives on Psychological Science
2021, Vol. 16(3) 639–648

DOI: 10.1177/1745691620958012
www.psychologicalscience.org/PPS
SAGE

Daniël Lakens
Department of Industrial Engineering and Innovation Sciences, Eindhoven University of Technology

THE AMERICAN STATISTICIAN
2016, VOL. 70, NO. 2, 129–133
http://dx.doi.org/10.1080/00031305.2016.1154108
Taylor & Francis
Taylor & Francis Group

EDITORIAL

The ASA's Statement on *p*-Values: Context, Process, and Purpose

THE AMERICAN STATISTICIAN
2019, VOL. 73, NO. S1, 1–19: Editorial
https://doi.org/10.1080/00031305.2019.1583913
Taylor & Francis
Taylor & Francis Group

OPEN ACCESS Check for updates

EDITORIAL

Moving to a World Beyond "$p < 0.05$"

Effect sizes

But literally H0 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)

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

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

$$r = \sqrt{\frac{\chi^2(1)}{N}} \qquad r = \sqrt{\frac{t^2}{t^2 + df}} \qquad r = \sqrt{\frac{F(1,-)}{F(1,-) + df_R}}$$

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

UNIVERSITA' DEGLI STUDI DI MILANO BICOCCA

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_{pool} =1.25

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

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

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

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 d family)		Measures of association (the r 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	r	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		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
d	Cohen's d: the uncorrected standardized mean difference between two groups based on the pooled standard deviation		
Δ	Glass's delta (or d): 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 2×2 contingency table
g	Hedges' g: the corrected standardized mean difference between two groups based on the pooled, weighted standard deviation	C	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	V	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 d family)	Measures of association (the r 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
	$_{adj}R^2$	Adjusted R squared, or the coefficient of multiple determination adjusted for sample size and the number of predictor variables
	f	Cohen's f: quantifies the dispersion of means in three or more groups; commonly used in ANOVA
	f^2	Cohen's f 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

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

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

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

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

➢ Effect sizes go up when "**signal**" (*numerator*) increases relative to "**noise**" (*denominator*)

Effect size: useful tools

Read these

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

Evaluating Effect Size in Psychological Research: Sense and Nonsense

Advances in Methods and Practices in Psychological Science
2019, Vol. 2(2) 156–168
© The Author(s) 2019
Article reuse guidelines: sagepub.com/journals-permissions
DOI: 10.1177/2515245919847202
www.psychologicalscience.org/AMPPS
SAGE

David C. Funder and Daniel J. Ozer
Department of Psychology, University of California, Riverside

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)

Some 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 (ρ < .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

Hypothesis testing II

Testing for Significance

In practice, each test of significance requires:

- We should know what is the **null hypothesis**
- The variability of the parameter we are testing: standard error
- The conventional cut-off is 0.05 (or 0.01)

And software produces for us:

- The calculation of the test
- The probability of our result in the population described by the null hypothesis (**p**)

Significance of Correlation Coefficient

- Does the correlation coefficient differ from zero?

Correlation Matrix

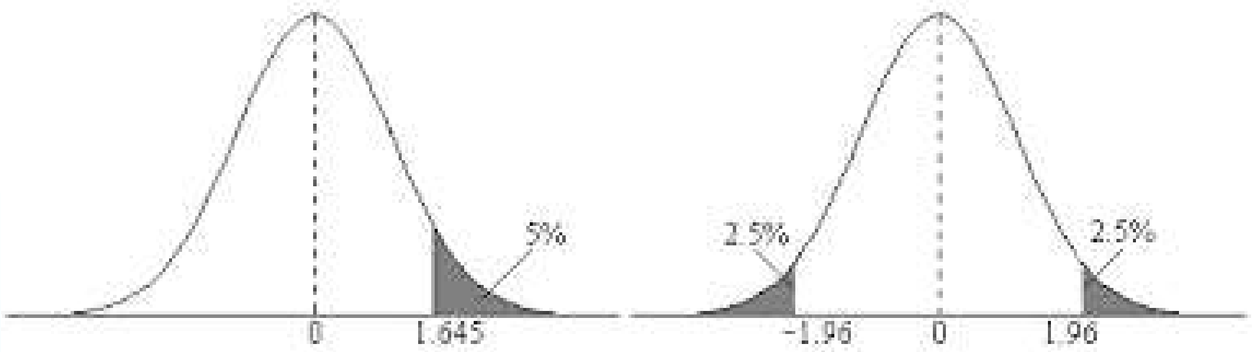
		v1	v19
v1	Pearson's r	—	
	p-value	—	
	95% CI Upper	—	
	95% CI Lower	—	
	N	—	
v19	Pearson's r	0.370***	—
	p-value	< .001	—
	95% CI Upper	0.459	—
	95% CI Lower	0.273	—
	N	334	—

Note. * p < .05, ** p < .01, *** p < .001

Because p<.05, we can reject the null hypothesis that r is equal to zero

One-tailed vs two-tailed t-test

- One-tailed test: we test that r=0 against r>0
- Two-tailed test: we test that r=0 against r≠0



(a) One-tailed test (b) Two-tailed test

Hypothesis testing among groups

- Theoretical hypothesis: Frustration increases aggression
- Empirical test: we need a situation that produces frustration (**independent variable, IV**) and a variable that measures aggression (**dependent variable, DV**)
- **IV**: Experimental manipulation. Ss are given a memory test and told that they have done poorly (**frustration**) or well (**control**)
- **DV**: Ss are asked to evaluate the confederate for suitability to an assistant research post

Defining and testing hypotheses

- **H1**: Frustration increases aggression
- **H0**: Frustration does not affect aggression
- To test it, we analyze sources of variation in the scores (DV)
- Two sources of variation

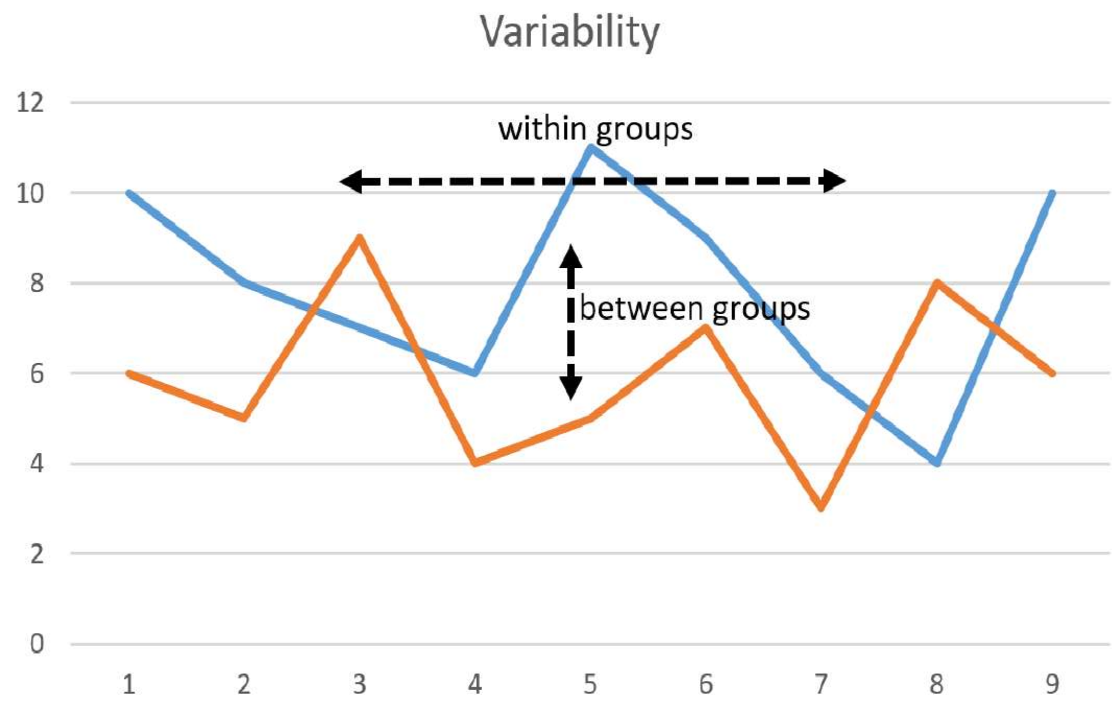
  a) due to the experimental effect (systematic variance, effect variance, variation **between** groups)

  b) due to other sources (non-systematic random variance, error variance, variation **within** groups)
- We need a test statistics with a known probability distribution

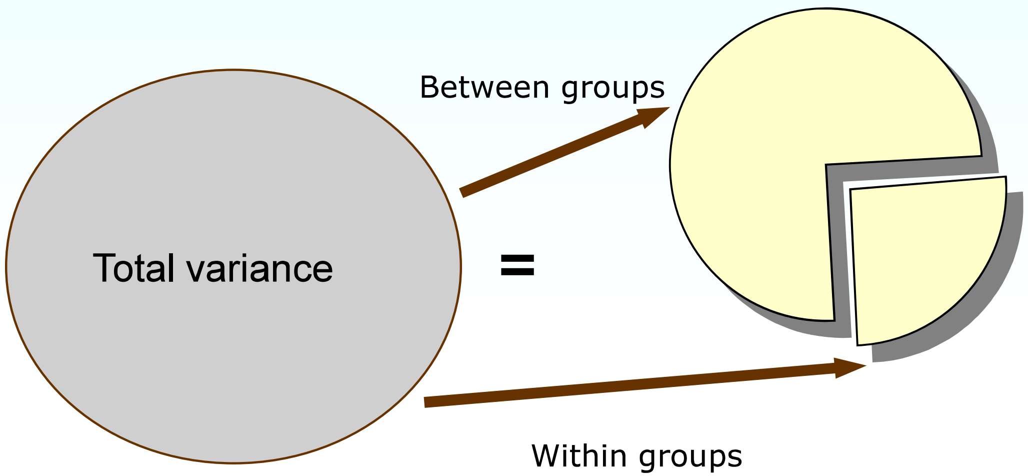
$$\textit{test statistic} = \frac{\textit{variance explained by the effect}}{\textit{error variance}} = \frac{\textit{variance between groups}}{\textit{variance within groups}}$$

The two sources of variance

Group A	Group B
10	6
8	5
7	9
6	4
11	5
9	7
6	3
4	8
10	6
7.89	**5.89**



Partitioning of variance



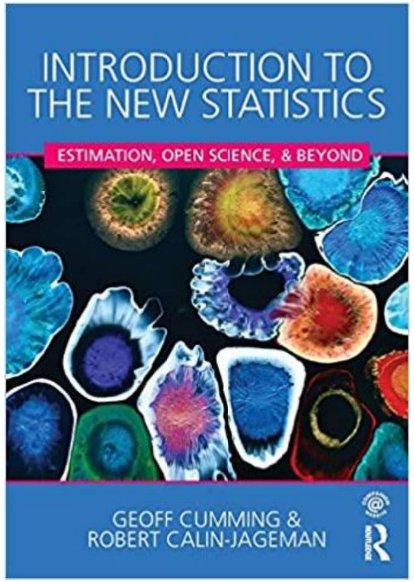
$$\textit{test statistic} = \frac{\textit{variance explained by the effect}}{\textit{error variance}} = \frac{\textit{variance between groups}}{\textit{variance within groups}}$$

The greater this value, the better; its significance is evaluated against a probability distribution of the the test statistic considering a certain threshold (e.g., p=.05)

Errors of statistical inference

Errors of inference

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



NHST= Null Hypothesis Significance Testing
(what you have been taught as a student)
H0 vs. H1

Errors of inference in NHST

Conclusion of the significance test (**SAMPLE**)	Real World (**POPULATION**): Null is true (H0 is correct)	Real World (**POPULATION**): Null is false (H1 is correct)
Null is true	Correct decision ($1-\alpha$)	Type II error (β)
Null is false	Type I error (α)	Correct decision ($1-\beta$)

UNIVERSITA' DEGLI STUDI DI MILANO BICOCCA

Errors of inference in NHST

- **Type I error**: ***Erroneously rejecting the null hypothesis (<u>False positive</u>)***.
  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 (<u>False negative</u>)***. 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 that one uses in a significance test (*p value threshold*).

- 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.

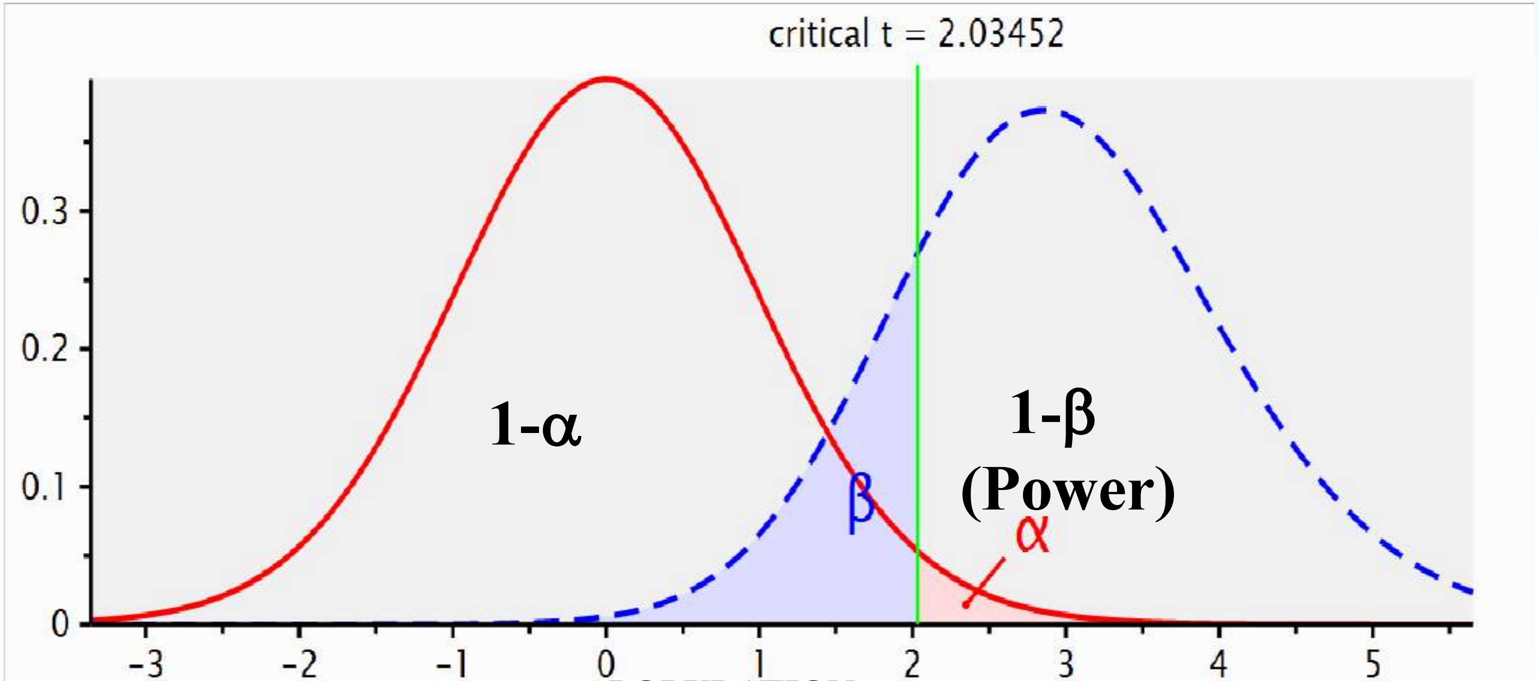
- The test can be two-tailed (more common) or one-tailed (directional)

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 – β**
- Conventionally a power of at least .80 (**β=.20)** is considered as acceptable

Power Analysis

What is power?



POPULATION

SAMPLE	Null is true (H0 is correct)	Null is false (H1 is correct)
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

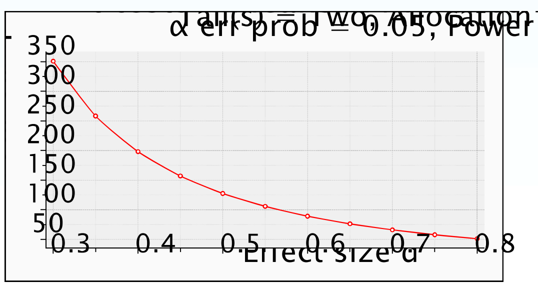
- Power is determined by four elements

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 α=.05 and (1- β)=.80
- Plot sample size and effect size for a two sample t-test



UNIVERSITA' DEGLI STUDI DI MILANO BICOCCA

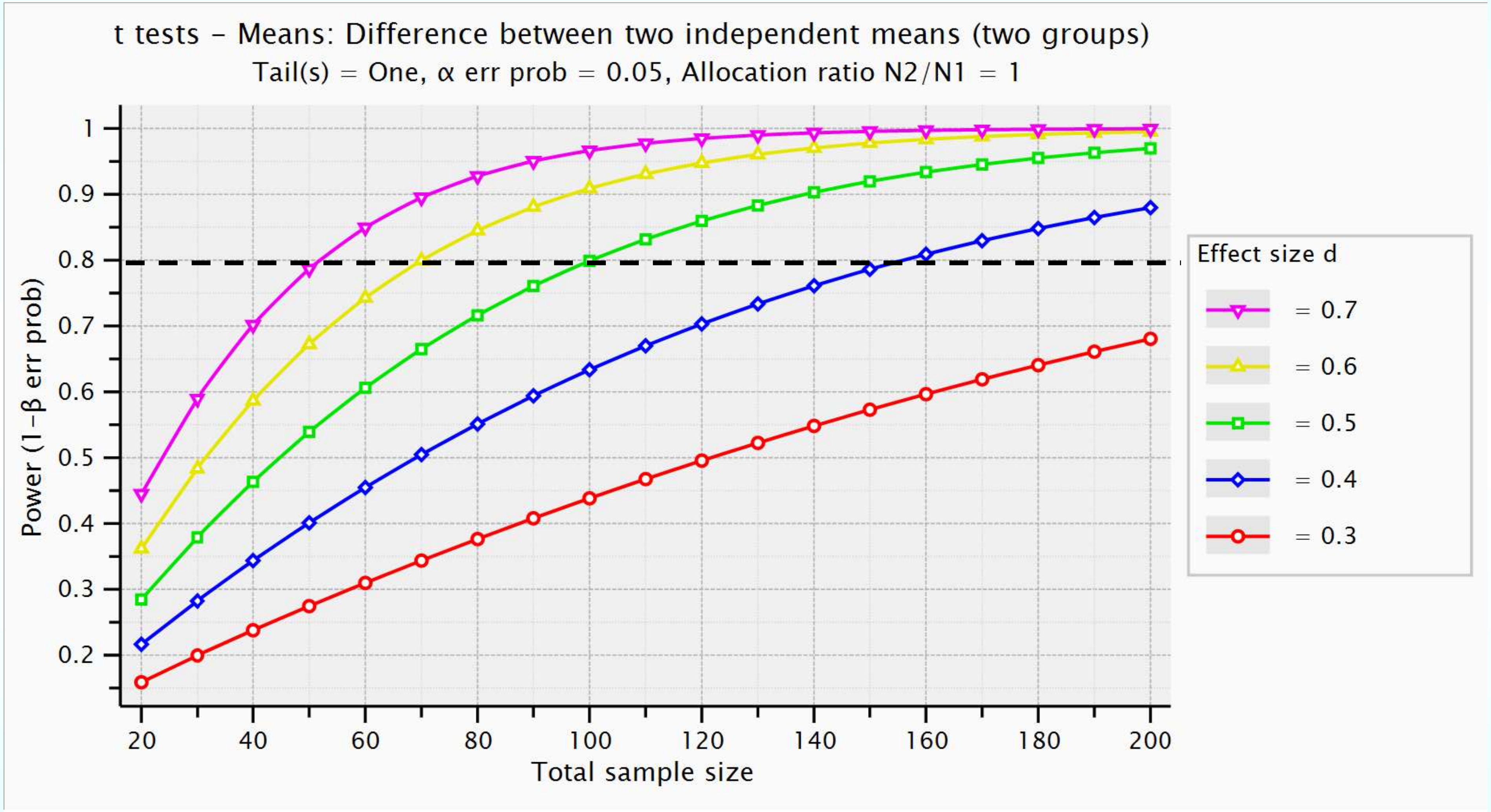
What affects power?

- Power goes up with larger effect sizes and sample sizes, given a certain decision criterion (e.g., α=.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

$$d = \frac{M_1 - M_2}{pooled\ SD} \qquad \hat{\eta}^2 = \frac{SS_{Effect}}{SS_T}, \qquad r(v,x) = \frac{\mathrm{cov}(v,x)}{sd(v) * sd(x)}$$

Power as a function of ES and N



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***

Is low power a real problem for Psychology?

PLOS | BIOLOGY

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. Ioannidis[2]

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-

Psychological Science
2016, Vol. 27(8) 1069–1077

Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0956797616647519
pss.sagepub.com
SAGE

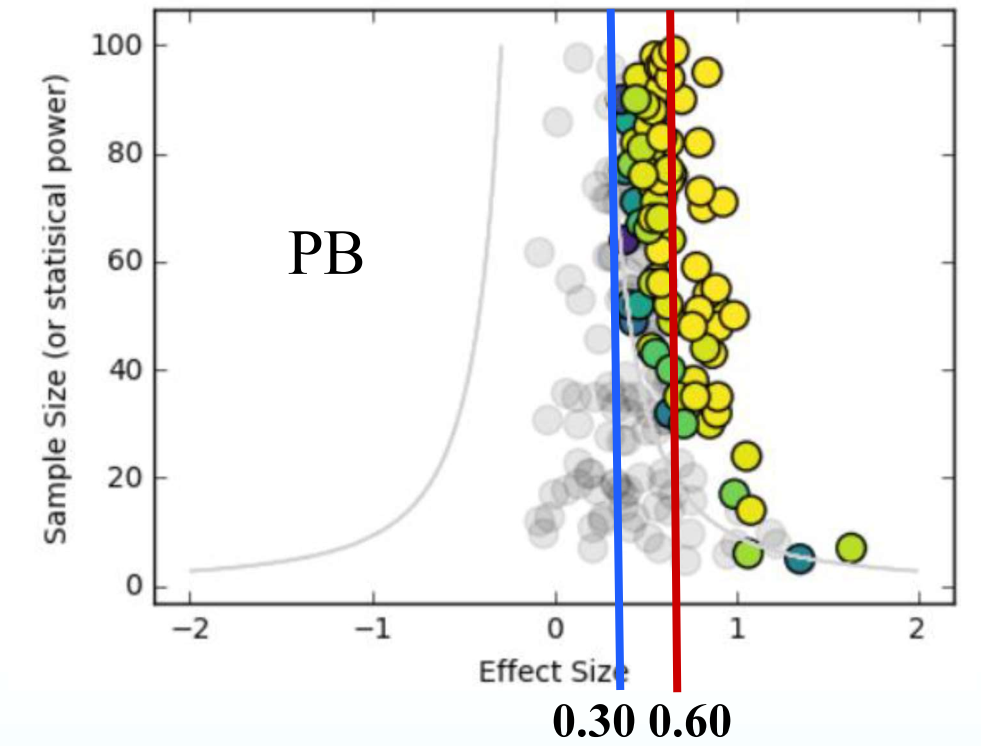
Researchers' Intuitions About Power in Psychological Research

Marjan Bakker[1], Chris H. J. Hartgerink[1], Jelte M. Wicherts[1], and Han L. J. van der Maas[2]
[1]Department of Methodology and Statistics, Tilburg School of Social and Behavioral Sciences, Tilburg University, and [2]Department 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-

Yes!

Publication bias and low power



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"**

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, 2008 Ann Rew Psych)
- 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

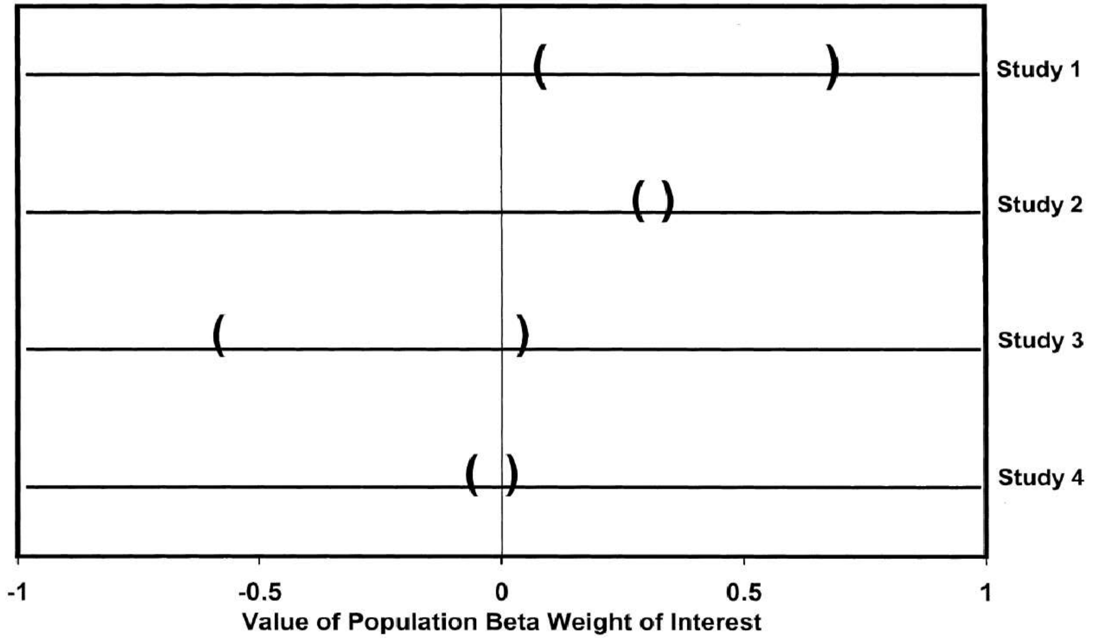
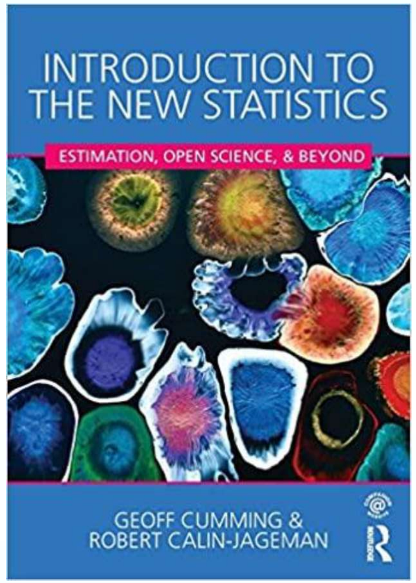


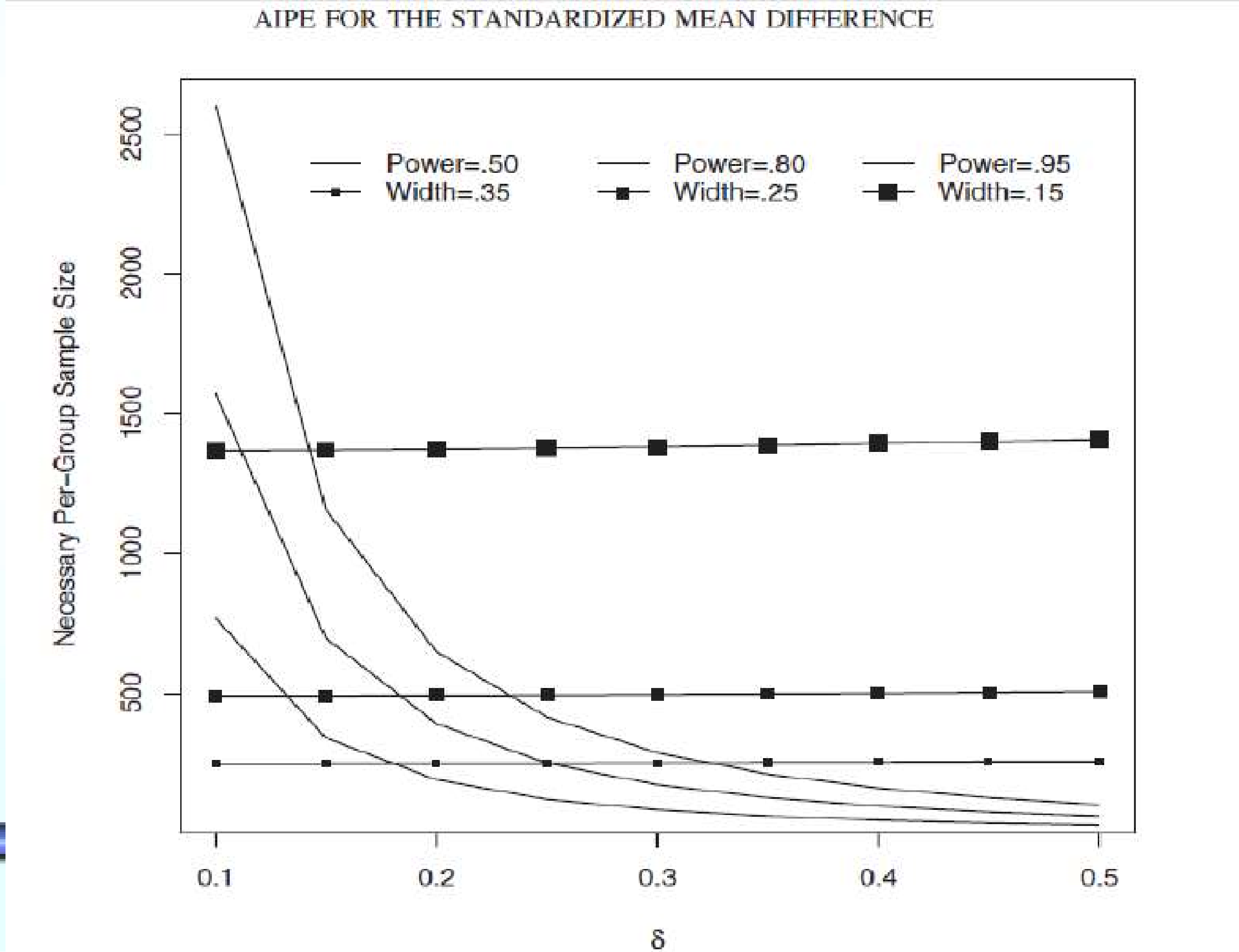
Figure 1. Illustration of possible scenarios in which planned sample size was considered a "success" or "failure" according to the accuracy in parameter estimation and the power analysis frameworks. Parentheses are used to indicate the width of the confidence interval.



- Precision is valuable no matter everything else

a MINOR practical problem…

- **<u>Big</u>** 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, χ2 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.

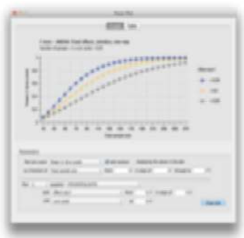
Kontrast

Suchbegriff

Screenshots (click to enlarge)

Main Window

Main Window (Table)



Power Plot

Power Plot (Table)

Power analysis calculations

- Examples of calculation of power analysis for some simple designs
- Also based on

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

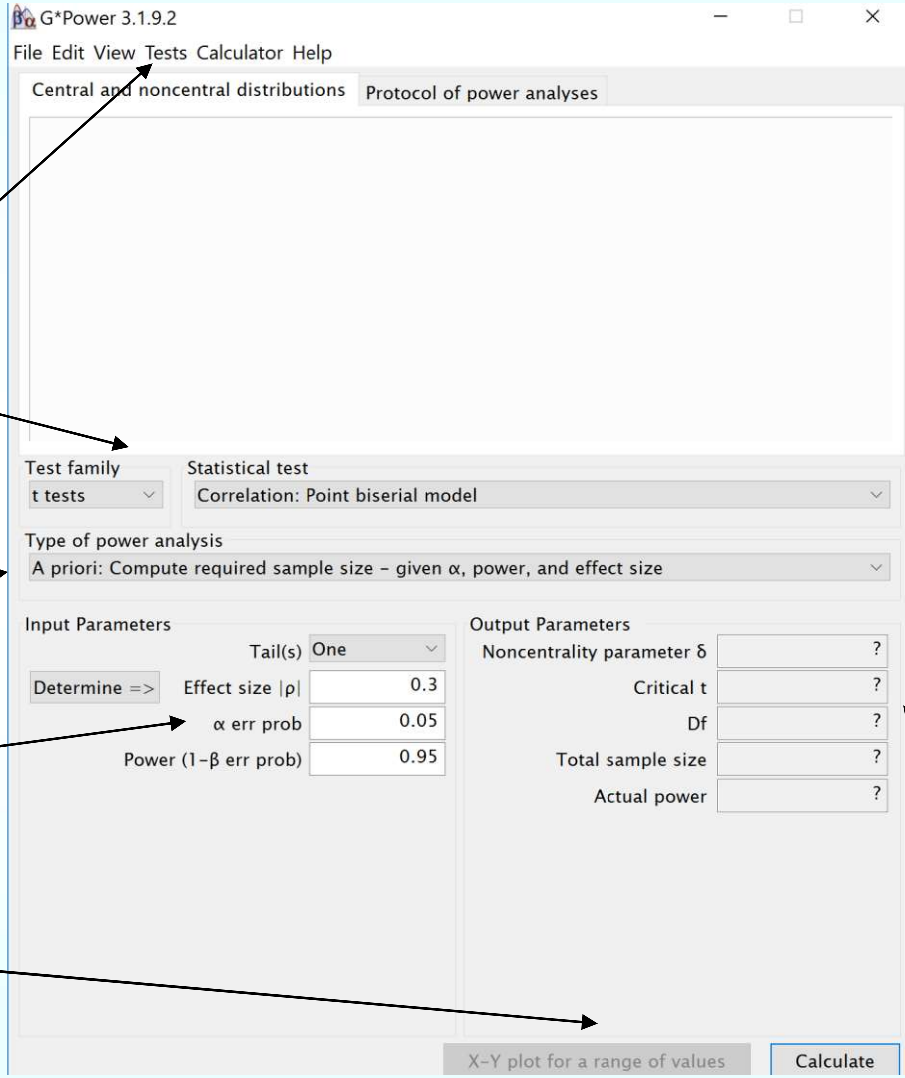
RESEARCH ARTICLE

A Practical Primer To Power Analysis for Simple Experimental Designs

Marco Perugini, Marcello Gallucci and Giulio Costantini

SM and routines available at https://github.com/mcfanda/primerPowerIRSP

G*Power



Analysis

Type of power analysis

Inputs

Action buttons

Output

Example: Two independent groups

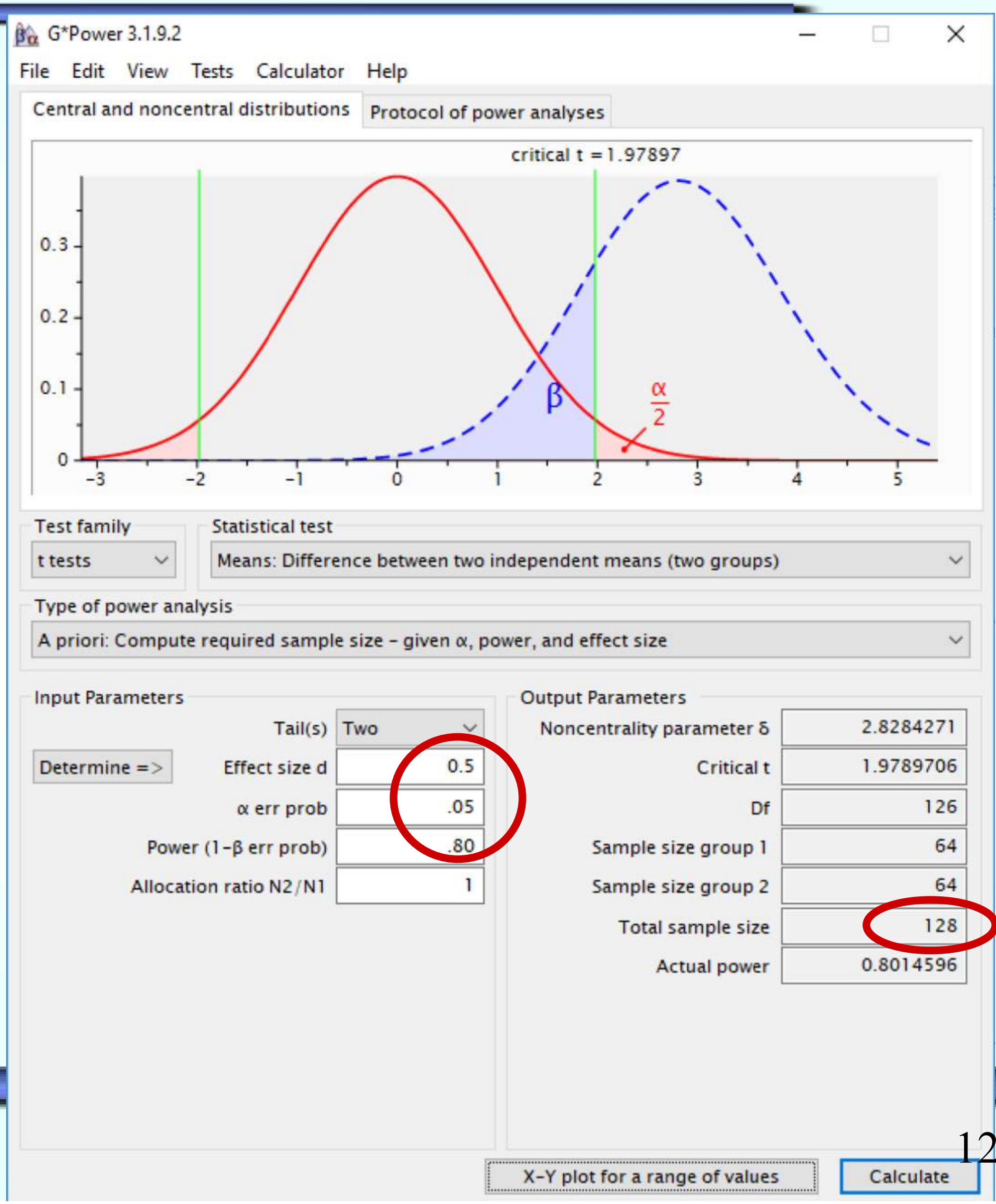
Standard pre-study planning approach

Fix ES, α, 1-β

Calculate needed N

Good practice

We consider also sensitivity analysis



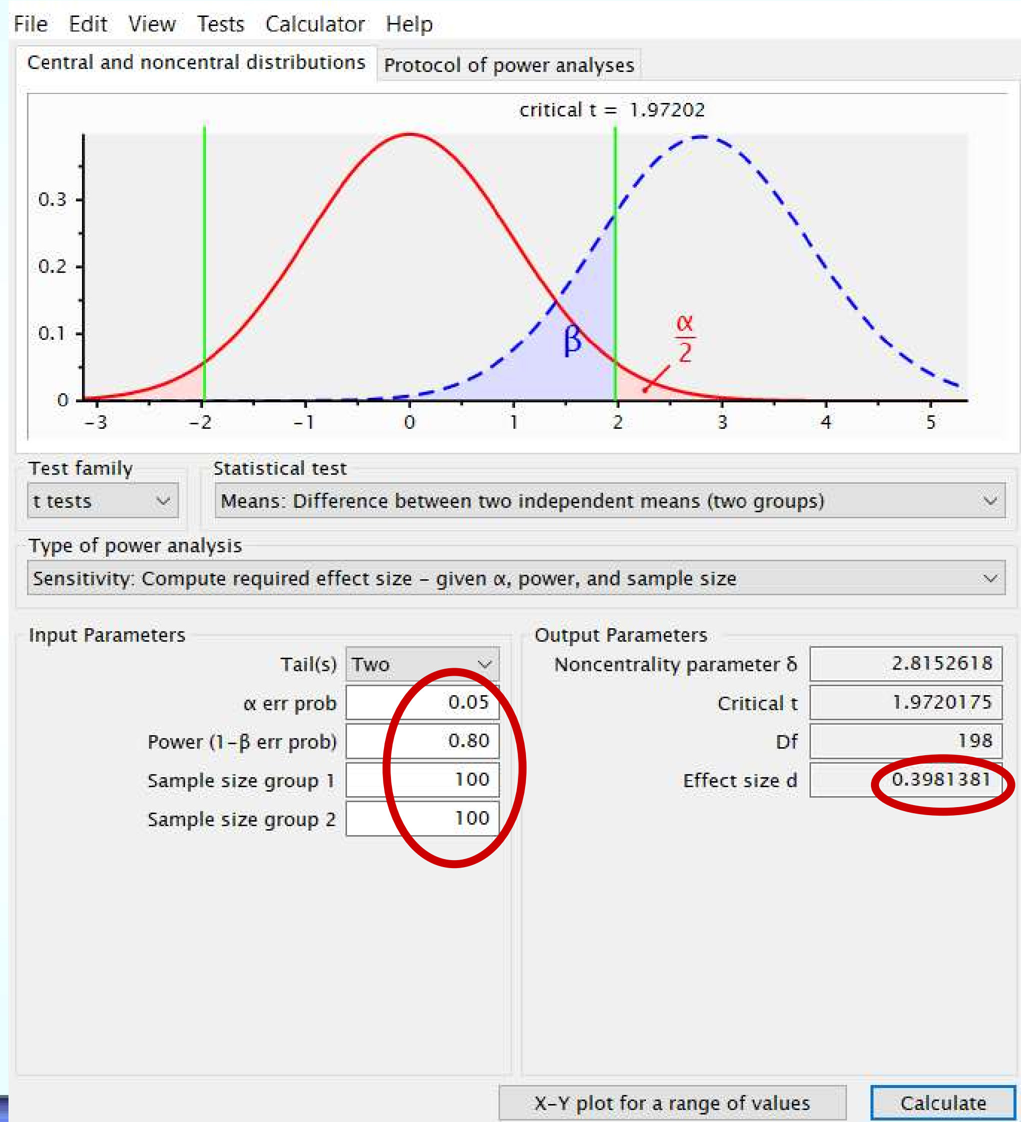
Sensitivity analysis: Starting from N

"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

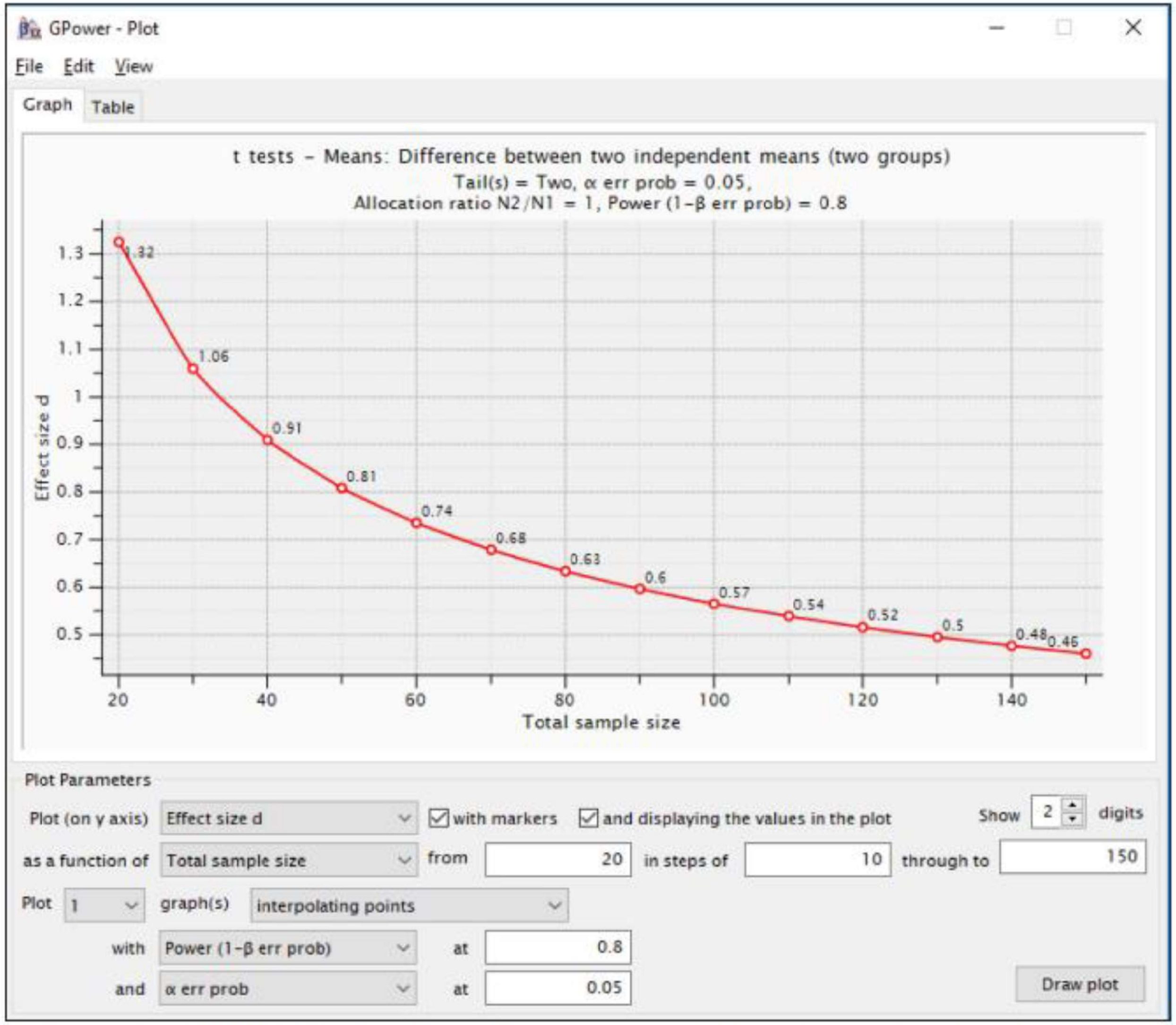


Figure 2: Sensitivity Plot of G*Power calculating the power of a two independent samples t-test: Lowest detectable effect size as a function of required *N*.

Sensitivity plot: N by Power

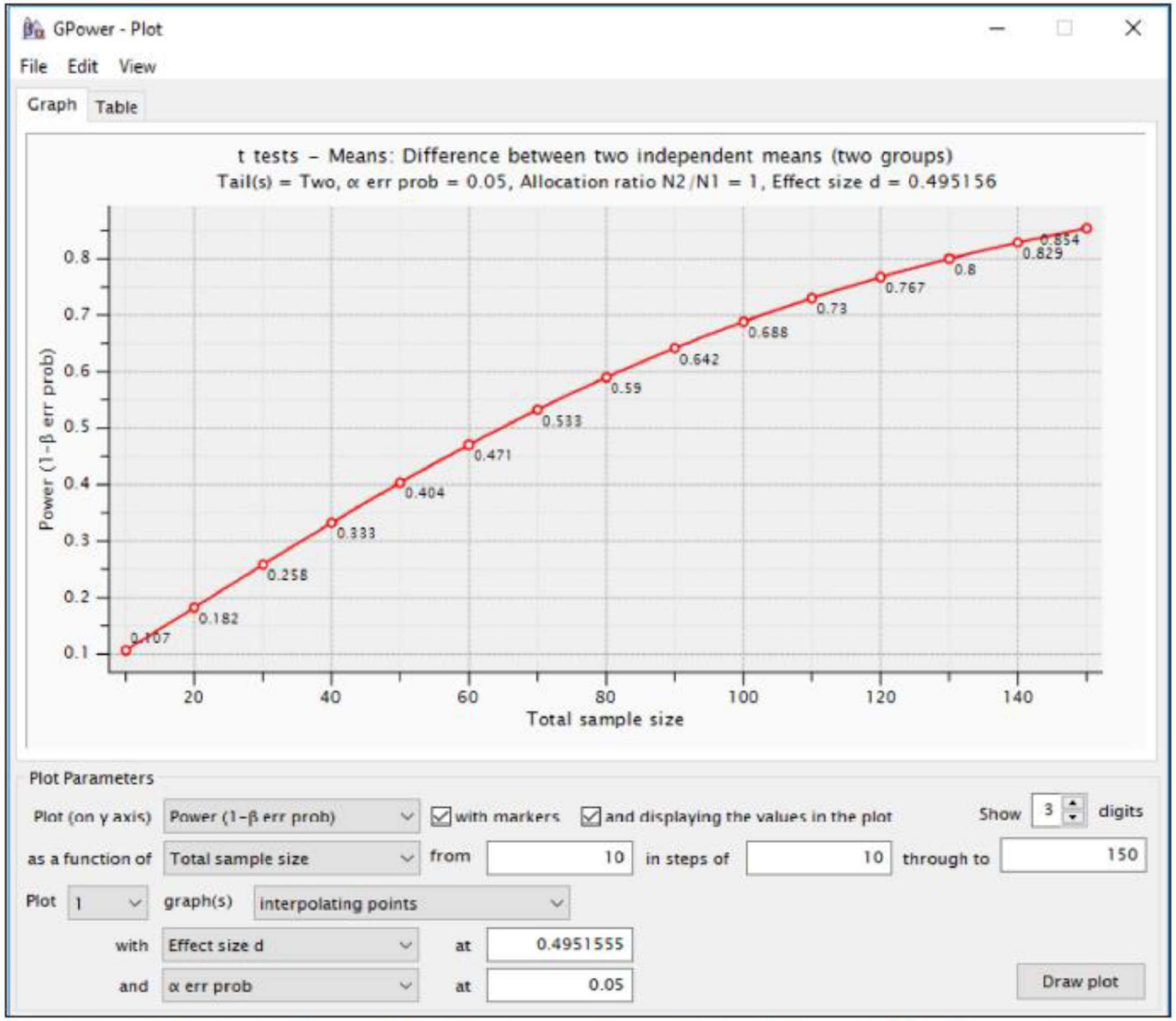
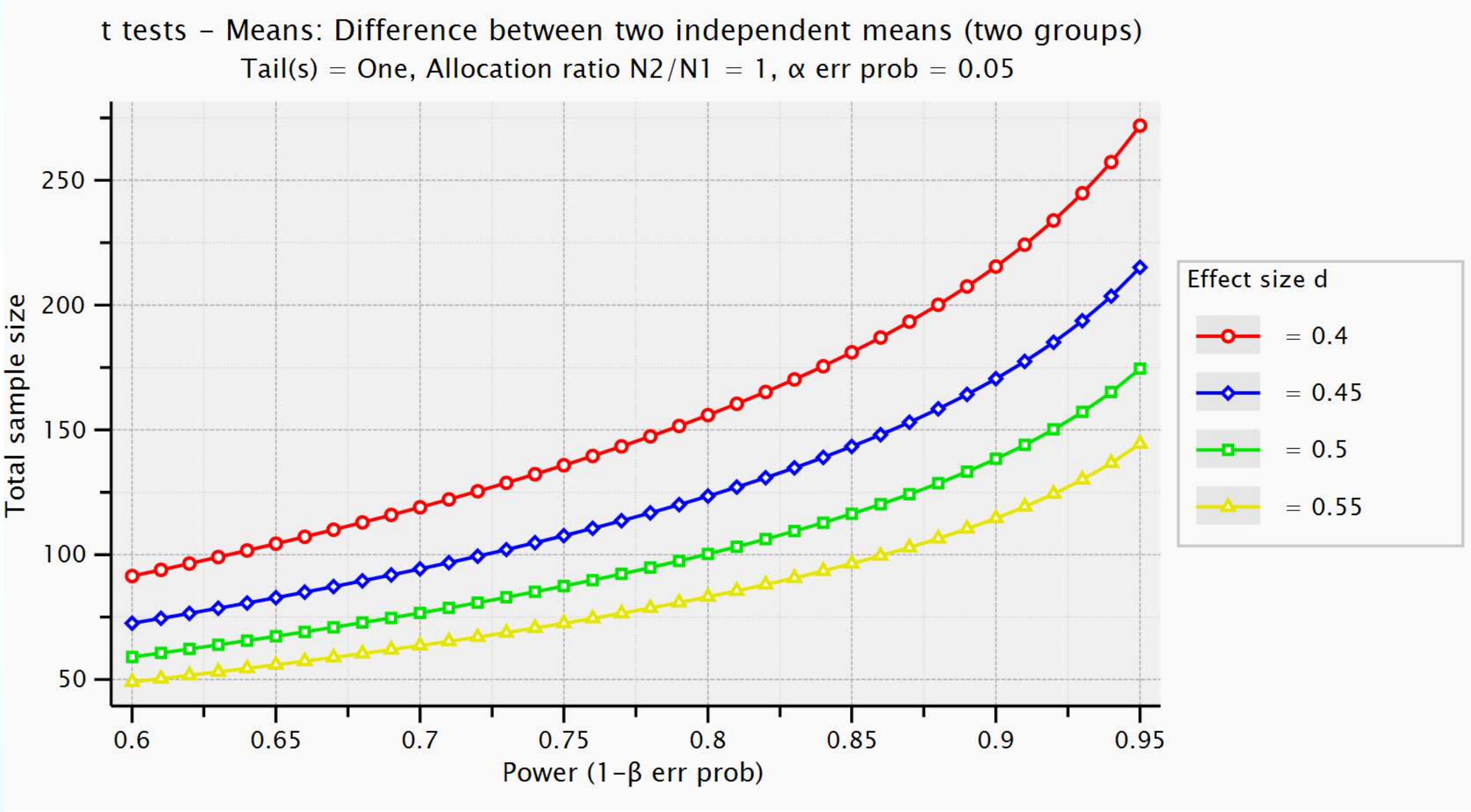
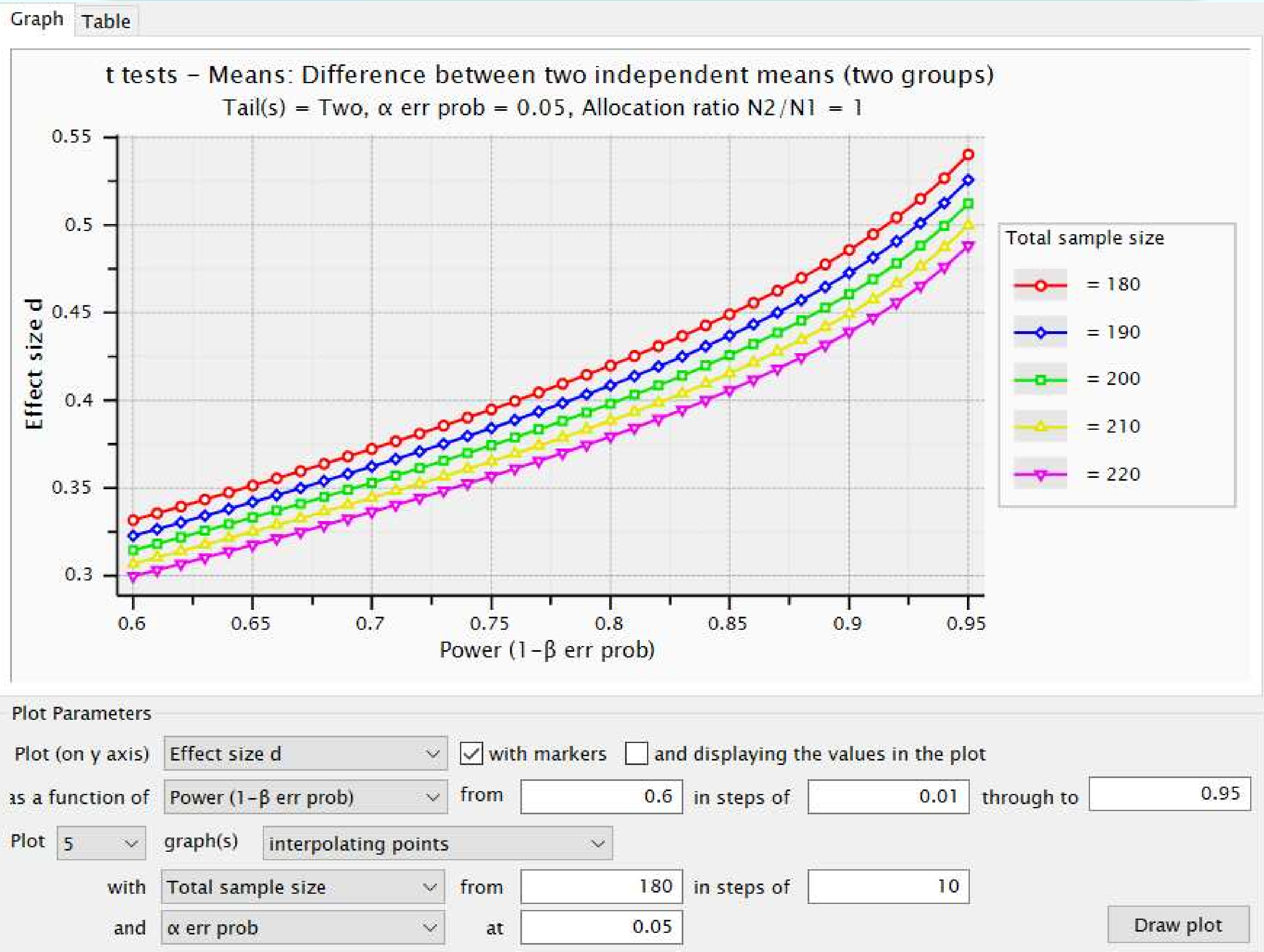


Figure 3: Sensitivity Plot of G*Power calculating the power of a two independent samples t-test: Power as a function of required N for fixed effect size.

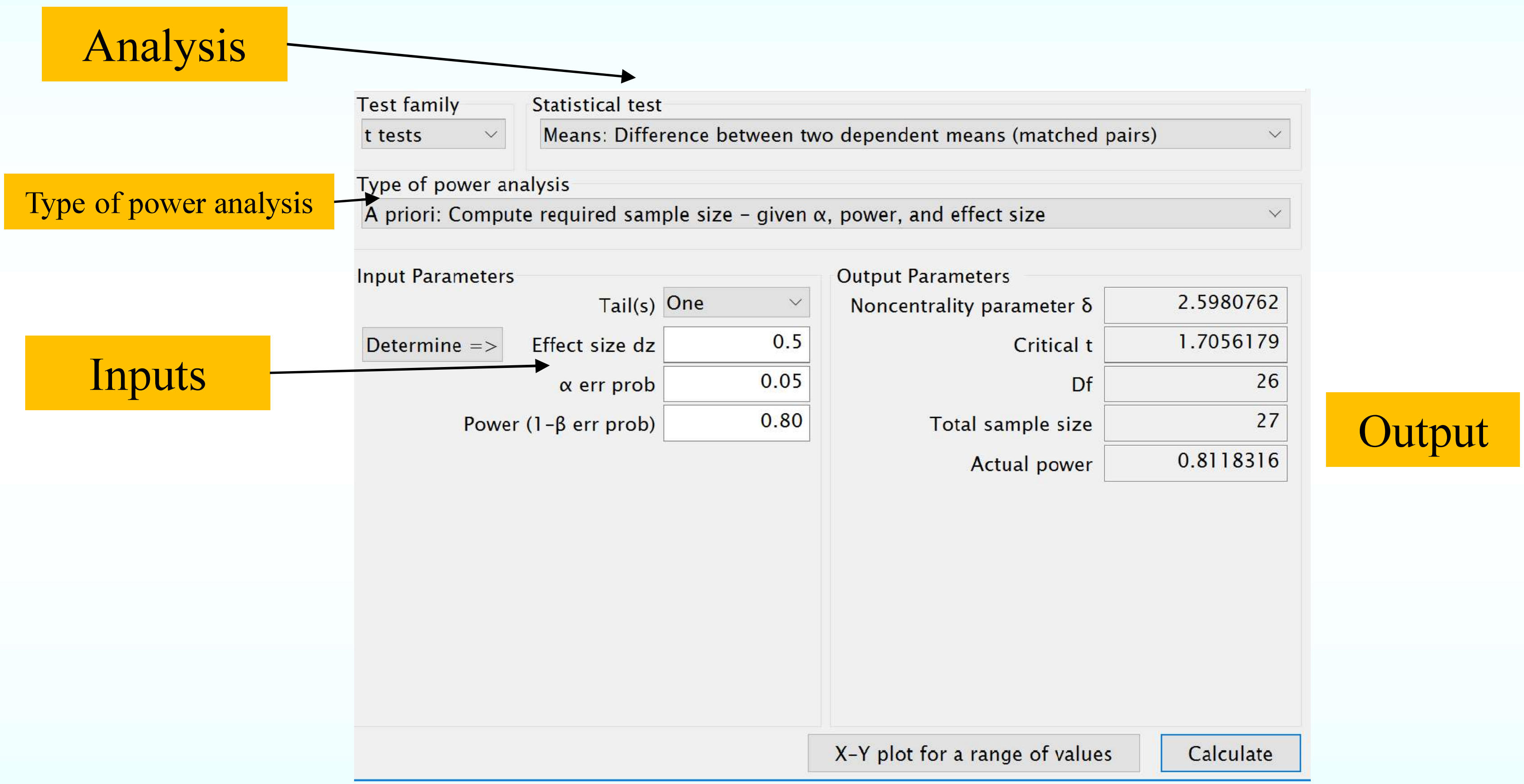
Inspecting scenarios around effect sizes



Inspecting scenarios around N



Two repeated measures



Two repeated measures

The ES for a paired means design is $d_z = \Delta / sd$.

This is not the same as Cohen's d, but calculated with the difference score divided by its SD. To double check with available previous results $d_z = \frac{t}{\sqrt{N}}$

If no previous results, could guess ES as if the two are independent groups and how much the measures are correlated (r/ρ)

$\boldsymbol{d_z} = \frac{\boldsymbol{d}}{\sqrt{\mathbf{2}(\mathbf{1}-\boldsymbol{r})}}$, e.g., with d=0.5 and r=.55, $d_z = \frac{0.5}{\sqrt{2(1-0.55)}} = 0.527$

and the other way round

$\mathbf{d} = \boldsymbol{d_z} * \sqrt{\mathbf{2}(\mathbf{1}-\boldsymbol{\rho})}$, e.g. with d_z=0.527 and r=.55,

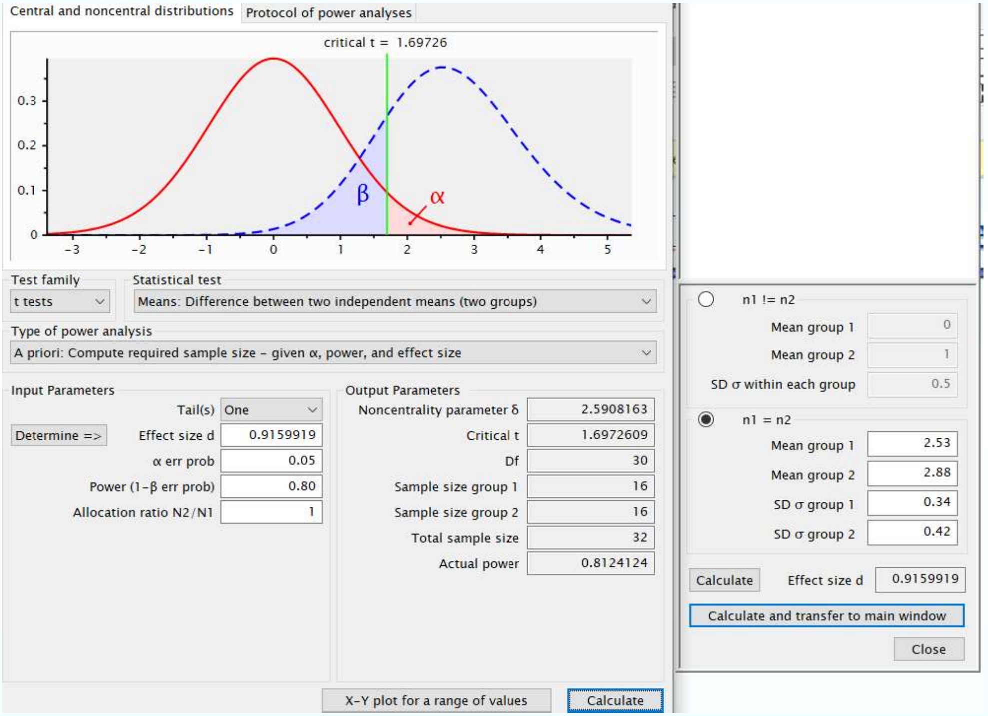
$$\mathrm{d} = 0.527 * \sqrt{2(1-.55)} = 0.527 * 0.95 = 0.50$$

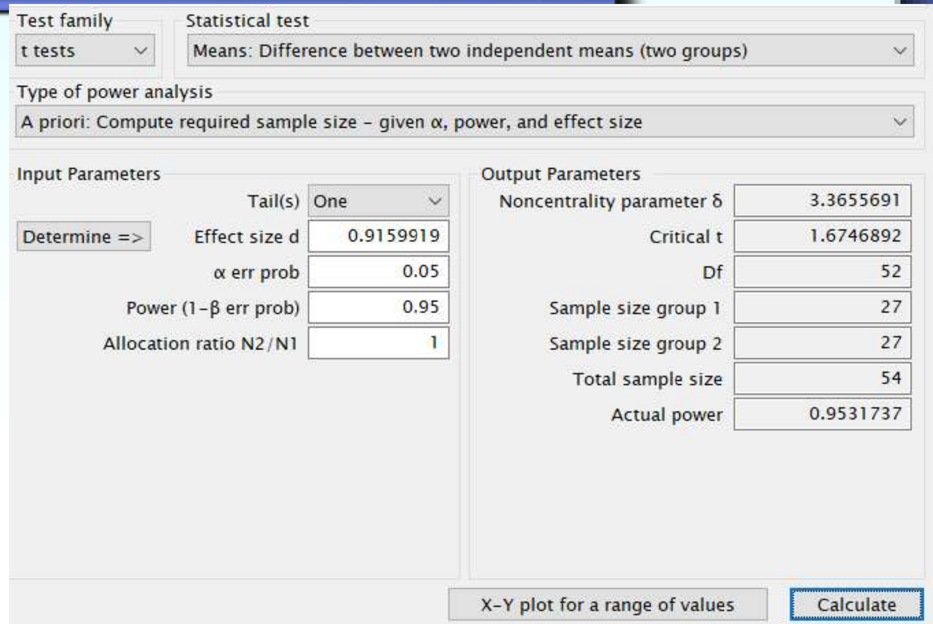
Let's see it together !

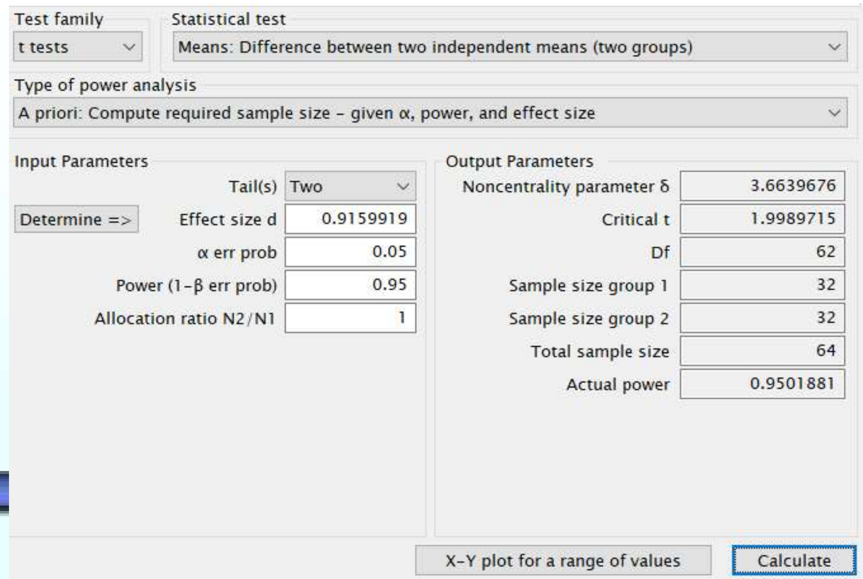
1) We want to run a study to replicate a previous research. They had two groups and found these results:
   N1=40, M=2.53, SD=0.34
   N2=40, M=2.88, SD=0.42.
   We want to estimate N with power at 80% and at 95% (α=.05, one-tail, two-tails). **N?**
2) We expect a subtle but theoretically important effect (*d*=0.15). How many N with power at 90% or 80% (α=.05, one-tail, 2 groups). **N?**
3) We can run a study with about 120 Ss. What ES can we detect at power 80% (α=.05, two-tails) for two-groups between design? **ES?**

Let’s see it together !

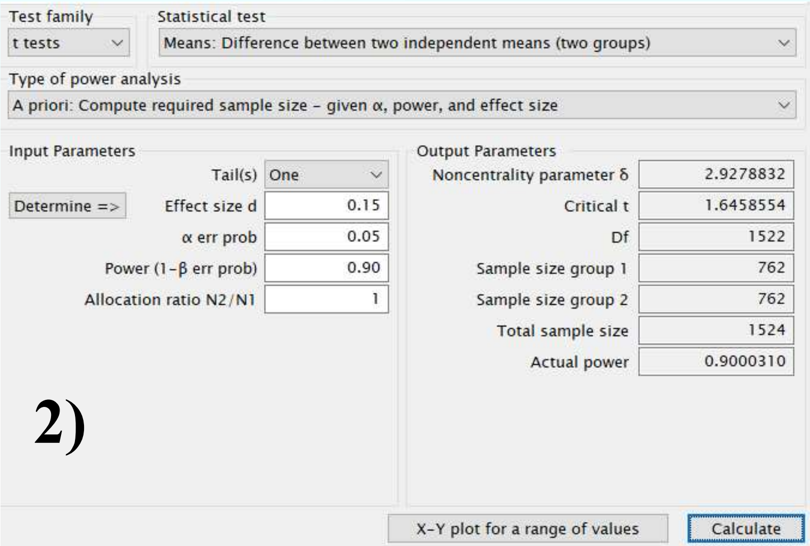
1)



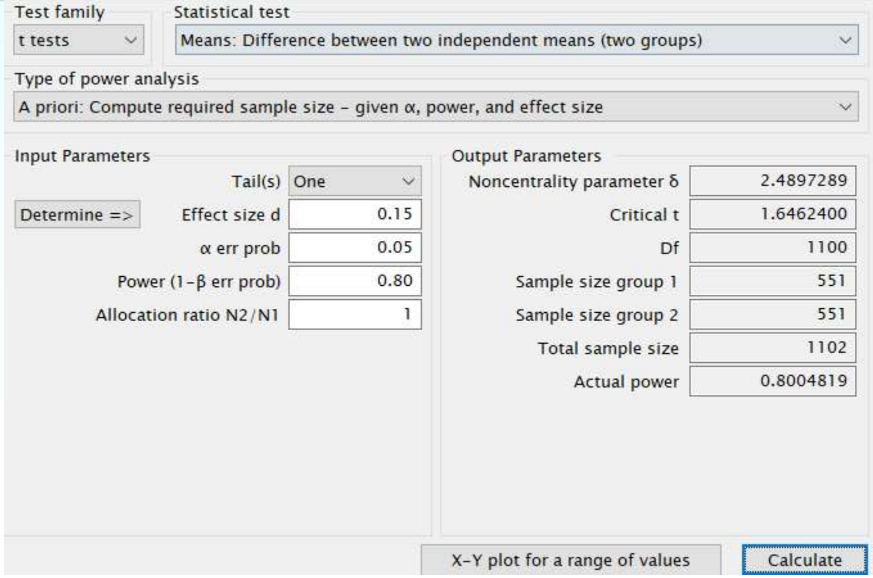




Let’s see it together !



2)



3)