Biostatistics BootCamp Lecture 3:

Understanding of the Statistical Power and Sample Size in Protocol Development

August 23, 2023

Wonsuk Yoo, PhD

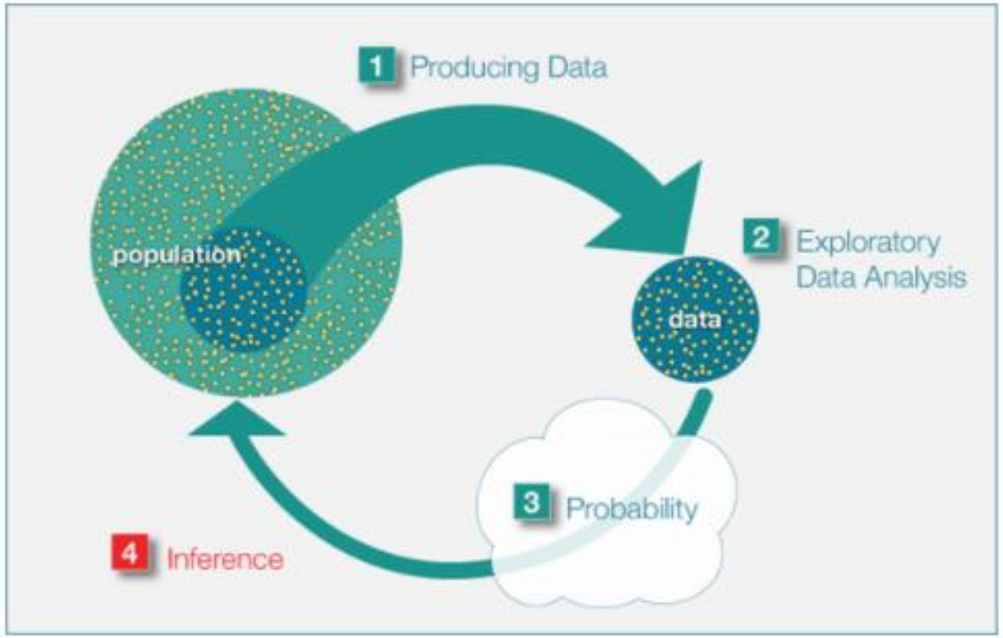
Dept. of Translational Neuroscience

wonsuk.yoo@barrowneuro.org

Ivy Brain Tumor Center | Barrow Neurological Institute

Lecture 1. Uncertainty &Hypothesis Tests:

- An ideal approach for a scientific research is to investigate the total population.
  - ➡ Since it is not usually feasible, we do not have population characteristics.
  - ➡ The need for Statistics.
- Statistics uses sample data to determine population parameters. ➡ Statistical inference and uncertainty.
- We need to perform the hypothesis testing procedure to quantify the uncertainty.



Open learning Initiative at Carnegie Mellon University, https://oli.cmu.edu/courses/concepts-of-statistics/

- A good design through hypothesis tests is related to the statistical uncertainty, the two types of errors, α and β.

Lecture 2. Control of α up to 5%

- When do we encounter the multiplicity?

> When multiplicity is present, the usual statistical approach may necessitate an adjustment to the type 1 error. Multiplicity may arise, for example, from multiple primary variables, multiple comparisons of treatments, repeated evaluation over time, and/or interim analyses – ICH E9

- The multiplicity (multiple tests) leads to the inflation of type 1 error rate.

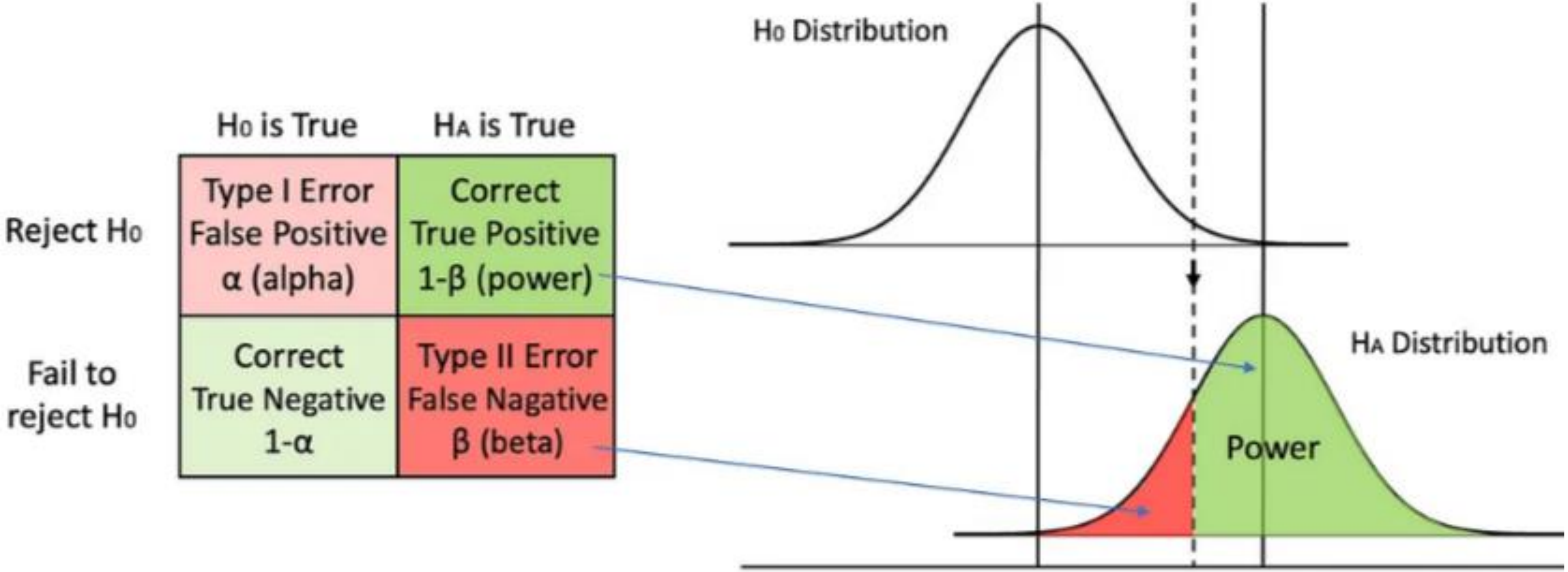
  ➡ Inflated type 1 error (α): α = 0.098 for k=2, α = 0.19 for k=4.

- How can we overcome the multiplicity?

  ➡ Bonferroni method: $\boldsymbol{\alpha}/k$

  ➡ Other methods like Holm method, Hochberg method : $\boldsymbol{\alpha}/k$

Power 1: What’s Power? graphical understanding



	H_0 is True	H_A is True
Reject H_0	Type I Error False Positive α (alpha)	Correct True Positive 1-β (power)
Fail to reject H_0	Correct True Negative 1-α	Type II Error False Nagative β (beta)

Source: Statistical power from Towards data science (AY Yao)

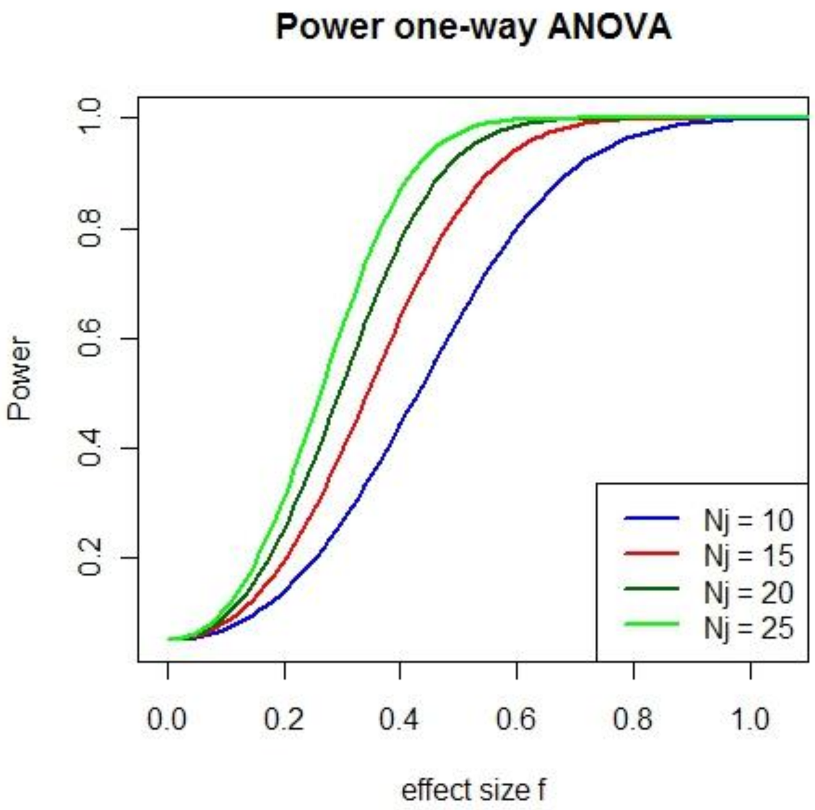
Power 2: what does it mean?

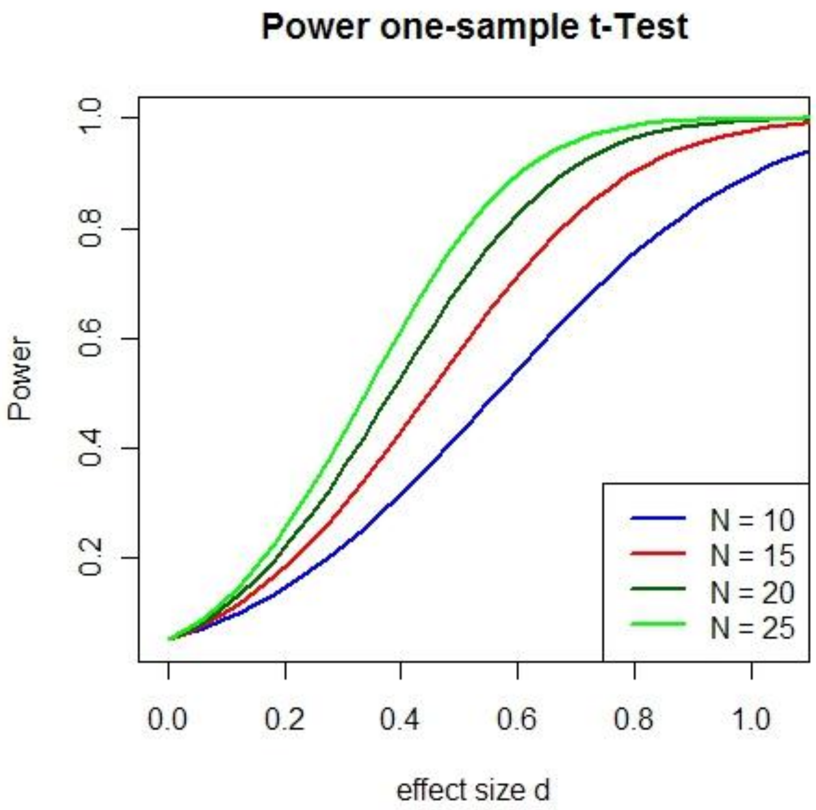
- In hypothesis tests, the H_0 always means that no scientific findings exist and the H_1 says that the study group shows better over the control group.
  - ➡ The investigators/researchers typically hope to prove the alternative hypothesis. That is, they hope to reject the null hypothesis (with p<0.05).
- The power of a test tells us how likely we are to find a significant difference given that the H_1 is true ➡ The true mean (μ) is different from μ_0.
- If the power is too low, then we have little chance of finding a significant difference even if the true mean is not equal to μ_0 ➡ The detection power
- During the planning stage, we need how many subjects are needed in order to have a desired power of detecting a clinical meaningful difference.

Power 3: Elements affecting the power

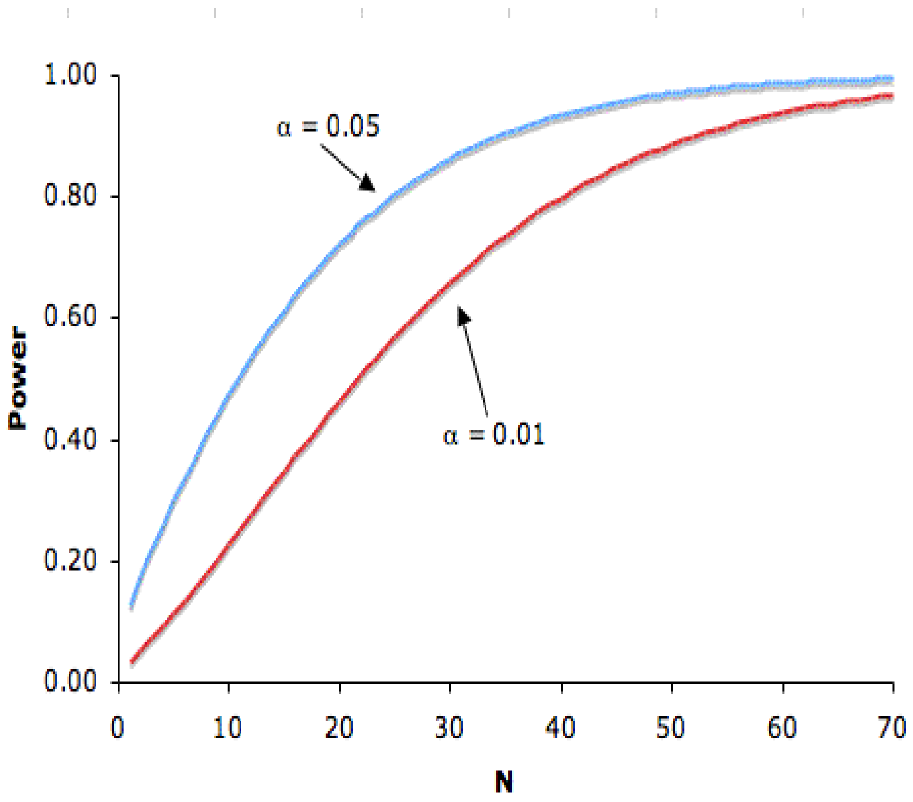
- The sample size (N): As N increases, the power increases.
- The effect size (ES): As the ES increases, the power also increases with a fixed sample size ➡ Clinically/scientifically meaningful mean difference between control and treatment groups.
- Two types of errors (α and β):
  ➡ As β increases, the power decreases (power = $1 - \beta$)
- Precision of the measurements:
  ➡ As the standard deviation of the measurement decreases, the power increases.

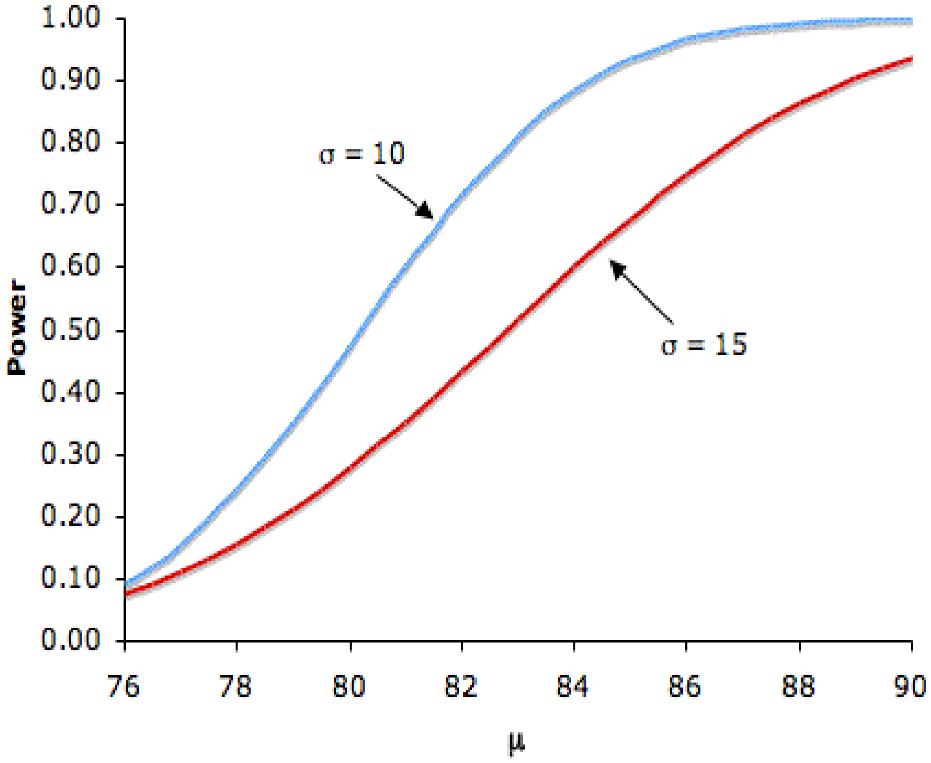
Sample Size (N) & ES .vs. Power (1- β)





Error (α) and Precision (σ) .vs. Power (1- β)





Sample size: General formula

$$n_i = 2\left(\frac{Z_{1-\alpha/2} + Z_{1-\beta}}{ES}\right)^2$$

- This is a sample size formula for two independent samples with continuous outcome where n_i is the sample size for each group, α and β are two types of errors, and ES is the effect size.
- Hypothesis: $H_0(\mu_1 = \mu_1)$ versus $H_a(\mu_1 \neq \mu_1)$
- As the sample size increases, the power increases.

Effect size 1: What it is?

- The “**effect size (ES)**" means the degree to which the phenomenon is present in the population
  ➡ Clinically meaningful difference in means, proportions, and
  odds ratios of the primary endpoint.
  ➡ The larger this value, the greater the degree to which the phenomenon under study is manifested.
- The effects sizes are determined by the study design: primary endpoints (PEs), study purpose and design, and repeated measurements.
- In hypothesis testing, the H_0 always means that the effect size is zero.
- As the effect size increases ➡ The power also increases with a fixed sample size.

Effect size 2: α, β, and ES

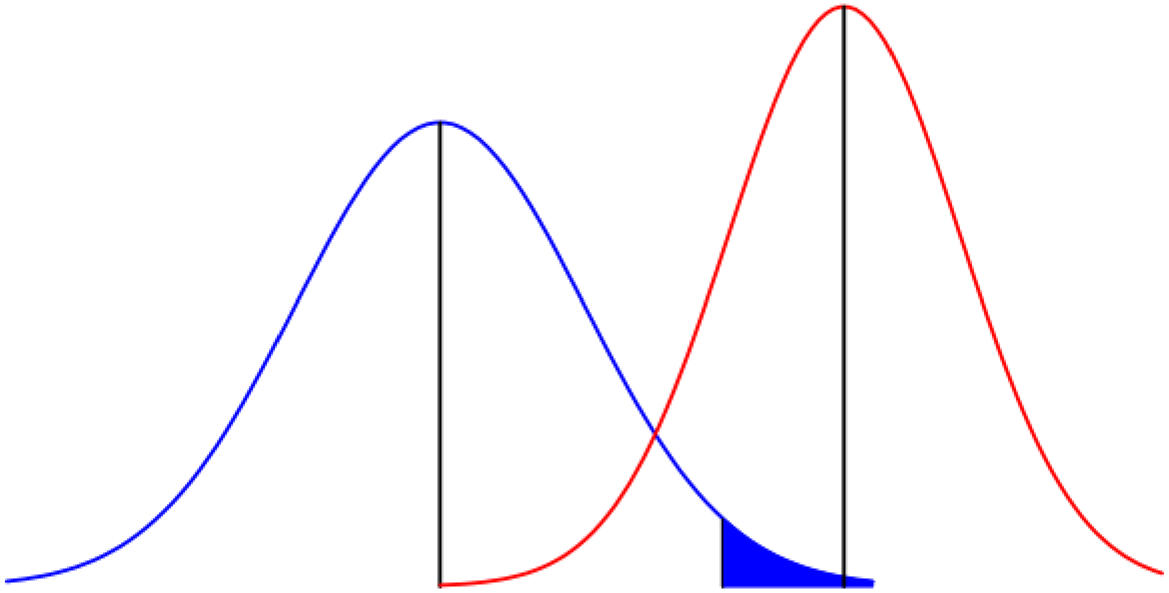


Figure 1. Distribution under the null hypothesis (blue), distribution under the alternative hypothesis (red), and the blue area: 2.5% of alpha level.

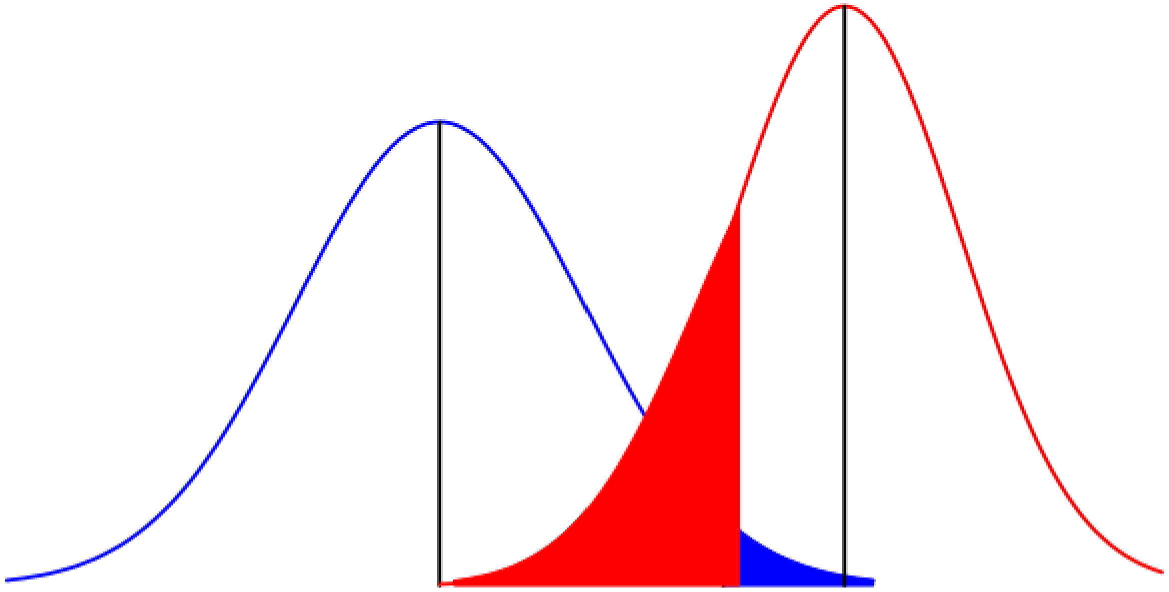
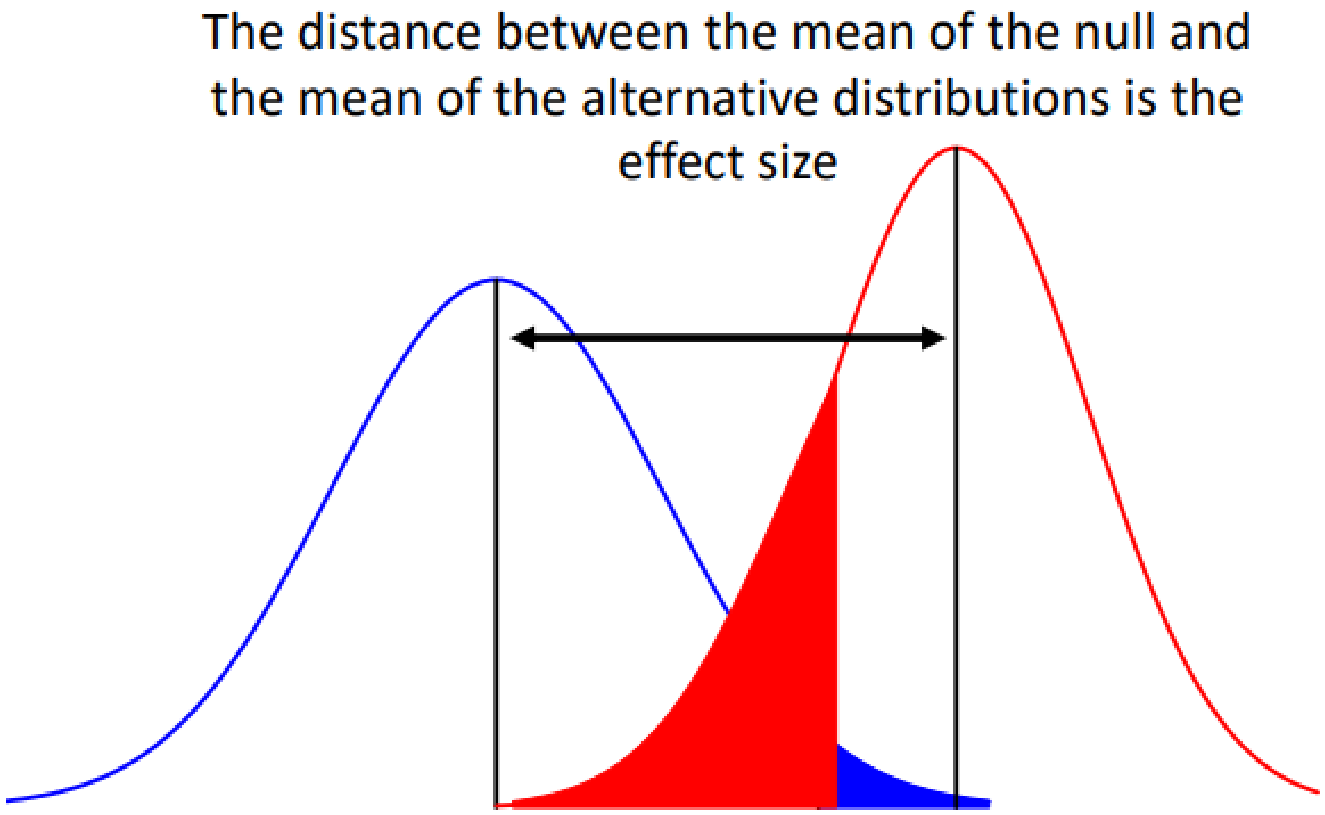


Figure 2. Distribution under the null hypothesis (blue), distribution under the alternative hypothesis (red), and the red area: beta level.

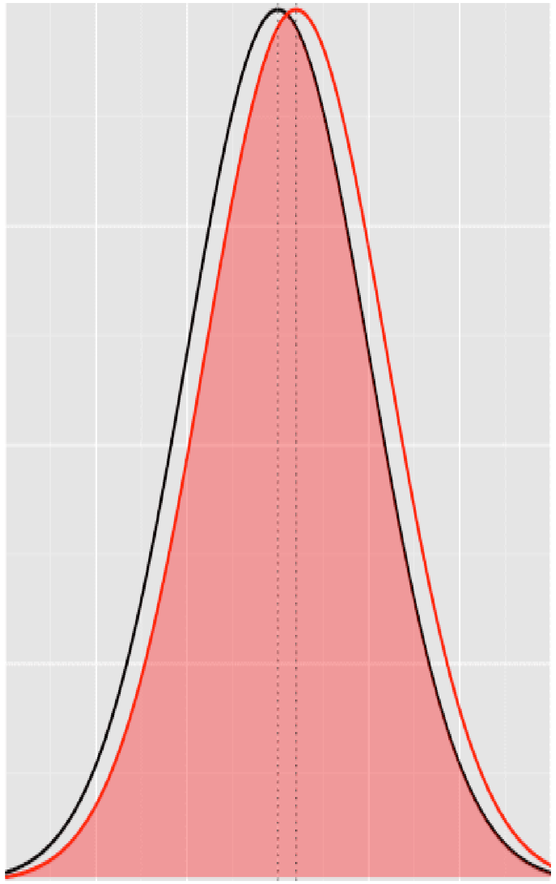


The distance between the mean of the null and the mean of the alternative distributions is the effect size

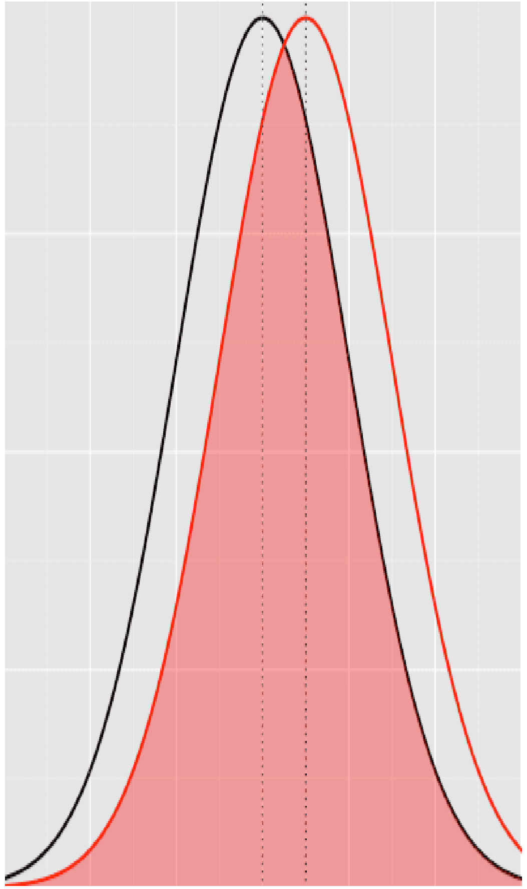
Source: Verhulst B (2022). Hypothesis Testing, effect size and statistical power.

Effect size 3: Cohen’s d

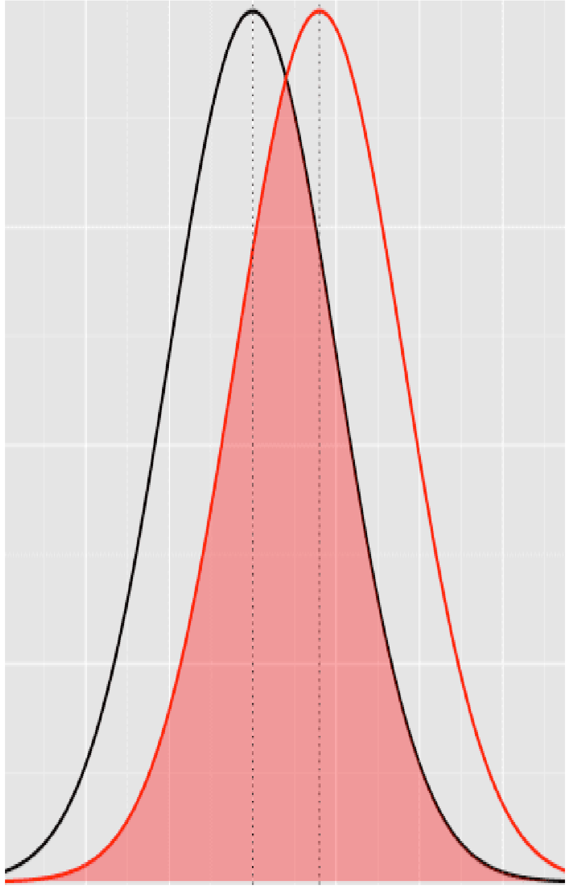
- Be careful with interpreting Effect Sizes. How can we interpret the ES?
- Cohen’s (1988) Standards for Interpreting Effect sizes
  - ➡ Cohens d is a standardized effect size for measuring the difference between two group means.
  - ➡ How to calculate? $d = \frac{\bar{X}_1 - \bar{X}_2}{S_p}$
- Cohen’s d interpretation:
  - ➡ d=0.2 ….. ES is small
  - ➡ d=0.5 ….. ES is medium
  - ➡ d=0.8 ….. ES is large



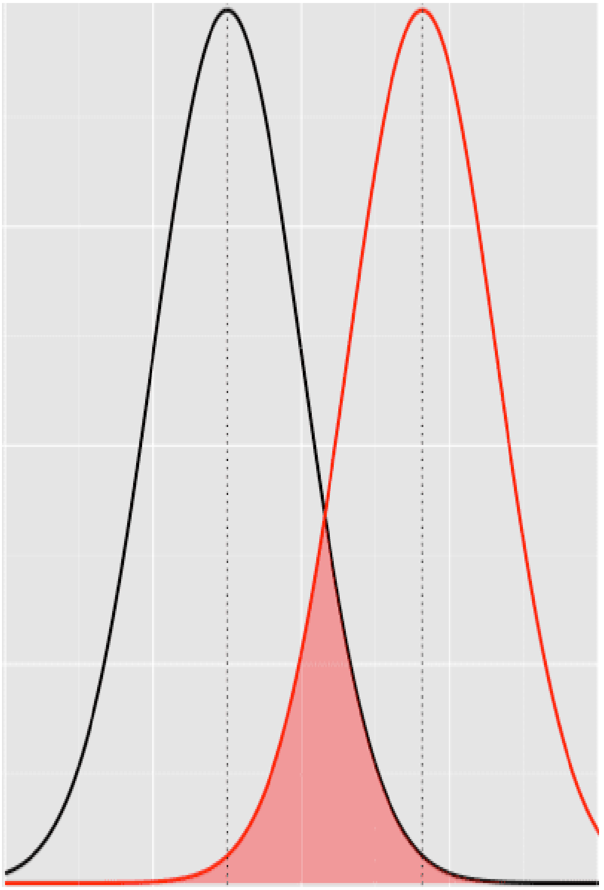
Small ES (Cohen’s d=0.2)



Medium ES (Cohen’s d=0.2)



Large ES (Cohen’s d=0.2)



Clark et al (2006): CT for anxiety disorder, social phobia (d=0.2)

Effect size 4: Examples

A retrospective study was performed to test whether a new biomarker is more effective to cognitive impairment over an existing biomarker. Data were extracted from the existing source (EMR or clinical charts). PE (test scores) were measured and compared between two groups (higher score is the better). What is an ES for this study?

A pilot study was performed to obtain preliminary data for a promising treatment on PD. Investigators could have at most 35 PD patients from the clinics. The PE would be measured at baseline and remeasured in 6 month after the treatment. This data will be used for a formal CT. What is an ES for this study?

A clinical trial was performed to test whether an experimental drug would show strong efficacy over the existing standard drug treatment. This was a phase 2 two-arm trial in Alzheimer’s disease (AD). Patients would be randomized after registration into either group (treatment and control group). The PE of a continuous biomarker would be repeatedly measured at baseline and in 6 month after the treatment. What is an ES for this study?

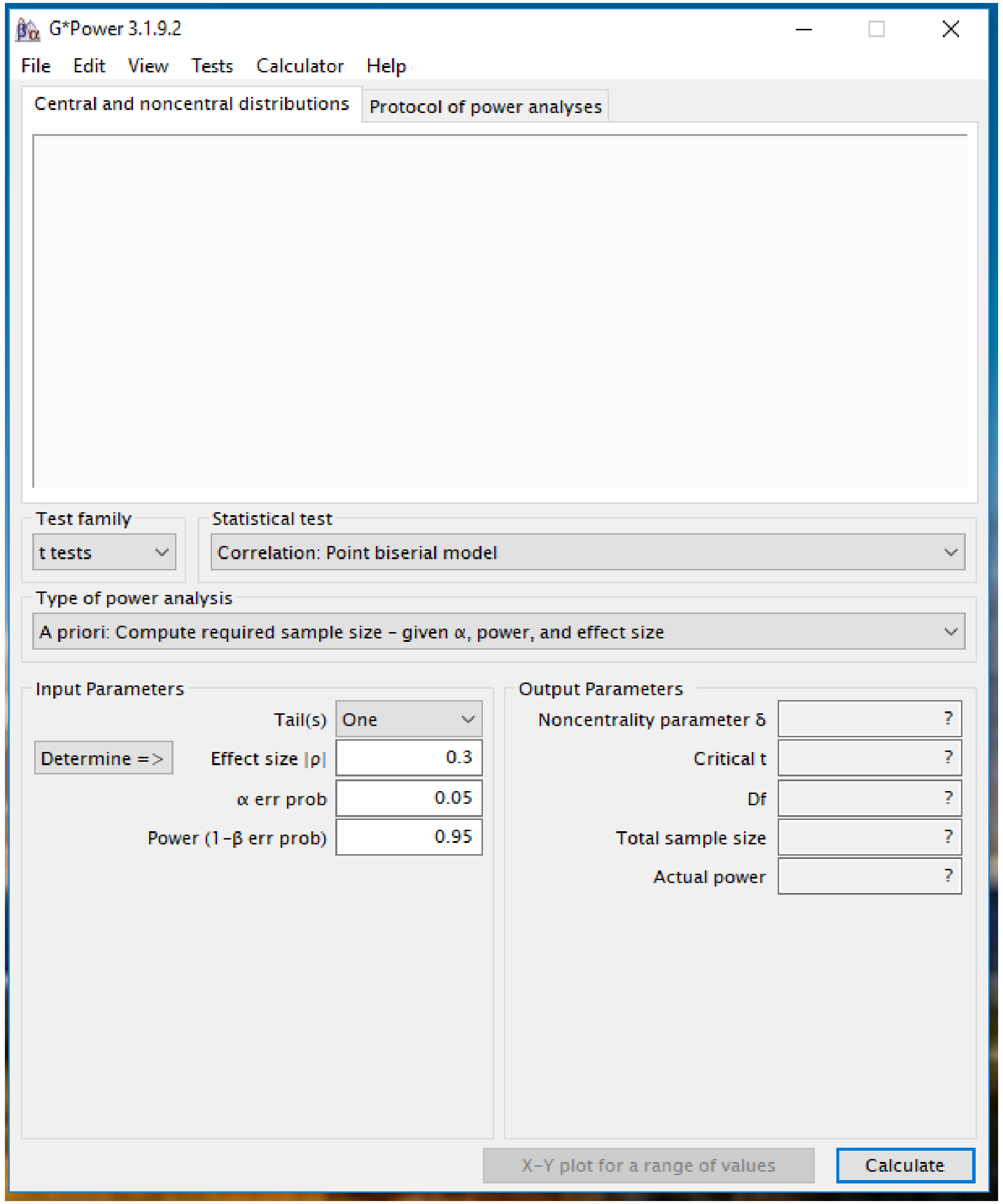
Software:

- **G*Power**: 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

  ➡ Manual and tutorial.

  ➡ Download: G*Power 3.1.9.7 for Windows and G*Power 3.1.9.6 for Mac OS X
- **R and Python**: a programming language for statistical computing and graphics. ➡ Manual and tutorial.
- **GraphPad**: It provides sample size calculation with power analysis. Introductory video is at https://www.graphpad.com/series/how-to-calculate-sample-size-using-power-analysis.
- Others: SAS, STATA, SPSS, etc

G*Power:


G*Power 3.1.9.2
File
Edit
View
Tests
Calculator
Help
Central and noncentral distributions
Protocol of power analyses
Determine =>
X-Y plot for a range of values
Calculate

Test family: t tests

Statistical test: Correlation: Point biserial model

Type of power analysis: A priori: Compute required sample size – given α, power, and effect size

Input Parameters

Tail(s): One

Effect size |ρ|: 0.3

α err prob: 0.05

Power (1-β err prob): 0.95

Output Parameters

Noncentrality parameter δ: ?

Critical t: ?

Df: ?

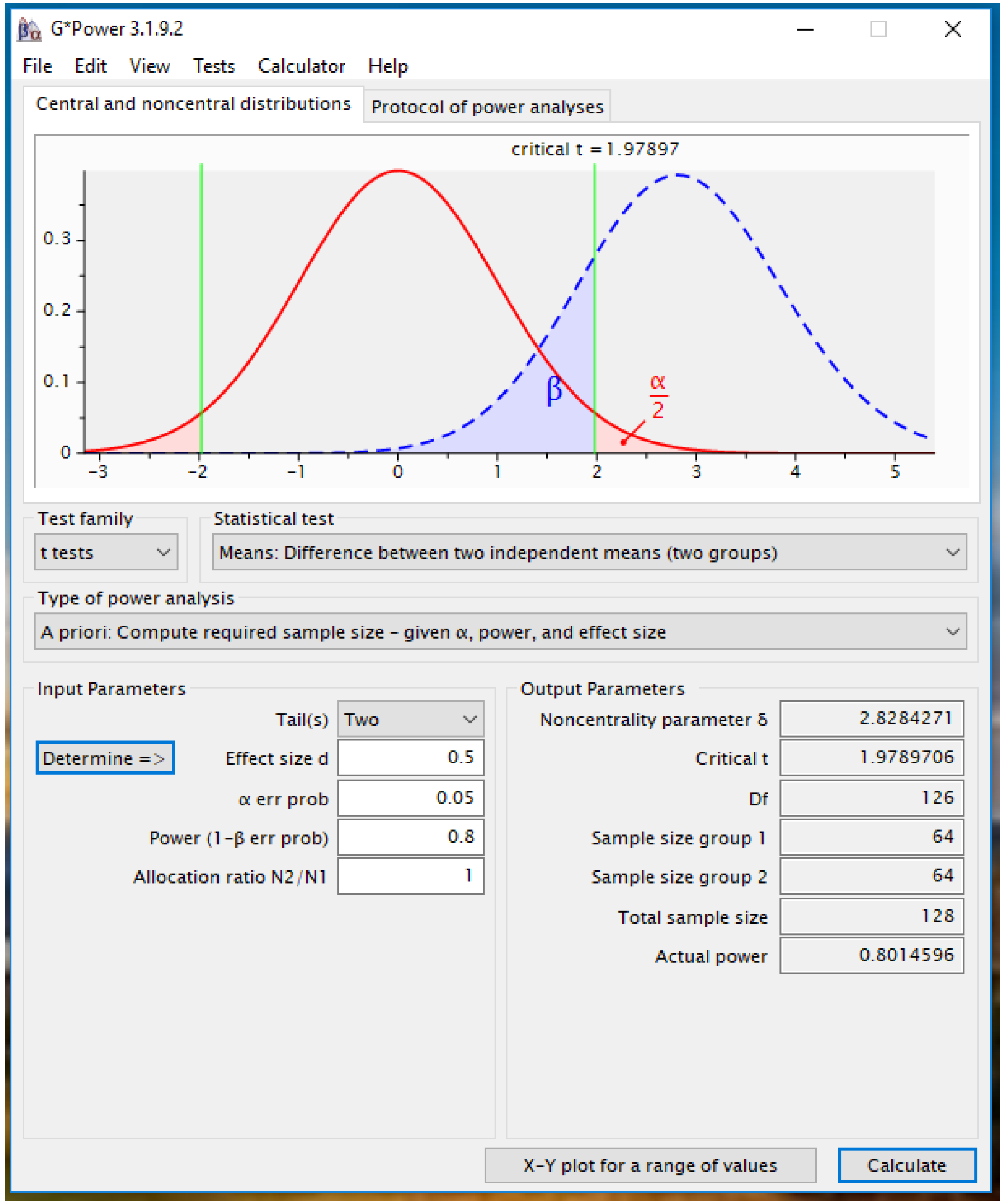
Total sample size: ?

Actual power: ?

Practical example 1:

Let us assume that researchers are planning a study to investigate the analgesic efficacy of 2 drugs. Drug A has traditionally been used for postoperative pain control and drug B is a newly developed drug. A pain score using a Visual Analog Scale (VAS) will be measured at 6 hours post-operation. The researchers want to determine the sample size for the null hypothesis to be rejected with a 2-tailed test, α=0.05, and β=0.2.

- What is the primary endpoint?
- What analysis method should be applied?
- What are two errors?
- How can we determine the effect size? ➡ Three cases …



Test family: t tests

Statistical test: Means: Difference between two independent means (two groups)

Type of power analysis: A priori: Compute required sample size – given α, power, and effect size

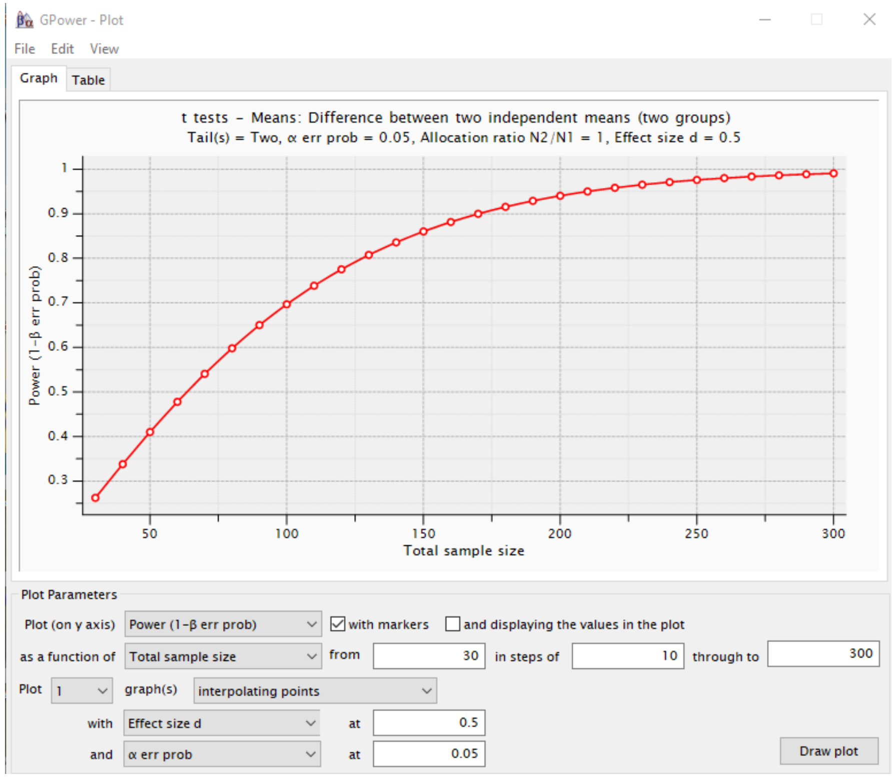
Input Parameters

Tail(s)	Two
Effect size d	0.5
α err prob	0.05
Power (1−β err prob)	0.8
Allocation ratio N2/N1	1

Output Parameters

Noncentrality parameter δ	2.8284271
Critical t	1.9789706
Df	126
Sample size group 1	64
Sample size group 2	64
Total sample size	128
Actual power	0.8014596

- [Step 1] Analysis method? From the top menu, click … Tests - Means – Two independent groups
- [Step 2] Input parameters from “Determine=>”: Two tails, ES=0.5, alpha=0.05, power=0.8, ratio=1
- [Step 3] Click “Calculate”. Then you can get the left window result.


GPower - Plot
File
Edit
View
Graph
Table
t tests – Means: Difference between two independent means (two groups)
Tail(s) = Two, α err prob = 0.05, Allocation ratio N2/N1 = 1, Effect size d = 0.5
Power (1−β err prob)
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
50
100
150
200
250
300
Total sample size
Draw plot

Plot Parameters

Plot (on y axis): Power (1−β err prob) ☑ with markers ☐ and displaying the values in the plot

as a function of: Total sample size from 30 in steps of 10 through to 300

Plot 1 graph(s) interpolating points

with Effect size d at 0.5

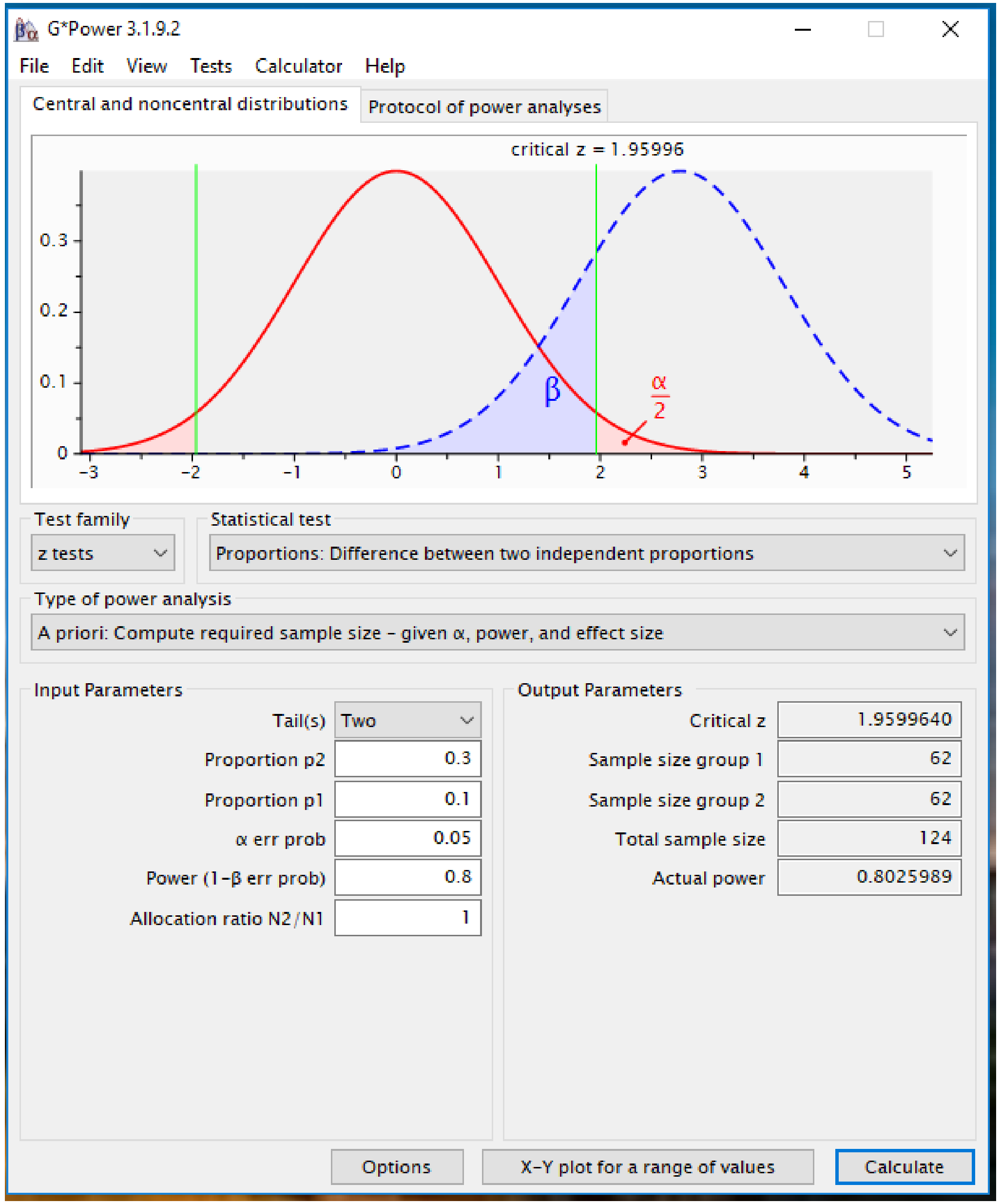
and α err prob at 0.05

Practical example 2:

Assume that a study examined the effects of 2 treatments, for which the measure of the effect is a **proportion**. Treatment A has been traditionally used for the prevention of post-herpetic neuralgia, and treatment B is a newly developed treatment. The researchers wanted to **determine the sample size** for the null hypothesis to be rejected using 2-tailed testing, α=0.05, and β=0.2. The number of patients was equal in both groups. Suppose the researchers conducted a pilot study examining the effects of treatments A and B. In the pilot study, the proportions of post-herpetic neuralgia development were 0.3 and 0.1, respectively.

- What is the primary endpoint?
- What analysis method should be applied?
- What are two errors?
- How can we determine the effect size? ➡ Three cases …

- [Step 1] Analysis method? From the top menu, click ... Tests - Proportions – Two independent groups: inequality, z-test.
- [Step 2] Input parameters from “Determine=>”: Two tails, p2=0.3, p1=0.1, alpha=0.05, power=0.8, ratio=1
- [Step 3] Click “Calculate”. Then you can get the left window result.



Test family: z tests

Statistical test: Proportions: Difference between two independent proportions

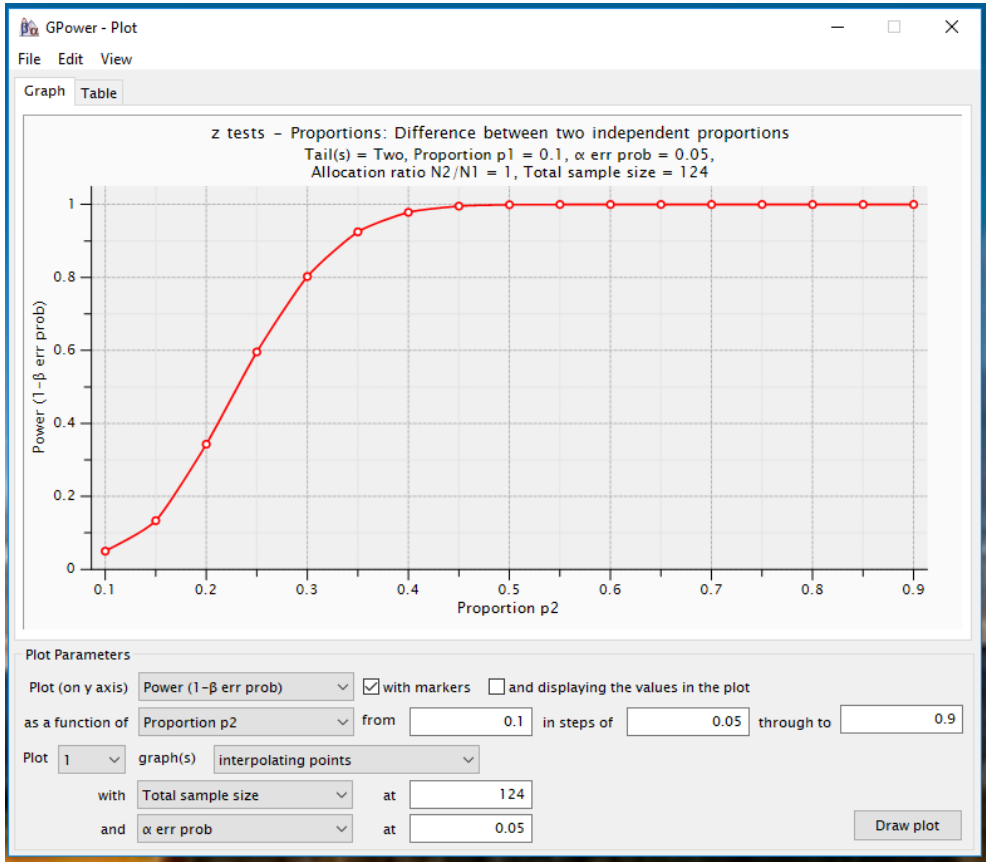
Type of power analysis: A priori: Compute required sample size – given α, power, and effect size

Input Parameters

Tail(s)	Two
Proportion p2	0.3
Proportion p1	0.1
α err prob	0.05
Power (1-β err prob)	0.8
Allocation ratio N2/N1	1

Output Parameters

Critical z	1.9599640
Sample size group 1	62
Sample size group 2	62
Total sample size	124
Actual power	0.8025989



z tests – Proportions: Difference between two independent proportions
Tail(s) = Two, Proportion p1 = 0.1, α err prob = 0.05,
Allocation ratio N2/N1 = 1, Total sample size = 124

Plot Parameters

Plot (on y axis) Power (1-β err prob) ☑ with markers □ and displaying the values in the plot

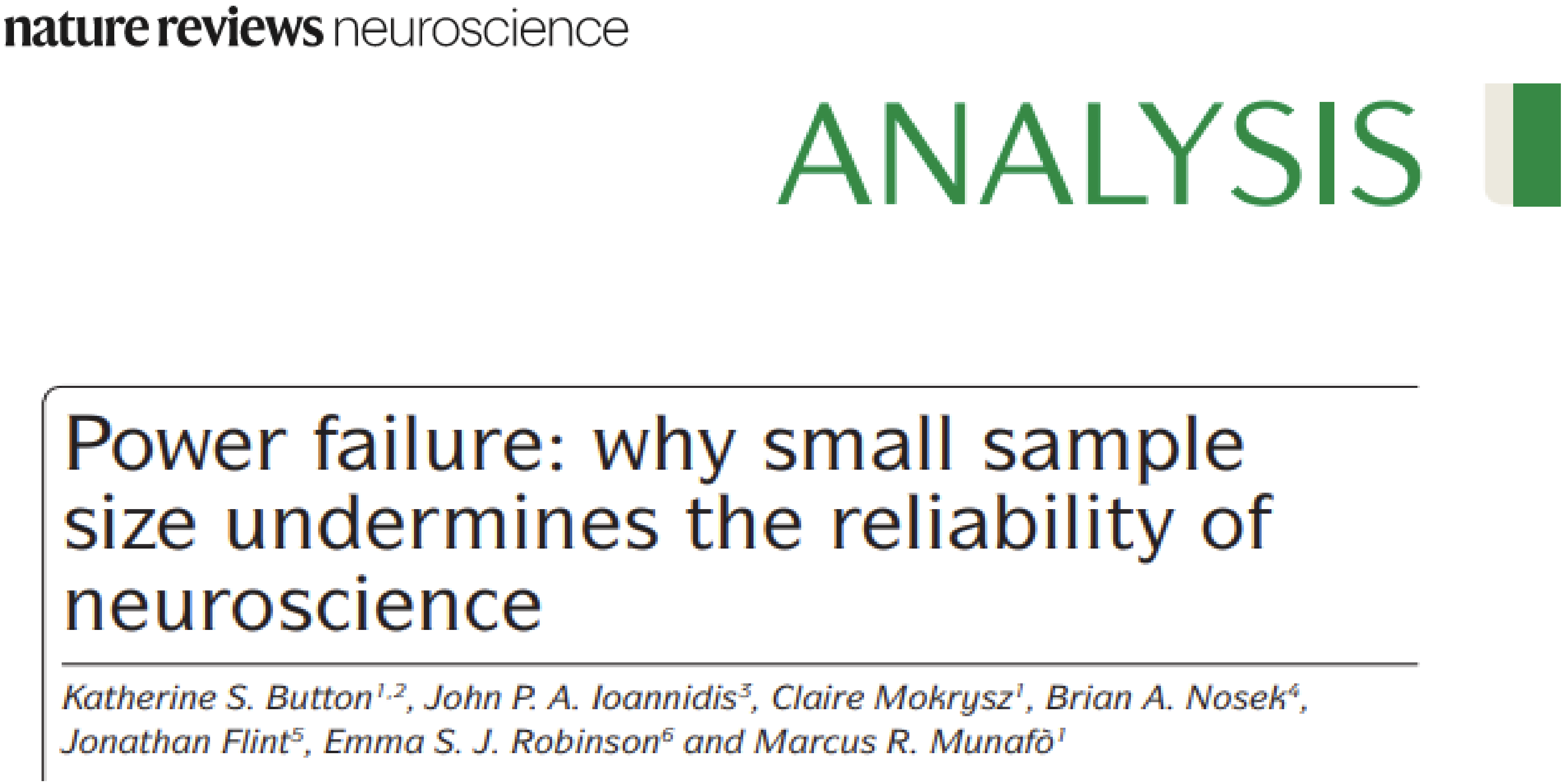
as a function of Proportion p2 from 0.1 in steps of 0.05 through to 0.9

Plot 1 graph(s) interpolating points

with Total sample size at 124

and α err prob at 0.05

Button KS et al. (2013): Power failure due to small sample size



nature reviews neuroscience

ANALYSIS

Power failure: why small sample size undermines the reliability of neuroscience

Katherine S. Button[1,2], John P. A. Ioannidis[3], Claire Mokrysz[1], Brian A. Nosek[4], Jonathan Flint[5], Emma S. J. Robinson[6] and Marcus R. Munafò[1]