httpswww.vox.comfuture-perfect21504366science-replication-crisis-peer-review-statisticsfbclid=IwAR3IlJXfXBVwFWaE5aw4RXHKY

NHST, CI, EFFECT SIZES AND STATISTICAL POWER

Ph.D Programme in Psychology, Linguistics and Cognitive Neurosciences

QUICK REFRESH ON GLIMs

- From the General Linear Model to the Generalised Linear Model
- Results form first application
- Bridge from GLIM to GLIMM
- Which assumption of the GLIM are violated when GLIMM is needed?
- Are there any ways for checking the need for a GLIMM instead of a GLIM?

PLAN OF THE LESSON

- Part I
  - Icebreakers: NHST and *p*-values

- Part II
  - Effect sizes, P-values, & Power
  - The language of power analysis
  - Types of power analysys

- **Part III**
  - Conducting, running and reporting a power analysis
  - Software for Power Analysis:
    - G*Power
    - R (introduction)

PART I

Icebreakers: NHST and *p*-values

A SHORT PREMISE: SOME HISTORY

- Ronald Fisher in the 20's, described the testing of a null hypothesis and used *p* values to this purpose. He did not set the 0.05, 0.01 criteria
- Neyman and Pearson, in the 30's, described the extension of Fisher's model by adding the notion of the alternative or research hypothesis
- The use of the *p*-value as compared with standards of 0.05 and 0.01 followed soon after that
- In early 90's there were suggestions to include effect sizes in reporting (APA Style Manual, 1994).
- In 1999, the Task Force on Statistical Inference (TFSI) was formed to report on the controversy about significance testing and to promote the use of alternative methods The TFSI also described the different effect sizes that could be used.
- Many of the methods for effect size were introduced a long time ago, pre-Fisher, as *r* and η
- Fisher described eta-squared or the correlation ratio as a measure of variance accounted for
- Cohen's *d* about 1962 Glass effect size about 1976 Hedges effect size about 1981.

NULL HYPOTHESIS SIGNIFICANCE TESTING (NHST)

- Critical value and corresponding *p* level of significance **(criterion of significance)**:
  - *p* value the proportion of null experiments, out of all possible experiments, that will turn out significant even when the null hypothesis is true (α usually 05 , more seldom .01)

- It is related to sampling (**false positive**) [Analogously for **type II error, false negative].**
  - We sample (evidence) from all possible values in the population (the sample space). In the long run, sooner or later, we will sample from the 'extremes', which are eccentric and in this case the evidence not the general 'true' behavior.

- ***p*-value:** probability of observing test statistics greater than the (absolute value of) the critical value, under the null hypothesis (**false positive**)

****
***
*

CAN *P*-VALUES BE MEASLEADING? (I WANT TO SEE THE STARS!)

*
***
****

- Statistical significance provides no information about size of effects and other aspects. Many *p*-values are not very meaningful if not read together with the effect size and in a replicability conceptual framework.

- *p*-values can be so specific that they don't tell researchers what they need to know
  - Information on p cannot be used to make decisions about how to use the results

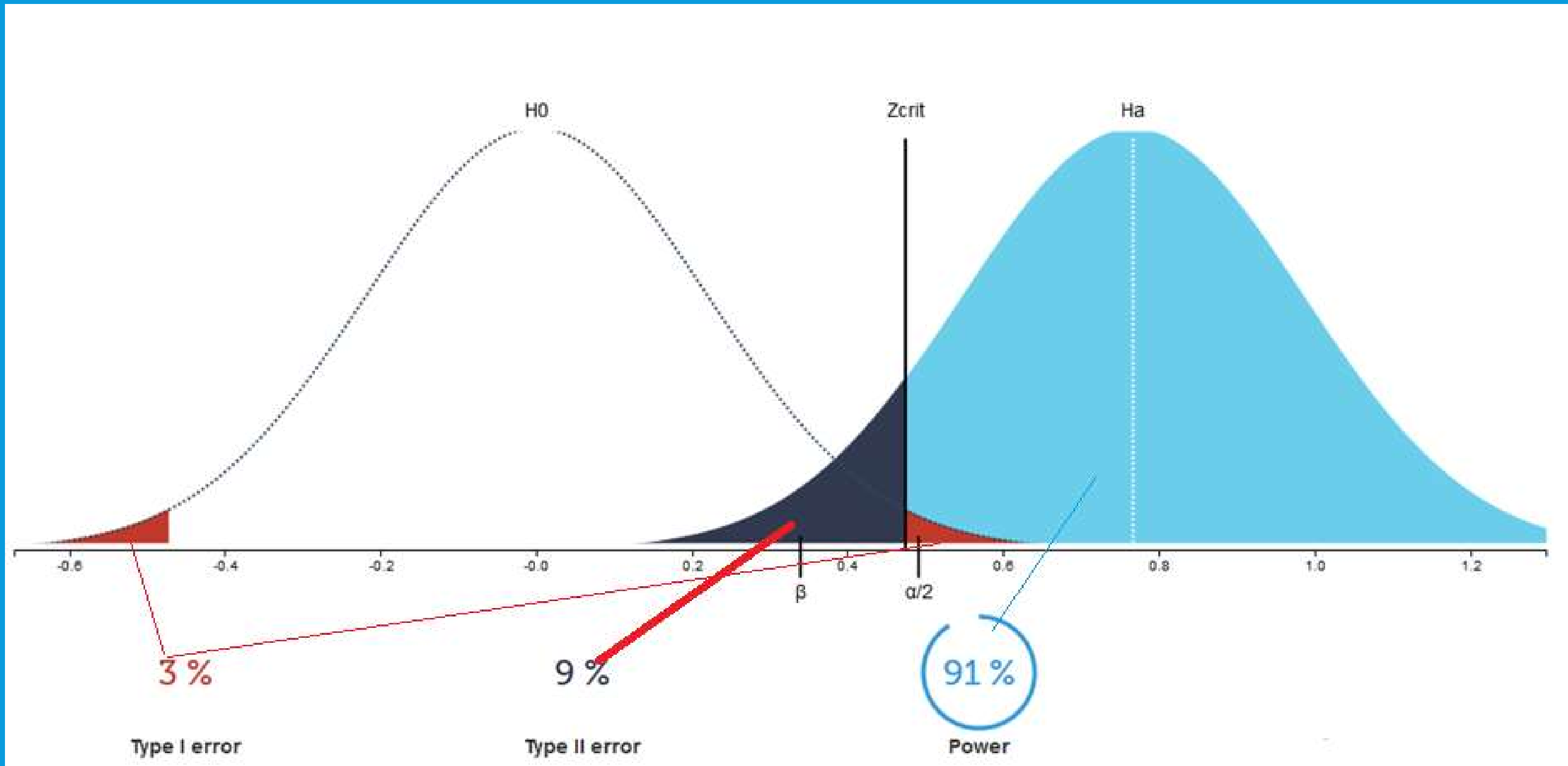
  - *p*-values indicate only that a difference could be or should not be attributed to chance in extracting that specific sample from all possible sample (this holds under the NHST).

NHST: THE LOGIC
REJECTING OR FAILING TO REJECT (FTR)

	Reality: NO EFFECT	Reality: EFFECT EXISTS
Research concludes: FAIL TO REJECT NULL; NO EFFECT	CORRECT FTR	TYPE 2 ERROR (β)
Researcher concludes: REJECT NULL; EFFECT EXISTS	TYPE 1 ERROR (α)	CORRECT REJECT (1-β)

Type I and II errors
Coeteris paribus, when type I decreases, type II increases

NHST: THE GRAPH



TYPE I ERROR RATE

- The first error rate, **the significance level**, is chosen by the experimenter and is conventionally one of 5%, 1% or 0.1% or (P=0.05) (P=0.01) or (P=0.001). Each error rate gives a **threshold value** (on the X axis) that must be exceeded for significance.
- Interpretation: in 5% of experiments in which there is no real treatment effect the t statistic will exceed the threshold value due to chance in sampling
- Choice of Type I error rate. The smaller the rate chosen the stronger the evidence that there is an effect of treatment.

The typical reference t-test for the difference in the means of two population
Example: difference in height between basket players and the general population, treatment vs control.

BEWARE

- *nominal alpha*: the probability of making a Type I error when all the assumptions are met
- *real alpha*: the probability of making a Type I error when one or more of the assumptions are violated
- Real alpha higher then nominal alpha: ***alpha inflation***

NHST VERSUS CI

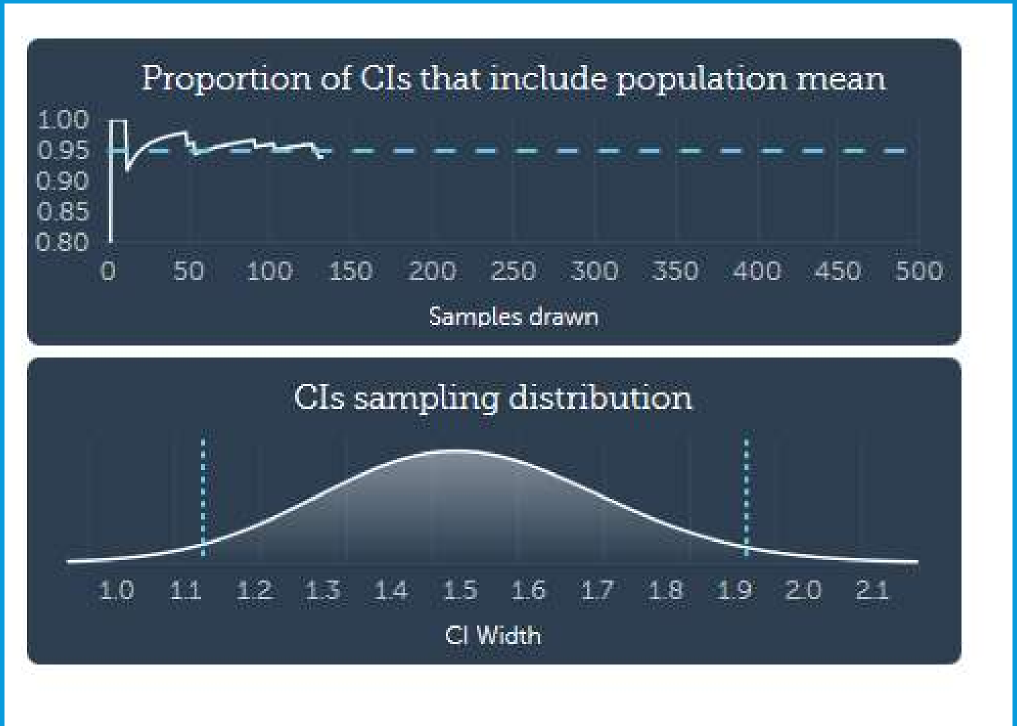
t test

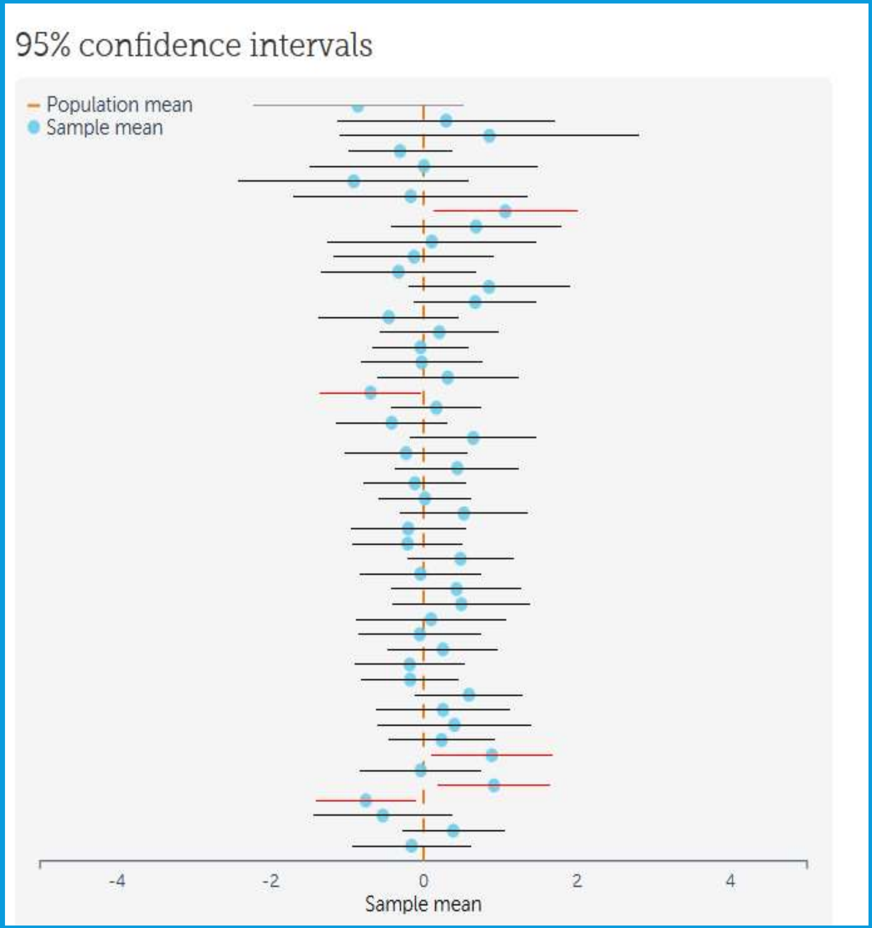
mean under the null hypothesys: 120 sentences

	t	df	Sign. (two tails)	mean diff	95% CI lower	95% CI upper
tot_frasi	8462673,000	101	,000	67049020,00	51332092,00	82765947,00

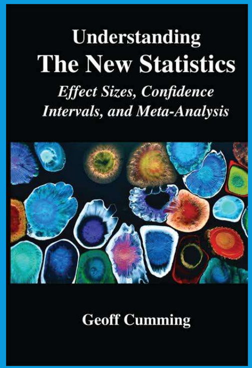
What is the main difference between NHST and CI?

- NHST: acceptance region is the bilateral CI under the null hypothesis, i.e. when the value of the null hypothesis is considered the TRUE mean in the population
- CI: no hypothesis or knowledge whatsoever about the value of the mean (or means difference) in the population
- Is the CI informative? When its precision its high, i.e. when the CI width is small: (upper limit-lower limit)


Proportion of CIs that include population mean
1.00
0.95
0.90
0.85
0.80
0 50 100 150 200 250 300 350 400 450 500
Samples drawn
CIs sampling distribution
1.0 1.1 1.2 1.3 1.4 1.5 1.6 1.7 1.8 1.9 2.0 2.1
CI Width

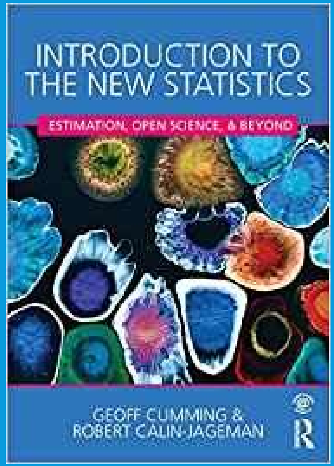

95% confidence intervals
Population mean
Sample mean
-4 -2 0 2 4
Sample mean

NEW VERSUS OLD (?)



The accent is on replicability. The 'novelty' lies in this perspective.

The need for a larger use of CIs has been underlined by APA and it does not imply NHST neglection.



Contents have been largely known and discussed (see article discussing NHST and CI, written in 1979 and later papers in *Psychological Science*)

PART II

Effect sizes, P-values, sample size and Power

TYPE I ERROR RATE AND THE EFFECT SIZE

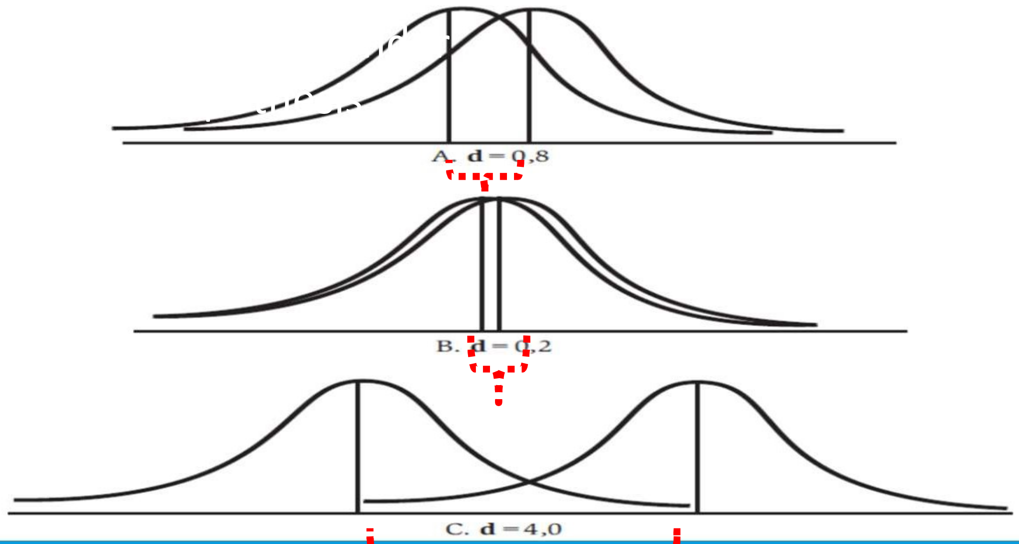
- In research, *p* values are one of the two measures used to report study results. The other one is the effect size (ES).
- An ES is what the result found, e.g. the difference found in the mean scores between two groups. It measures the strength of the result and it is pure, as it does not depend on sample size., unlike p-values.
- ES = (Meantreatment – Meancontrol)/Sdpooled *It is a pure number!*
- A general definition of ES is that it is a family of indexes that measure the magnitude of a treatment effect
- What is the probability of detecting an effect when the effect exists in the population?
- Real effects may be very important or very unimportant. Is the ES generalizable?

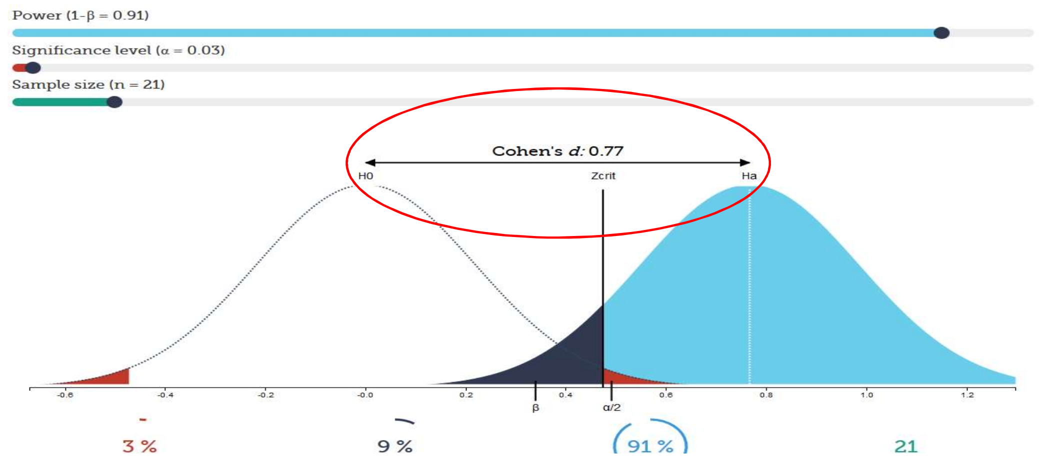
EFFECT SIZE

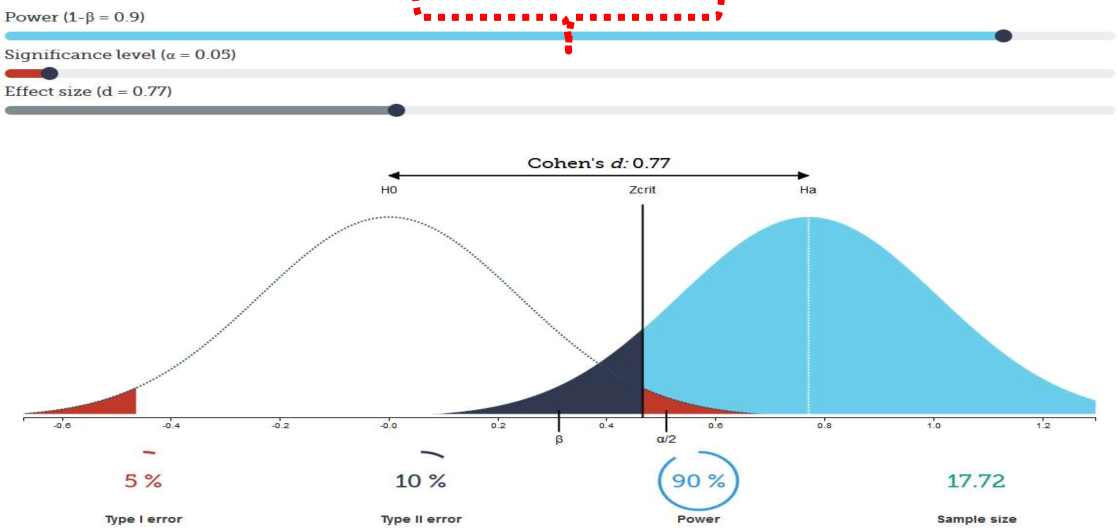
- Knowledge of the magnitude of a treatment effect is qualitatively different than knowing if the effect is real.
- Real ES may be very important or very unimportant (negligible).
- Two types of ES:

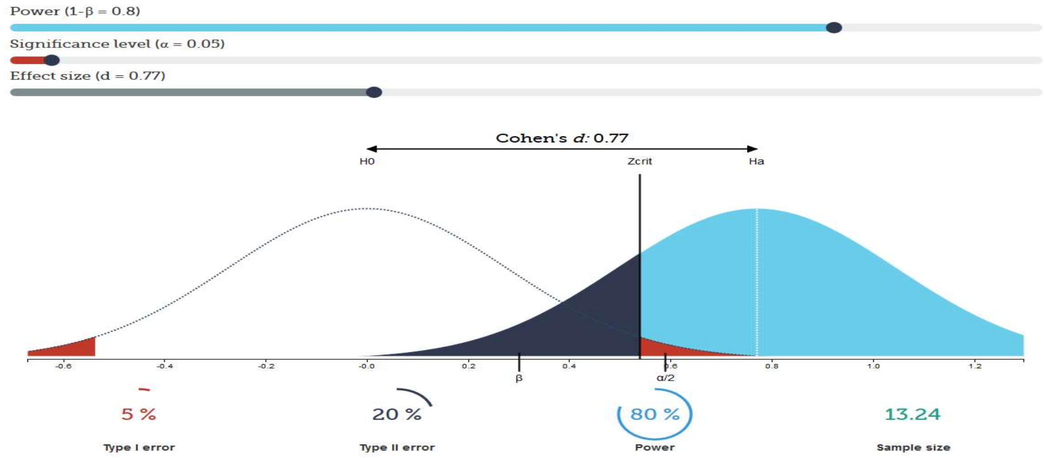
1. standardized units of difference
2. Variance-accounted-for statistics

UNDER THE NORMALITY ASSUMPTION: THE BIGGER D, THE BETTER



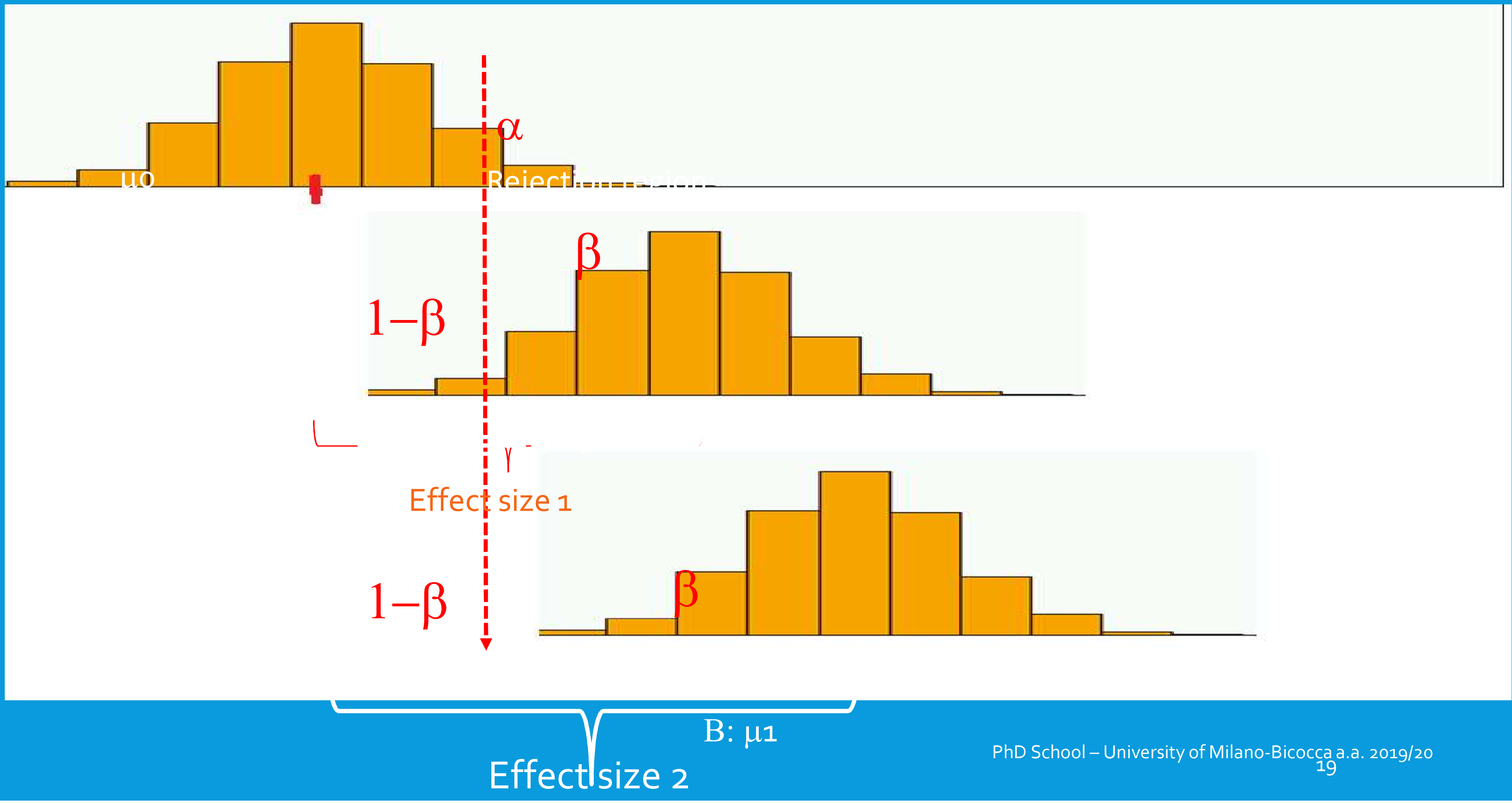






Bigger and bigger effect size *d*

A bigger effect size requires a smaller power to be detected



ES IN STANDARDIZED UNITS

An ES is the observed difference between means, proportions, etc.

To be useful for comparative purposes, this difference needs to be standardized.

Standardization relies on the *pooled variance*, i.e. on the variance computed on all subjects.

- Cohen’s d (most common)
- Glass’s g

EFFECT SIZES IN STANDARDIZED UNITS OF DIFFERENCE

- Observed difference on means, proportions, etc. The difference needs to be standardized
- Cohen's d (most common), Glass's g'

Cohen's d (1969)

Population Form

$$\delta = \frac{\mu_1 - \mu_2}{\sigma}$$

Statistic Form

$$d = \frac{\overline{X}_1 - \overline{X}_2}{s}$$

Glass's Effect Size (1978)

$$g' = \frac{\overline{X}_{Experimental} - \overline{X}_{Control}}{S_{Control}}$$

σ and s represent the total score standard deviation

First recognized effect size.
Mean differences in units of common population standard deviation (called pooled sd).
Cohen used this as the basis his research on power.

Glass proposed a modification of the Cohen *d* the common standard deviation replaced with the standard deviation of the control group.

VARIANCE-ACCOUNT-FOR STATISTICS

- Very similar to the correlation coefficient (r) and the coefficient of determination ($r2$). These provide indications of the proportion of variance that can be attributed to the treatment.
- For mixed models, the intraclass correlation coefficient, estimates the effect size. It measures the proportions of total variance in the second (higher) level of the model.
- In case of nominal variables, Cramer's V is applied, as a transform of the Chi-square

  - Eta-squared, $\eta 2$
  - Intraclass Correlation, ICC
  - Cramer's V

EFFECT SIZES AS VARIANCE-ACCOUNTED-FOR STATISTICS

Model	Effect size
Regression, anova	Eta-squared, $\eta 2$ More often used in meta-analysis A η^2 of 0.25 would indicate that 25% of thetotal variation is accounted for by the treatment variation. $\eta^2 = \frac{SS_{Treatment}}{SS_{Total}}$
Mixed models	intraclass Correlation, ICC Interaction can be tricky: Results showed that power varies significantly as a function of model type and whether or not the model is the main model for the study $\rho = \frac{\text{population variance between level-two units}}{\text{total variance}}$
chi-square applications	Cram where $df^* = \min(r - 1, c - 1)$ and r = number of rows and c = number of columns $V = \sqrt{\frac{\chi^2}{n \cdot df^*}}$

Mathieu, J

Relating η2 to Effect Size *d (Coehn, 1988)*

$$r = \frac{d}{\sqrt{d^2 + 4}}$$

EFFECT SIZES THRESHOLDS

Cohen's Standard	*d*	r	r^2
LARGE	0.8	.371	.138
	0.7	.330	.109
	0.6	.287	.083
MEDIUM	0.5	.243	.059
	0.4	.196	.038
	0.3	.148	.022
SMALL	0.2	.100	.010
	0.1	.050	.002
	0.0	.000	.000

In general

By study design

Cramer's *V*

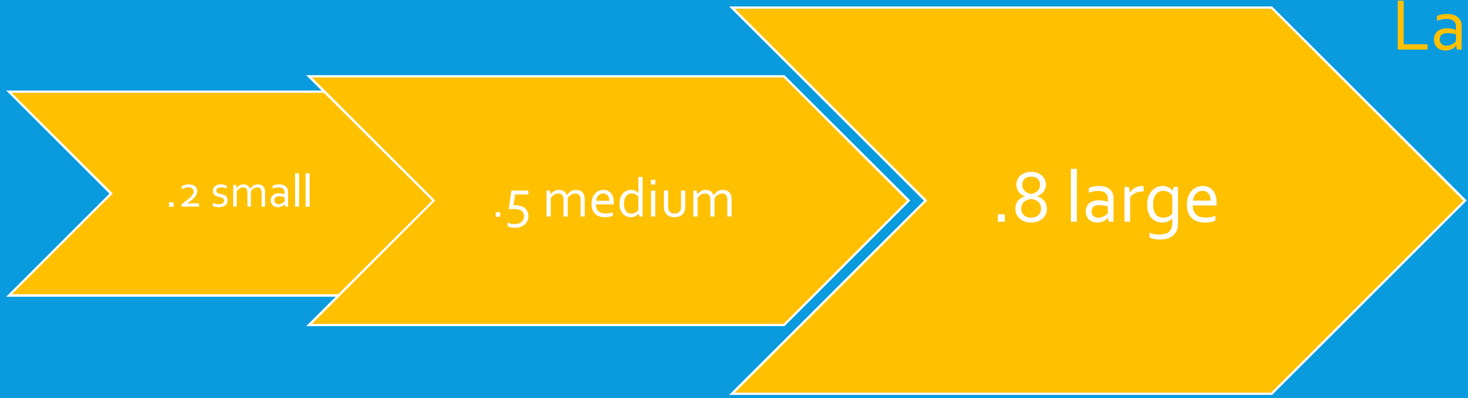
*df**	*small*	*medium*	*large*
1	.10	.30	.50
2	.07	.21	.35
3	.06	.17	.29
4	.05	.15	.25
5	.04	.13	.22

	Effect Size Benchmarks		
Statistic	Small	Medium	Large
Means - Cohen's d	0.2	0.5	0.8
ANOVA - f	0.1	0.25	0.4
ANOVA - eta squared	0.01	0.06	0.14
Regression f-test	0.02	0.15	0.35
Correlation - r or point serial	0.1	0.3	0.5
Correlation - r squared	0.01	0.06	0.14
Association - 2 x 2 table -OR	1.5	3.5	9
Association - Chi-square - w or Phi	0.1	0.3	0.5

INTERPRETING THE EFFECT SIZES/ STANDARDIZED MEASURES

- A *d* or *g*A *d* or *g*′ or *g* of 1.2 indicates that the range of difference among the means is one and two-tenths of the size of the standard deviation.

Large or small?



Cohen chose three values that had been used extensively as standards for effect size.

Beware: Cohen warned about using these standards in practice. The major problem is that effect sizes are influenced by the number of samples and the sample sizes.

:

POWER

- High error rate II β means that a real difference between treatments is unlikely to be detected.
- A rate of 0.5 means that 50% of all possible experiments will not detect effects.
- Power = $1 - \beta$ = : The probability of rejecting the null hypothesis when it is false, i.e. to detect a real difference.
- In everyday language, power is the probability of concluding that the group means differ on the basis of your sample evidence, when the groups means actually differ in the population.

SAMPLE SIZE, TYPE I ERROR RATE, EFFECT SIZE AND POWER
POWER ANALYSIS

- Type II error rate is related to:
  - sample size
  - Effect size
  - SD
  - Type I error rate a (the smaller type I error, the higher type II, coeteris paribus).

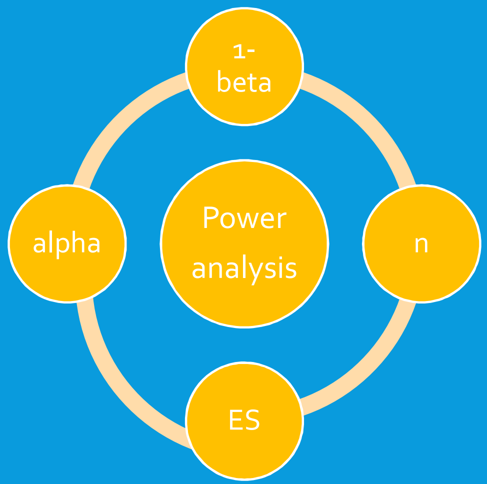
- These factors – ES, sample size, a , power, form a closed system. Once any three are established, the fourth is completely determined.

- What is a power analysis? A process by where one of several statistical parameters can be calculated given others.

- Usually, a power analysis calculates needed sample size given some expected effect size, alpha, and power.

A POWER ANALYSIS INVOLVES FOUR STATISTICAL MEASURES, 'FIXING' THREE OF THEM RESEARCHERS SOLVE FOR THE FOURTH

Probability of Type I error α
- Probability of finding significance where there is none
- False positive
- Usually set to .05



n
- The sample size - usually the parameter you are solving for

Power 1-b
- Probability of finding true significance
- True positive
- e beta is :
- Usually set to .80

ES
- The 'expected effect' is ascertained from:
- Pilot study results
- Published findings from a similar study or studies
- Sometimes calculated from results if not reported
- Field defined 'meaningful effect'
- knowledge of the field)

THE LANGUAGE OF POWER ANALYSYS

Factors Affecting Power

1. Size of the effect ↑
2. Standard deviation of the characteristic ↓
3. Bigger sample size ↑
4. Significance level desired ↓

Other factors in GLIMs

In a more complex model with more parameters or with more complex effects, power ↓ because SE gets larger

Bigger error (unexplained variance and therefore smalle R^2 power ↓ larger error means larger standard error of parameter estimates.

THE LANGUAGE OF POWER/2

desired power (typically .84 for 80% power).

Sample size in each group (assumes equal sized groups)

$$\boxed{n} = \frac{2\boxed{\sigma}^2 (\boxed{Z}_\beta + \boxed{Z}_{\alpha/2})^2}{\boxed{\text{difference}}^2}$$

Standard deviation of the outcome variable

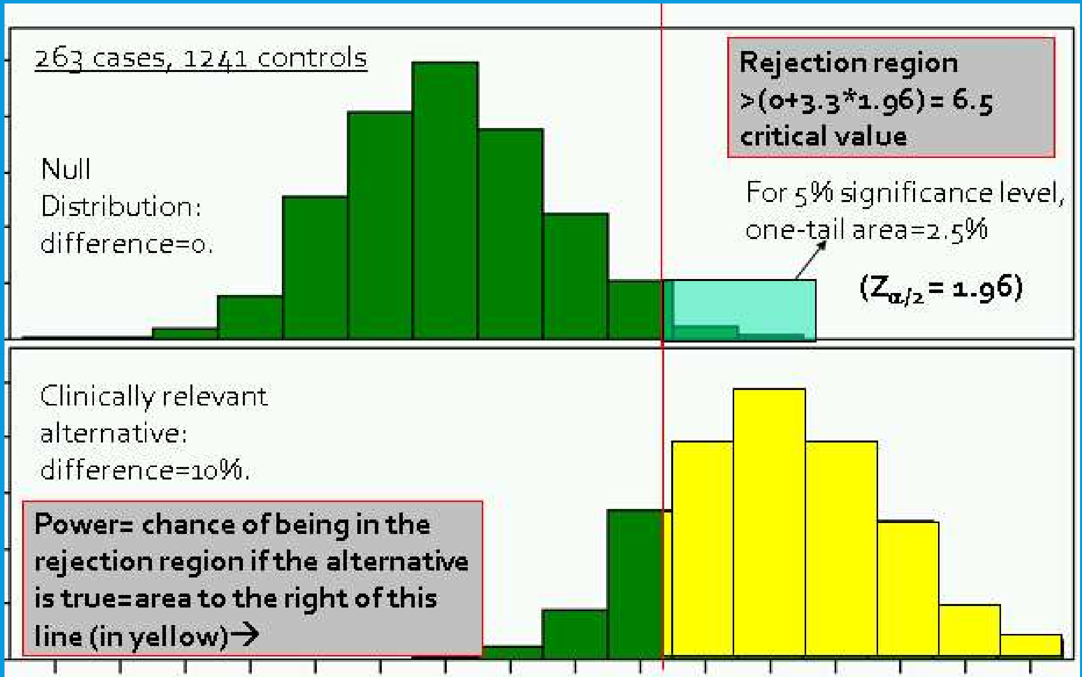
Effect Size (the difference in means)

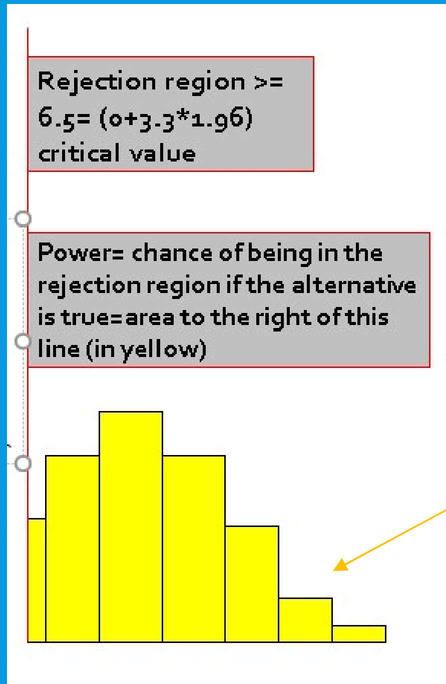
desired **level of statistical significance** (typically 1.96).

All-purpose power formula

$$Z_{power} = \frac{\text{difference}}{\text{standard error(difference)}} - Z_{\alpha/2}$$

HOW α (ES, ETC.) DETERMINES POWER



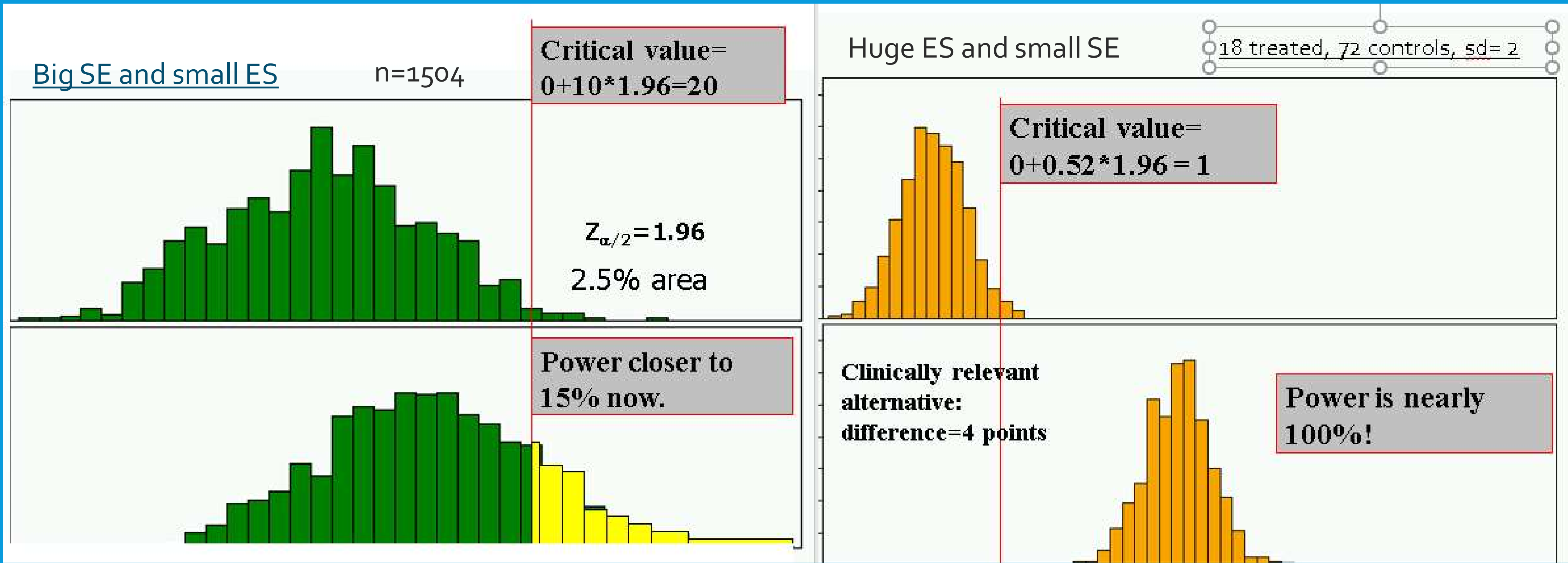


10: mean under H1
3.3=SE

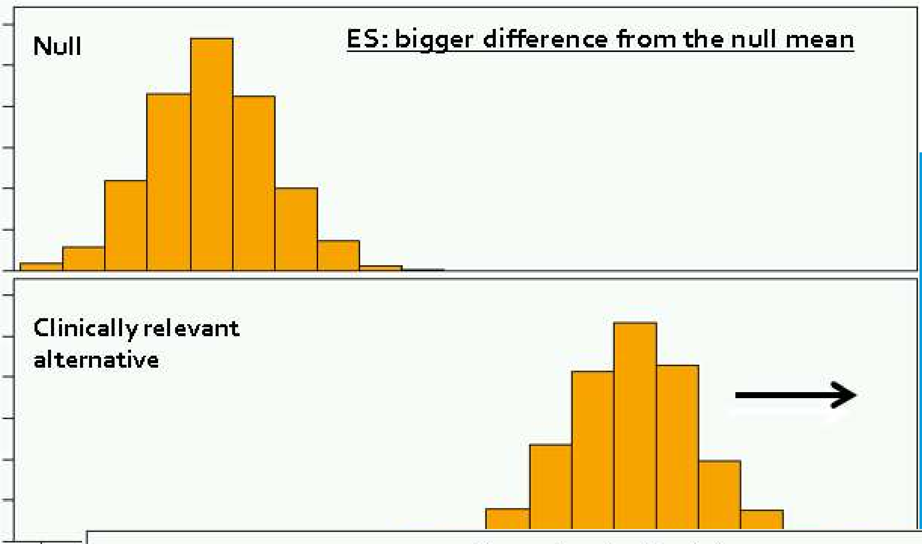
Power:

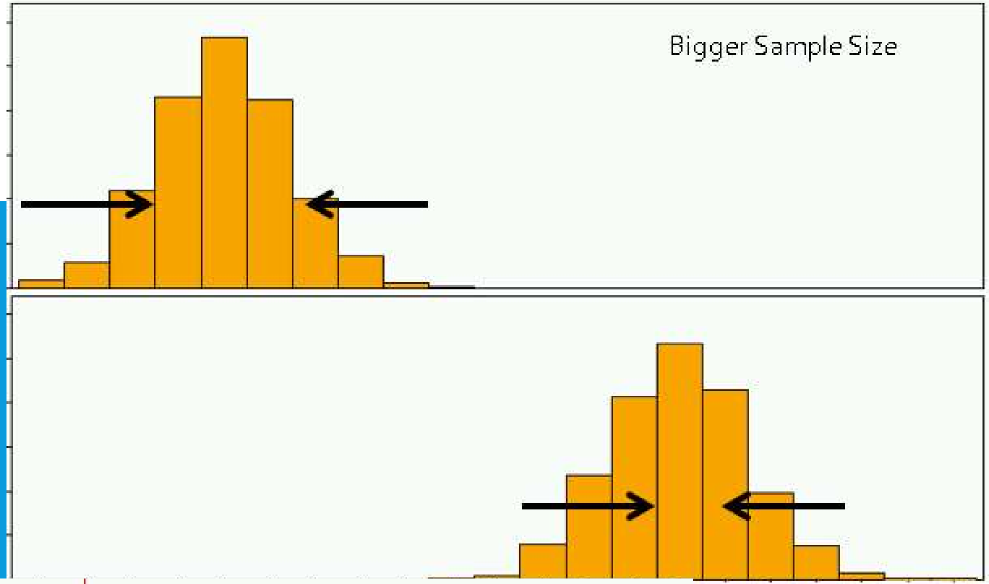
$P(Z>(6.5-10)/3.3$
$=P(Z>1.06)=$
85%

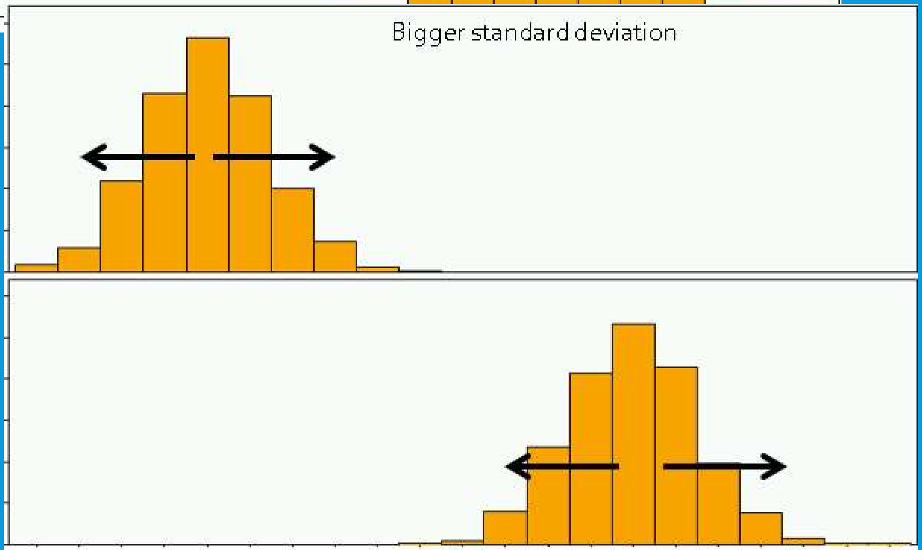
HOW ES AND VARIABILITY DETERMINES POWER

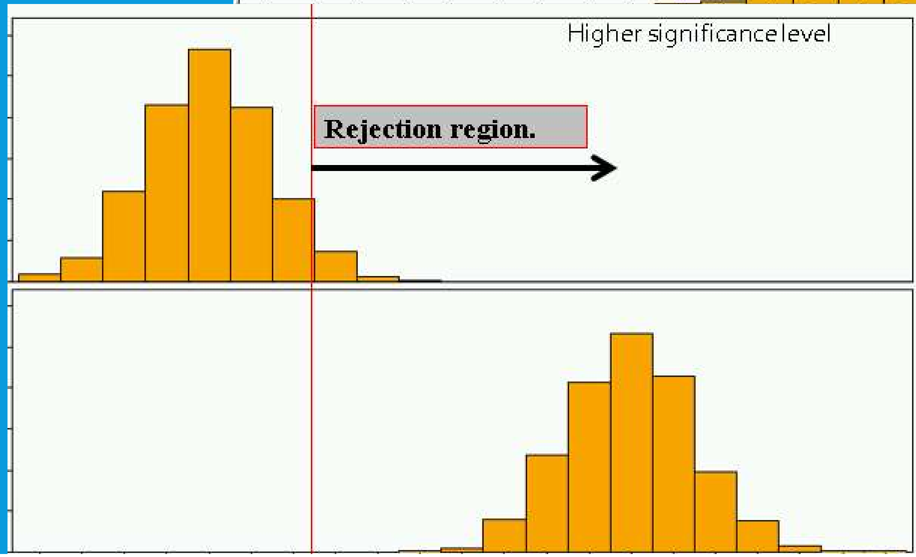


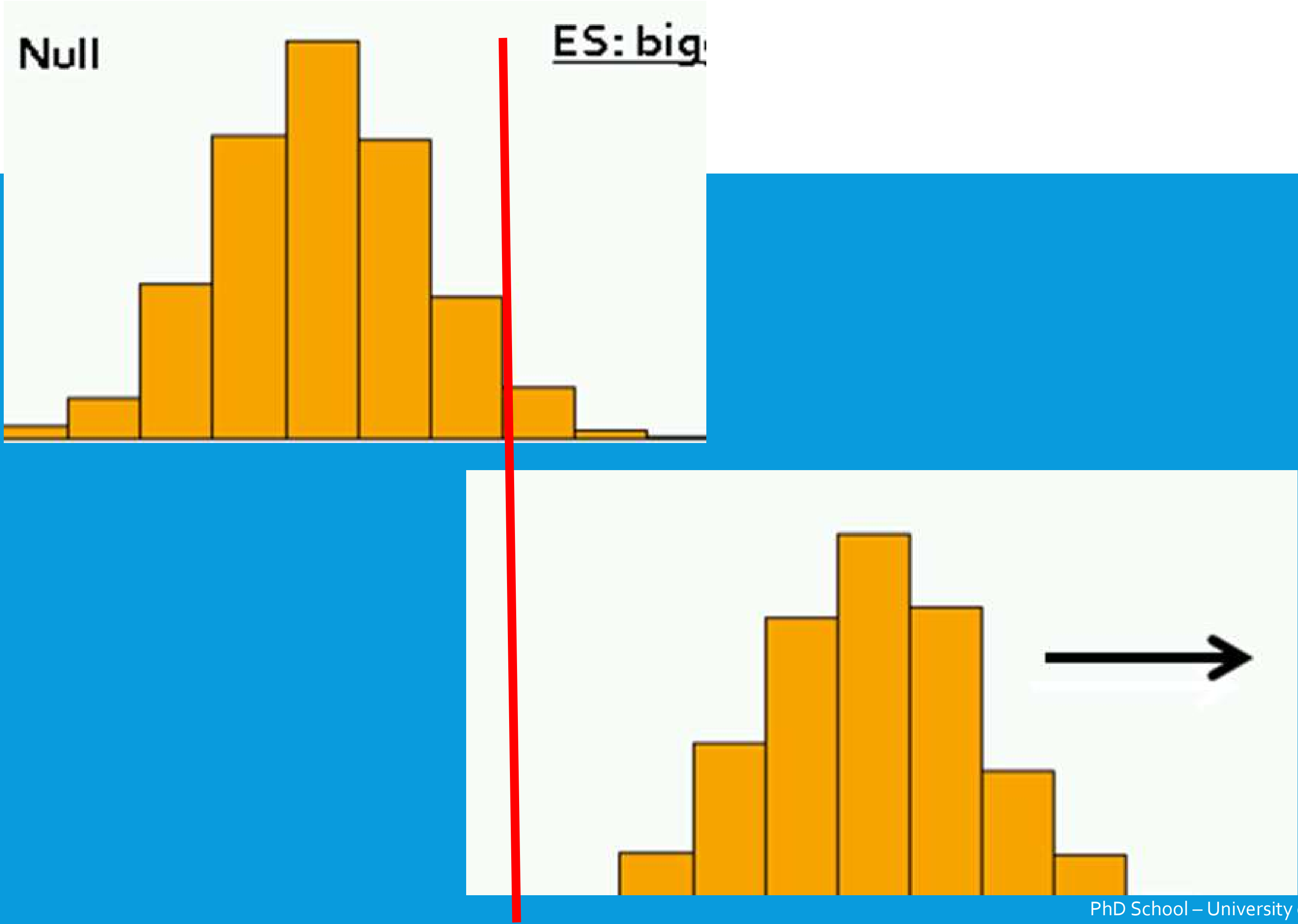
A BRIEF GRAPHICAL OVERVIEW










Null
ES: big

TYPES OF POWER ANALYSIS

Power conventions

- Desired level of power: the more the better, value of .80 minimum threshold standard
- Higher power means more precision in estimating ES (tighter CI)

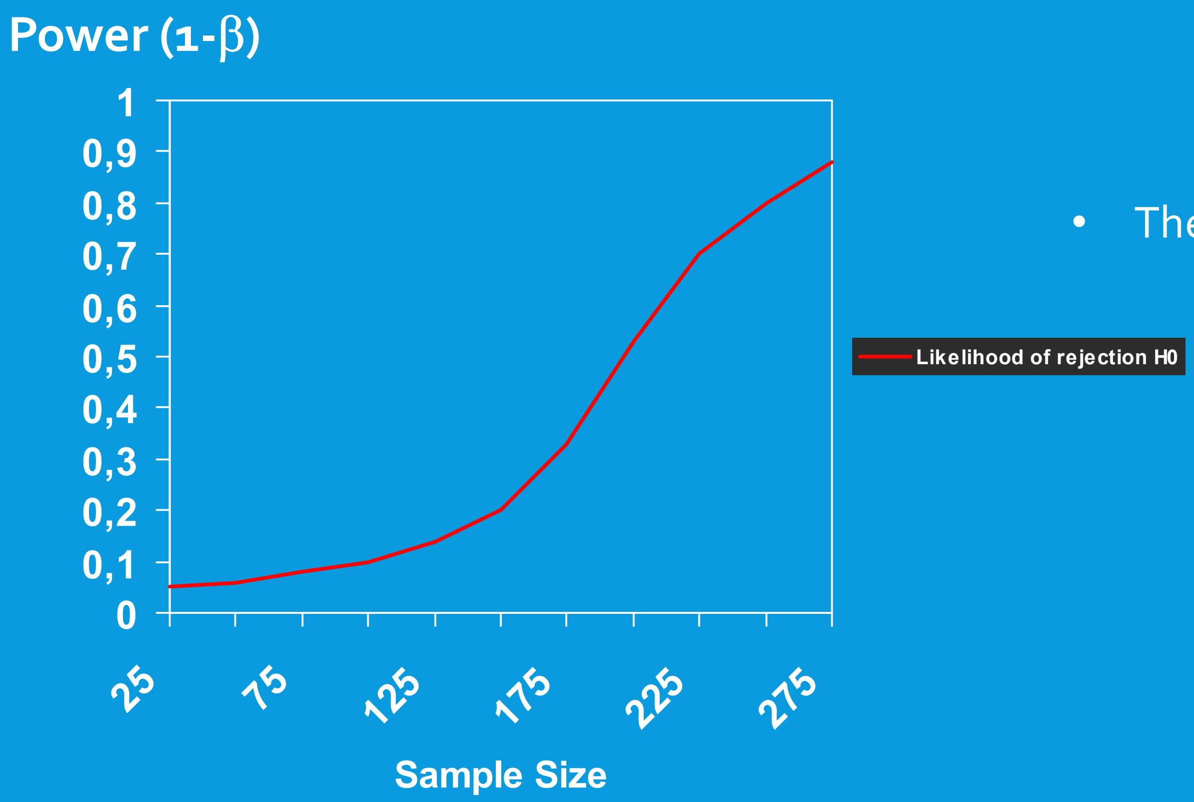
Strategies, determine:

- number of subjects needed (n*) for given level of power (e.g. .80)
- power for a given design (e.g., completed experiment with fixed n)

Types of power analysis

- A priori: compute n, given alpha, power, ES (advised)
- Post-hoc: compute power, given alpha, n, ES (controversial, in designing a study you need a priory)
- Criterion: compute alpha, given power, ES, n (not much used)
- Sensitivity: compute ES, given alpha, power, n (useful for Minimal Detectable Effect MDE)

POWER FUNCTIONS



POWER FUNCTION AND F DISTRIBUTION

- F is a ratio of observed effect to error
  - F = (True Effect + Error) / Error
- The larger the true treatment effect, the larger F you expect to find
  - If the null hypothesis is correct, E(F) = 1.0

- The power of most statistical tests in social sciences can be evaluated via the familiar F distribution (D. Lakens)

HOW TO INCREASE POWER

Increase n

- Effects of adding more subjects are not identical to those of adding more observations

Increase ES

- Choose a different research question
- Use stronger treatments or interventions
- Use better measures

Effects of implementing power analysis:

- Stronger studies: larger samples, better measures
- Fewer studies: adequate studies are harder to do than most people realize

CONDUCTING A POWER ANALYSIS

Are all tests the same in the face of power?

- Some statistical tests are more powerful (i.e., better at detecting real/non-zero population effects) than others.
- Parametric tests often are more powerful than non-parametric, because they work with more information from the data.
- GLIM is Minimum Variance (Estimator) when assumptions are met

Power can be calculated for tests of

- Effect for single regressor, subset of regressors controlling for other regressors in model, or all regressors in the model.

CONDUCTING A POWER ANALYSIS

A priori power analysis (sample size planning)

- set α level (max .05), and desired power (min .80)
- Specify (calculate) expected ES(conservative).
- n is a function of the above factors.

Software

For GLIMs G*Power http://www.gpower.hhu.de/

- Most popular by far, free download available for both the PC and Mac.
- It includes an effect size calculator
- On online tutorial manual

For (few) GLIMMs Optimal Design http://hlmsoft.net/od/

- It is somehow related to HLM, the free version is very basic

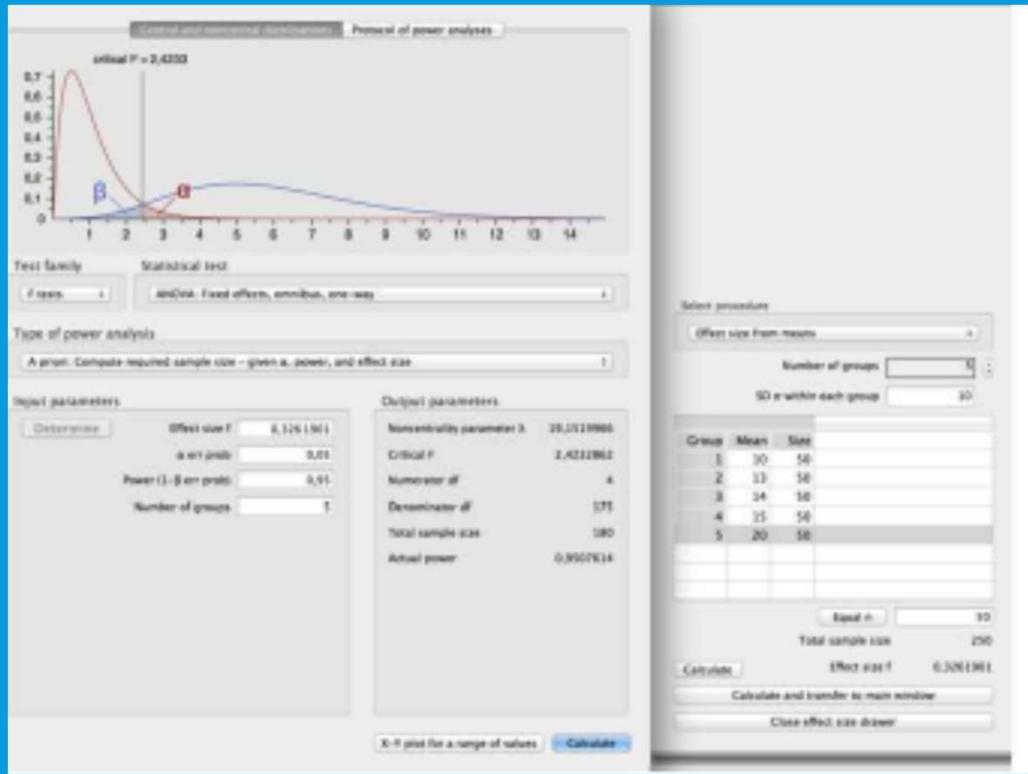
Commercial scientific software

- Mplus GLIMMs, Latent Models
- HLM GLIMMs

All models open source R packages

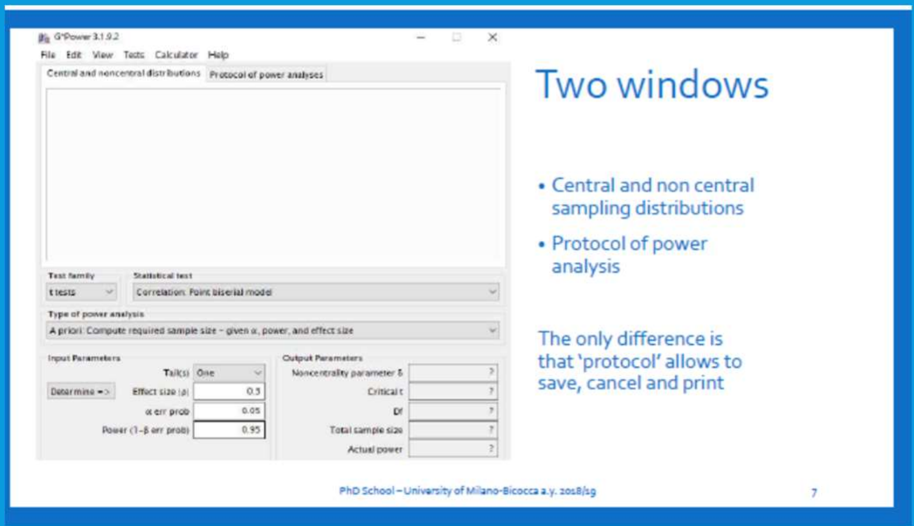
CONDUCTING A POWER ANALYSIS WITH G*POWER

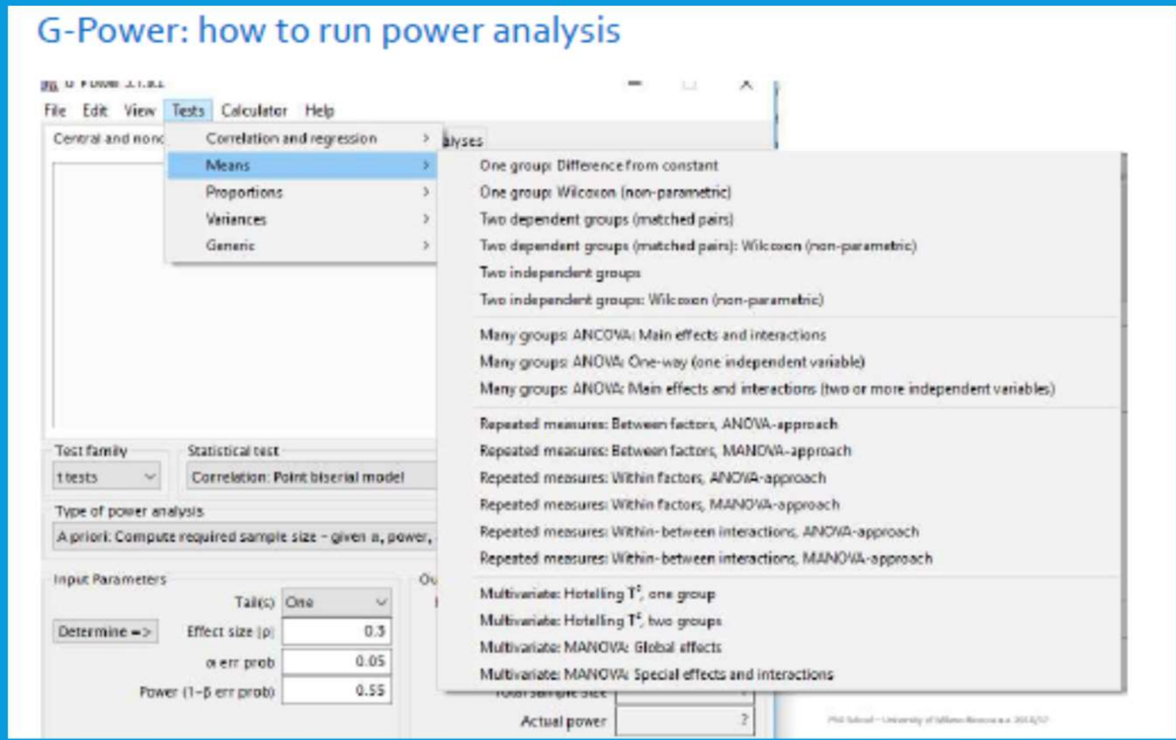
Detailed illustrations can be found on the related UCLA Idre web page: https://stats.idre.ucla.edu/other/gpower/



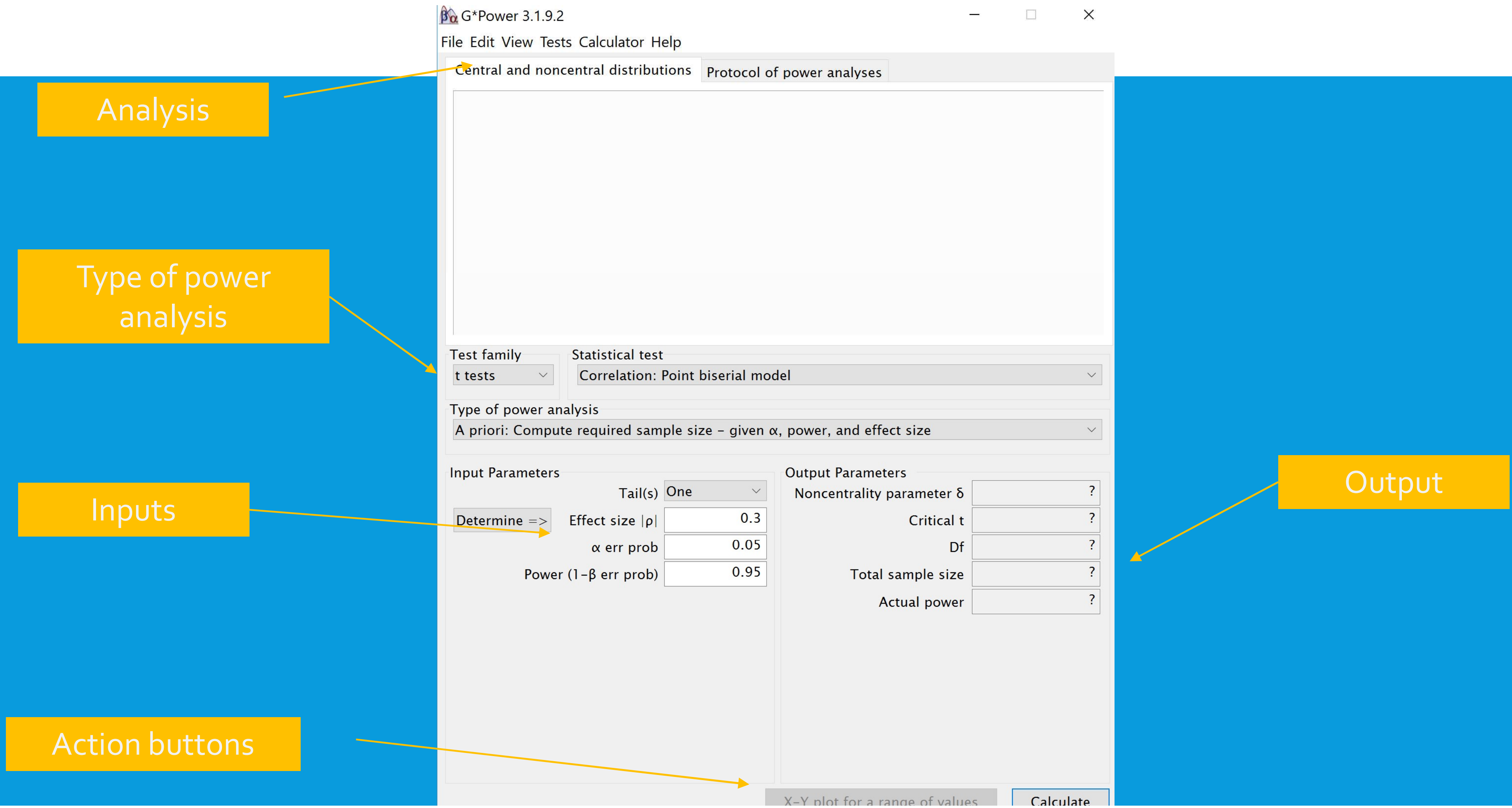
The UCLA idre web page provides also an introductory seminar to power analysis https://stats.idre.ucla.edu/other/mult-pkg/seminars/intro-power/

CONDUCTING A POWER ANALYSIS WITH G*POWER

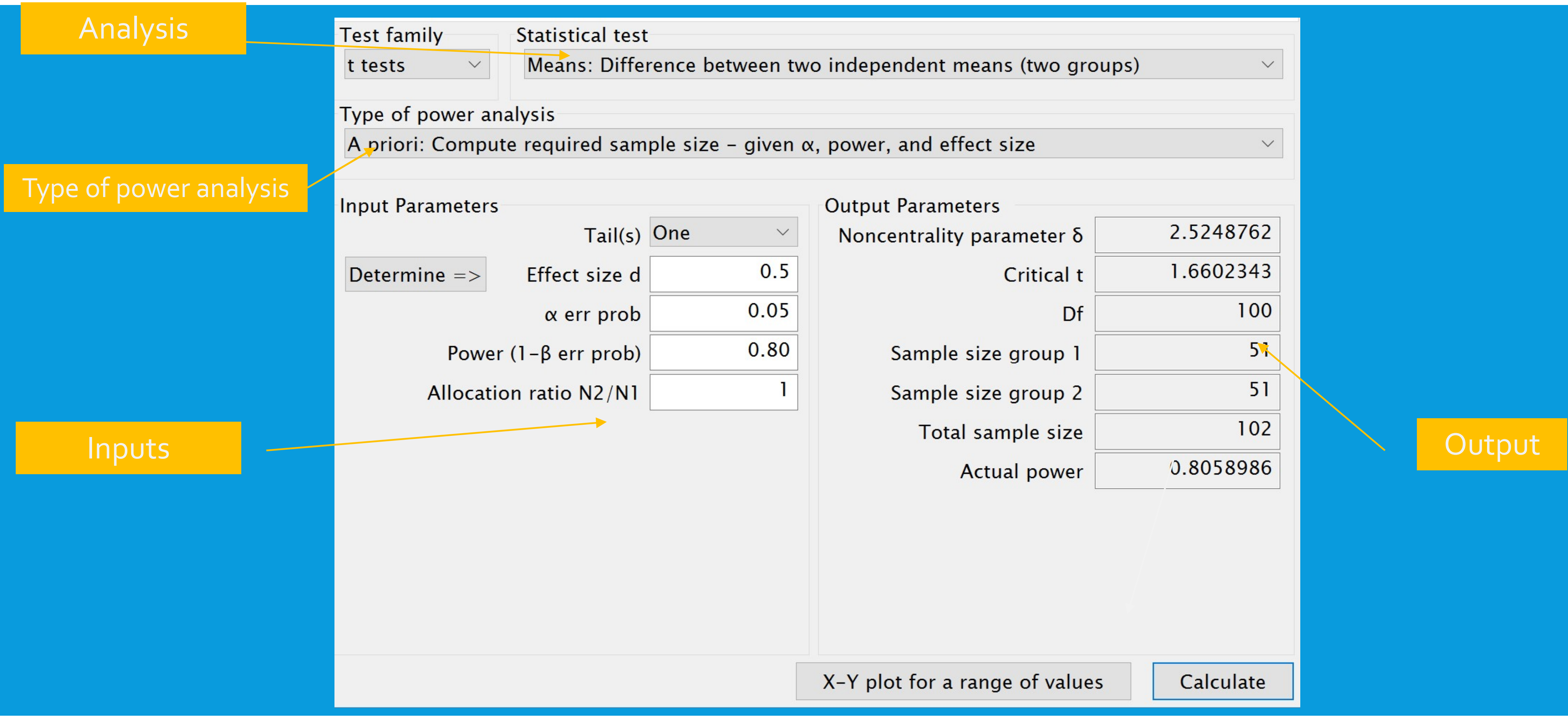




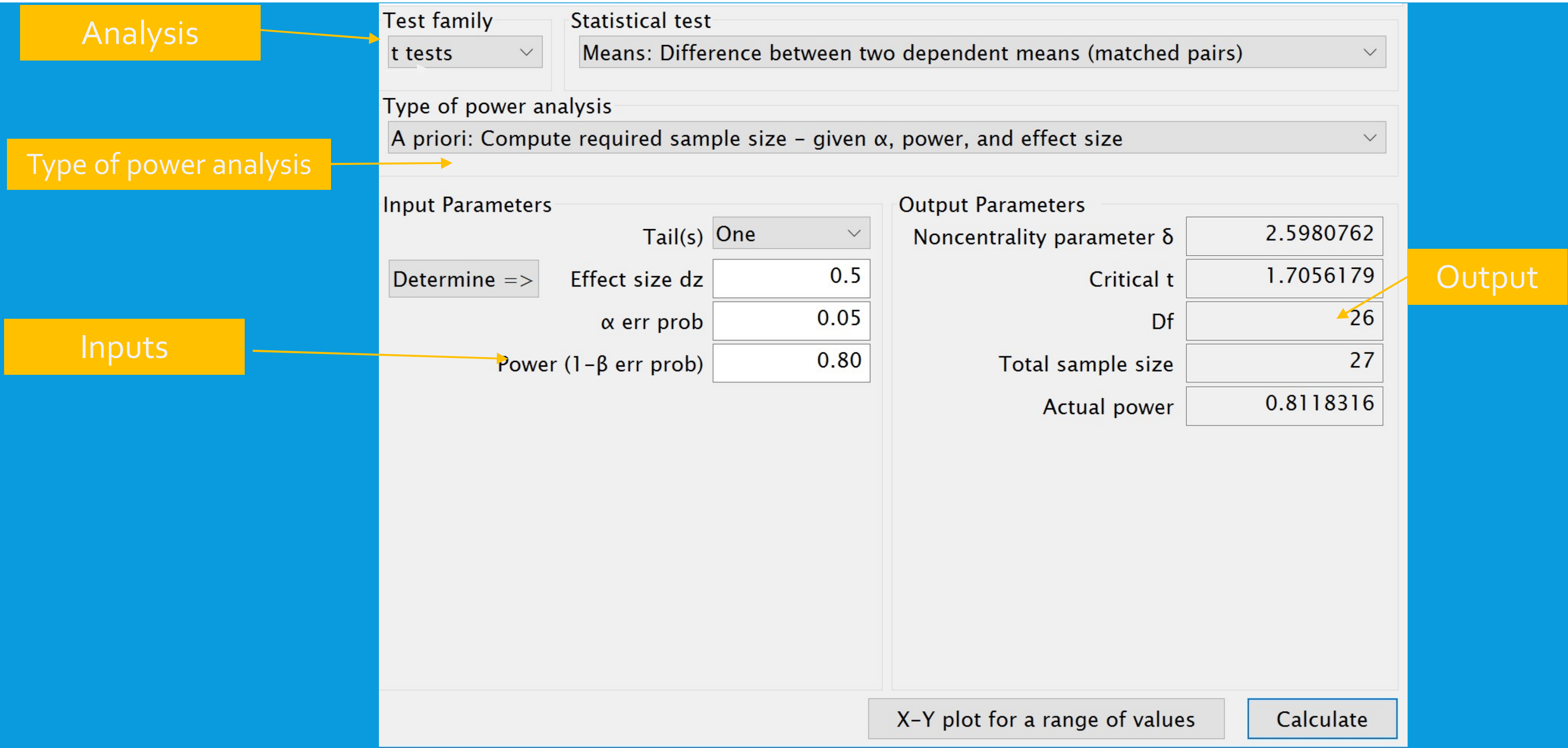
G*Power


Analysis
Type of power analysis
Inputs
Action buttons
Output
G*Power 3.1.9.2
File Edit View Tests Calculator Help
Central and noncentral distributions
Protocol of power analyses
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
Determine =>
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 ?
X-Y plot for a range of values
Calculate

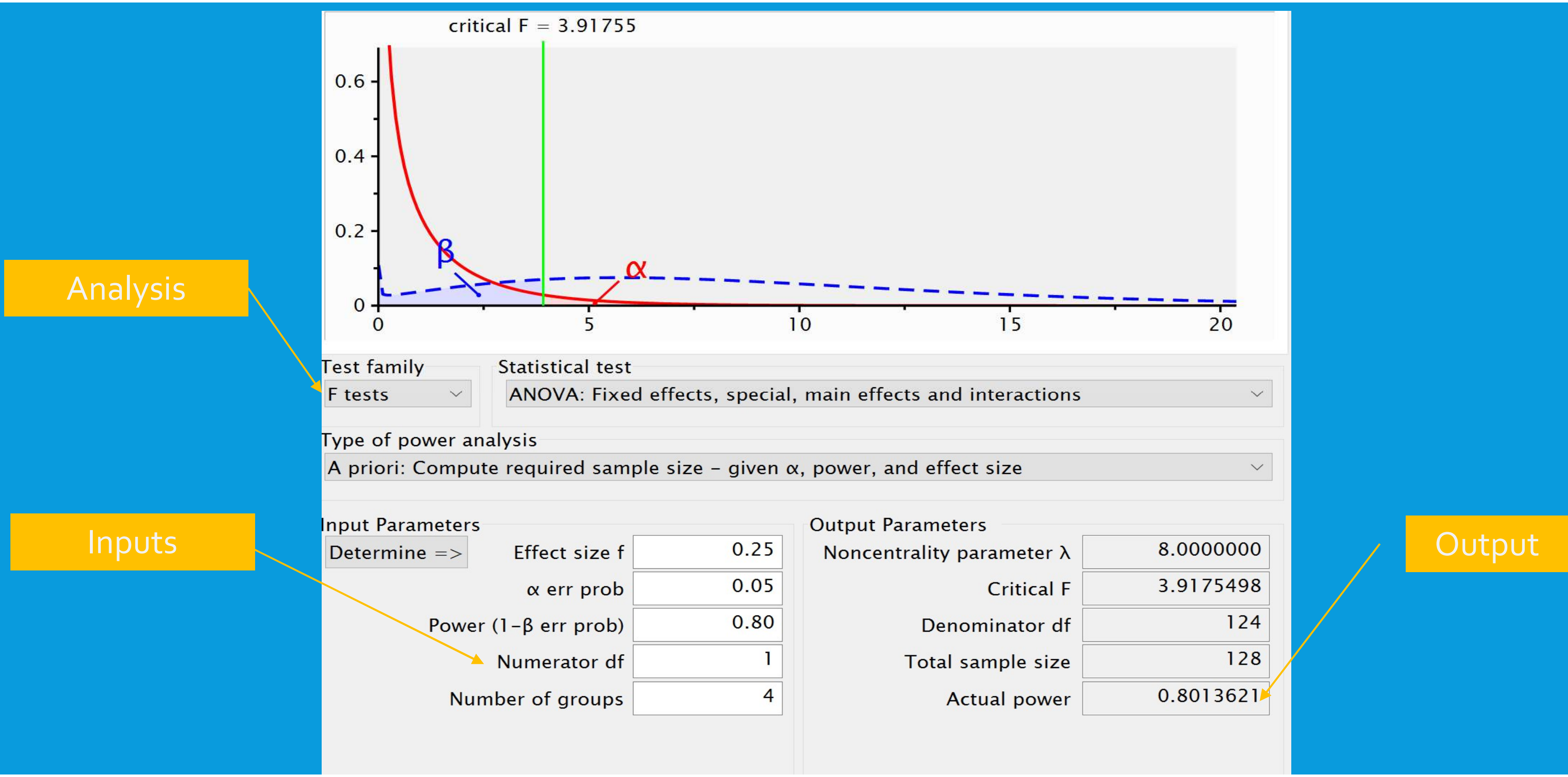
Two sample means



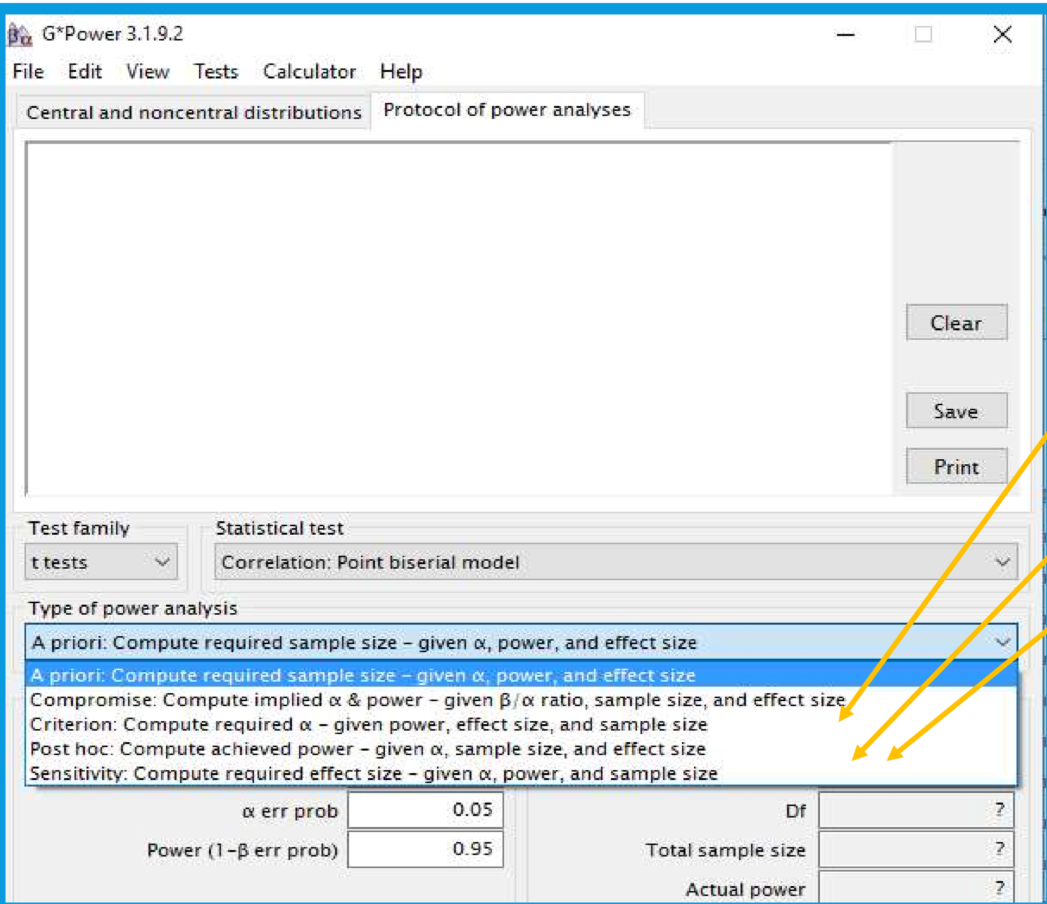
Example 2: Two repeated means



ANOVA 2 x 2


critical F = 3.91755
0.6
0.4
0.2
0
0
5
10
15
20
β
α
Analysis
Test family
F tests
Statistical test
ANOVA: Fixed effects, special, main effects and interactions
Type of power analysis
A priori: Compute required sample size – given α, power, and effect size
Inputs
Input Parameters
Determine =>
Effect size f 0.25
α err prob 0.05
Power (1–β err prob) 0.80
Numerator df 1
Number of groups 4
Output Parameters
Noncentrality parameter λ 8.0000000
Critical F 3.9175498
Denominator df 124
Total sample size 128
Actual power 0.8013621
Output

G-POWER: PROTOCOL OF POWER ANALYSIS. HAVE WE FOUND WHAT WE ARE LOOKING FOR?

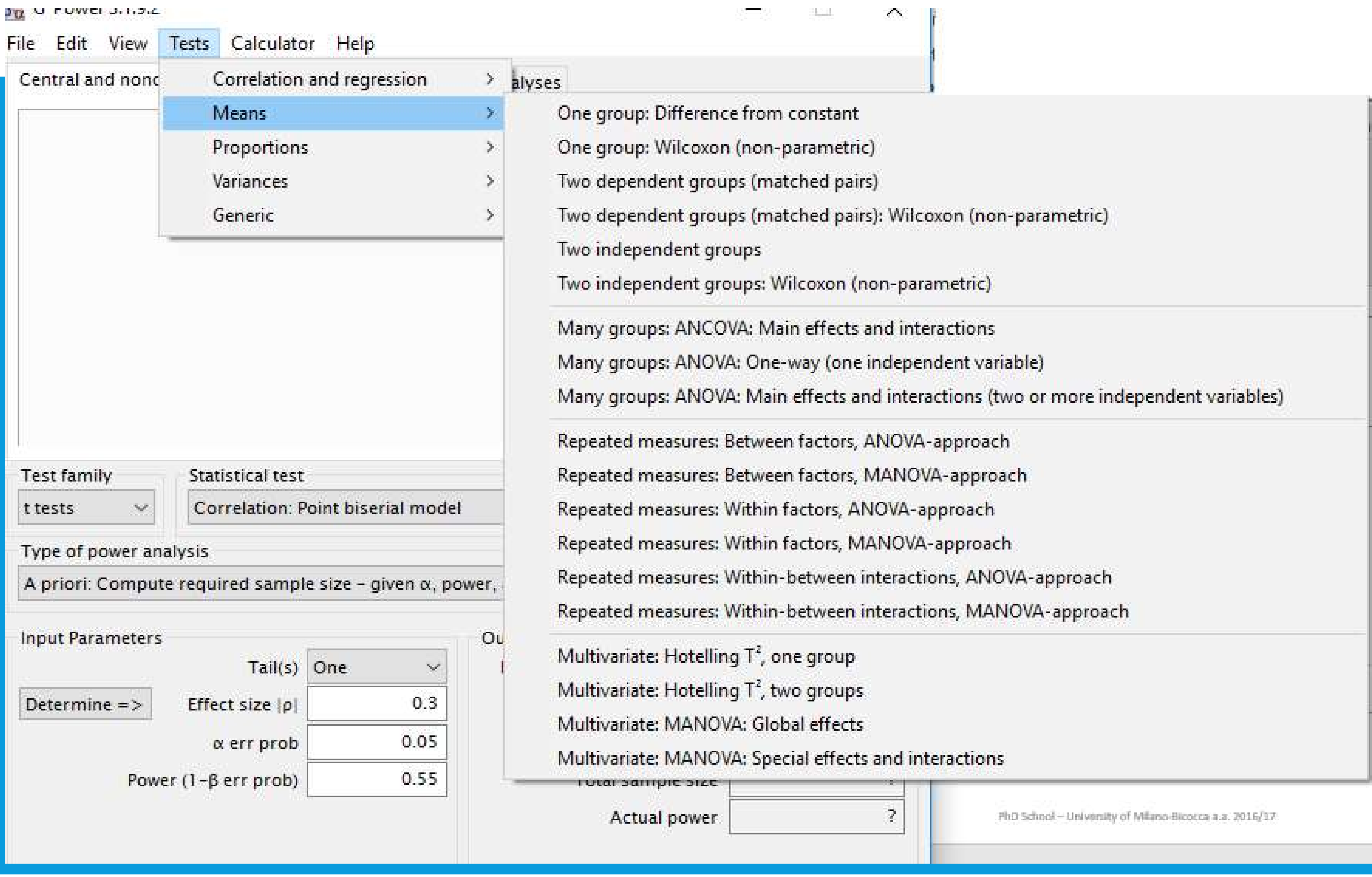


A priori: the 'ideal' way of determining power

Post hoc: the 'I could not help it' way of determining power

Sensitivity analysis: I have a feeling that there is an effect, but I do not have evidence so far.
How much do I have to increase my sample size to find (if I am wright) what I am looking for?

G-POWER: HOW TO CHOOSE THE GLIM



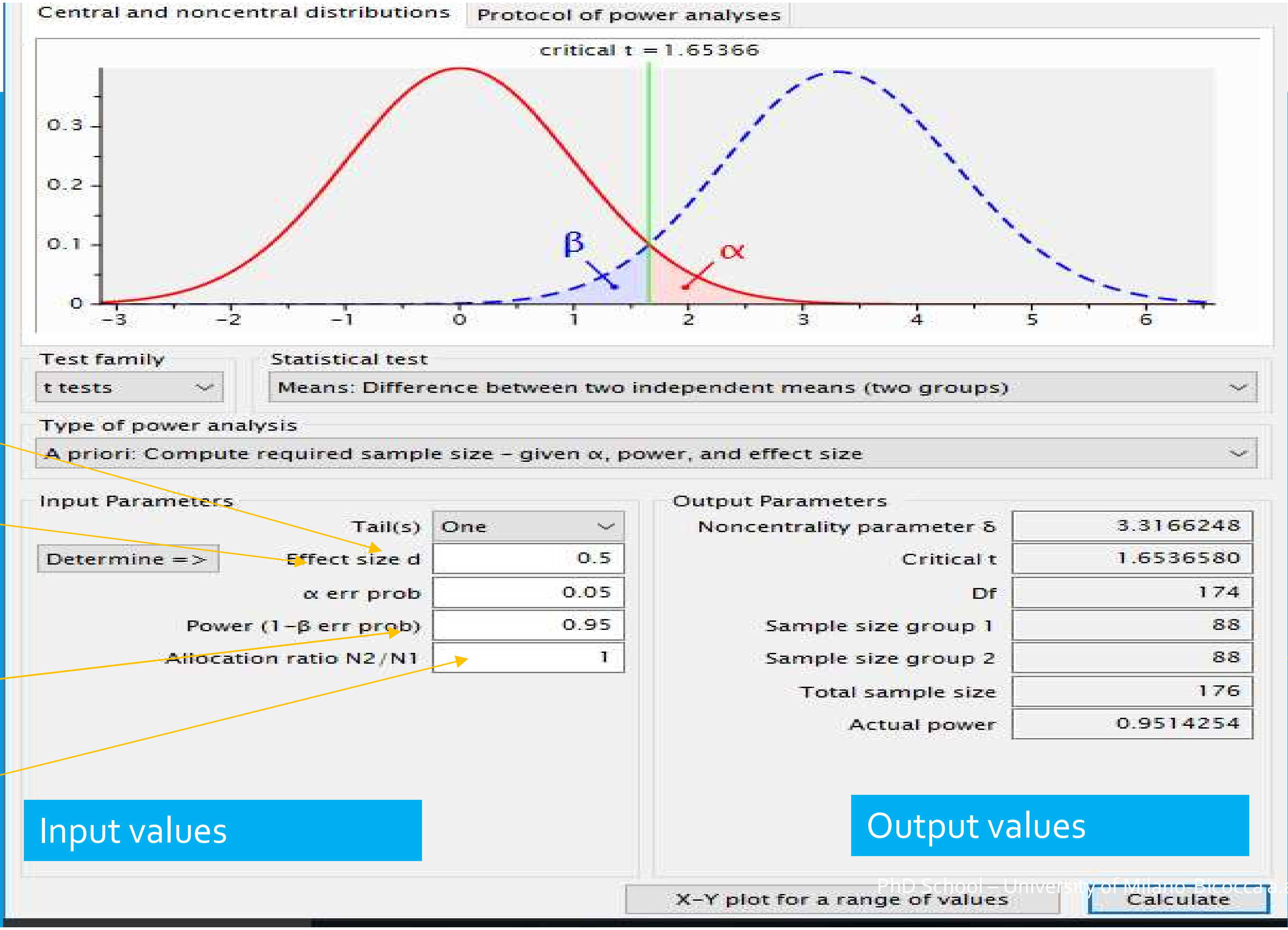
A COMPLETED EXERCISE : POWER ANALYSIS FOR TEST FOR TWO INDEPENDENT MEANS

Desired effect size d=0.5

Type I error, 0.05

Power wanted (in this case, very high, 0.95

Equal sample sizes,

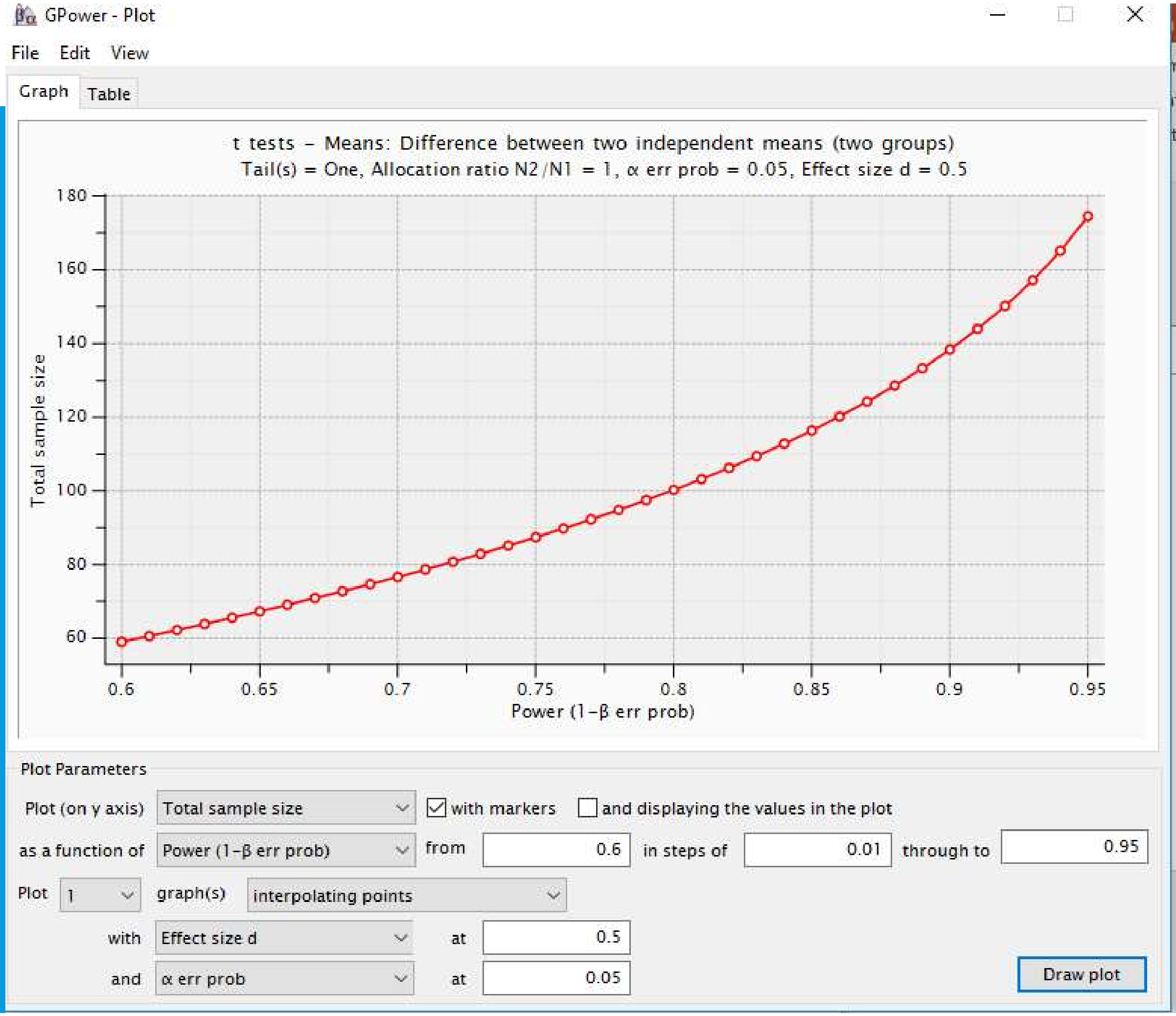


Input values

Output values

Sample sizes 88, in total 176 cases

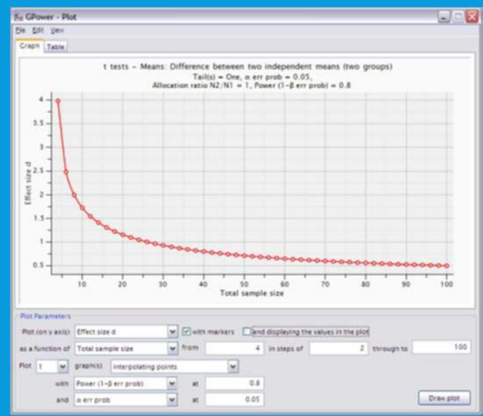
A COMPLETED EXERCISE: SENSITIVITY ANALYSIS FOR T TEST FOR TWO INDEPENDENT MEANS



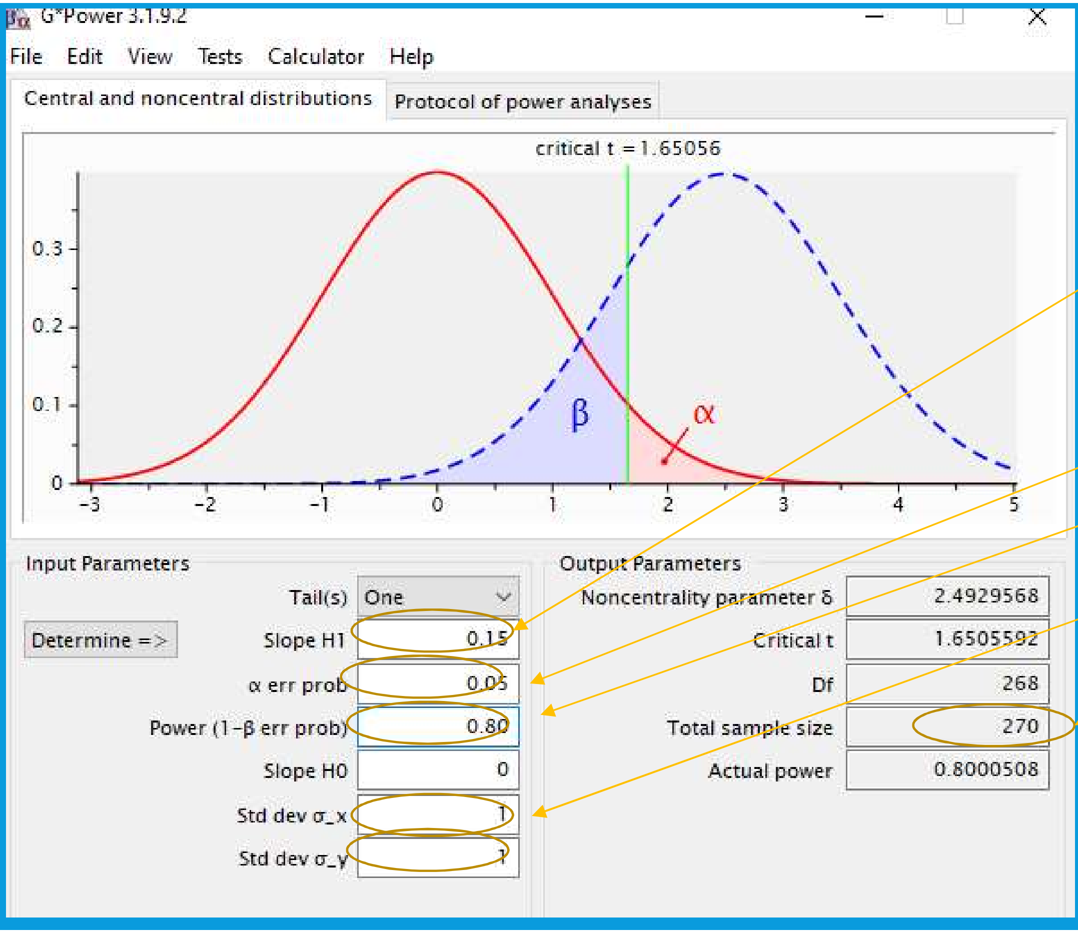
Same Gpower session as in the previous same analysis

Total sample size (*y* axis)for different power levels (*x* axis)

Below an example of curve for estimated ES as a function of the sample size n



POWER ANALYSIS: WHAT DO WE EXPECT IN A REGRESSION?



- Sample size in multiple regression
- Test for single coefficient (slope=0.15): bilateral t-test
- Alpha=.05
- Power .80 (minimum)
- SE for both variables

- Output: total sample size: 270

POWER ANALYSIS IN R

HTTPS://CRAN.R-PROJECT.ORG/WEB/PACKAGES/PAMM/PAMM.PDF

Search Results

The search string was **"pwr"**

Help pages:

pwr::pwr-package	Basic power calculations pwr
pwr::pwr.2p.test	Power calculation for two proportions (same sample sizes)
pwr::pwr.2p2n.test	Power calculation for two proportions (different sample sizes)
pwr::pwr.anova.test	Power calculations for balanced one-way analysis of variance tests
pwr::pwr.chisq.test	power calculations for chi-squared tests
pwr::pwr.f2.test	Power calculations for the general linear model
pwr::pwr.norm.test	Power calculations for the mean of a normal distribution (known variance)
pwr::pwr.p.test	Power calculations for proportion tests (one sample)
pwr::pwr.r.test	Power calculations for correlation test
pwr::pwr.t.test	Power calculations for t-tests of means (one sample, two samples and paired samples)
pwr::pwr.t2n.test	Power calculations for two samples (different sizes) t-tests of means

Random effects
ICC (effect size) in package sjstats
Power analysis for random effects in package pamm (and others)
effect size : Package 'effsize'

Simulations
Superpower in R
Lakens, D., & Caldwell, A. R. (2019). "Simulation-Based Power-Analysis for Factorial ANOVA Designs"

```
> pwr.t.test(n = 30, d = 0.5, sig.level = 0.05)

     Two-sample t test power calculation

              n = 30
              d = 0.5
      sig.level = 0.05
          power = 0.4778965
    alternative = two.sided

NOTE: n is number in *each* group
```

```
> pwr.t.test(d = 0.5, power = 0.80, sig.level = 0.05)

     Two-sample t test power calculation

              n = 63.76561
              d = 0.5
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

NOTE: n is number in *each* group
```

Very low power

'ex post', it is debatable theoretically for some methodologists.
We need at least .8 power (Lakens is requires .9!)

This is 'ex ante', we need 64 subjects in each group

d: effect size
n= sample size
We can either specify directly d value or the 'intended' difference between means D and the pooled standard deviation

```
> power.t.test(delta = 0.50, sd = 2.25, sig.level = 0.05, power = 0.8)

     Two-sample t test power calculation

              n = 318.8428
          delta = 0.5
             sd = 2.25
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

NOTE: n is number in *each* group

>
```

Now, we need a different es. We know that the pooled sd is 2.25, we need to detect a difference between the means equal or smaller than .50. This difference is called delta in library pwr

We need at least 638 subjects

Power curves (power functions)

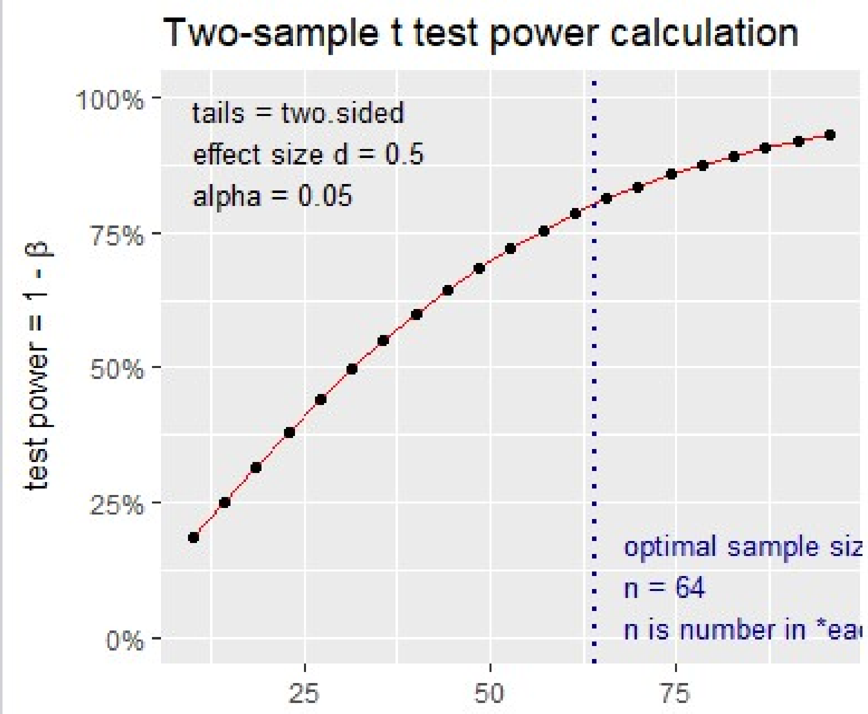
```
> power.t.test(delta = 0.50, sd = 2.25, sig.level = 0.05, power = 0.8)

     Two-sample t test power calculation

              n = 318.8428
          delta = 0.5
             sd = 2.25
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

NOTE: n is number in *each* group

> p.t.two <- pwr.t.test(d=0.5, power = 0.8, type= "two.sample", alternative = "two.sided")
> plot(p.t.two)
> plot(p.t.two, xlab="sample size per group")
>
```



pwr - developed by Stéphane Champely- power analysis as outlined by Cohen (1988)

- https://www.statmethods.net/stats/power.html

- powerAnalysis: power for experimental design https://cran.r-project.org/web/packages/powerAnalysis/powerAnalysis.pdf
- simr - Power Analysis for Generalised Linear Mixed Models (lme4) by Simulation https://cran.r-project.org/web/packages/simr/simr.pdf

CONSEQUENCES OF LOW POWER

- Low probability of finding true effects: ow power means that the chance of discovering effects that are really true is low. Low-powered studies produce more false negatives than high-powered studies.
- Low positive predictive value : the lower the power of a study, the lower the probability that an observed “significant” effect (among of all significant effects) actually reflects a true non-zero effect in the population (vs. a false positive).
- When an underpowered study discovers a true effect, it is likely that the estimate of the magnitude of that effect will be exaggerated. Effect inflation is worst for small, low-powered studies, because they can only detect sample parameter estimates effects when they are large.

ENJOY BEING POWERFUL!