10th International Course
Care and Use of Laboratory Animals
May 23, 2024

Experimental design
Statistical analysis and Interpretation of results

Konstadia (Dina) Lika
Biology Department
Univ of Crete
lika@uoc.gr

ΠΑΝΕΠΙΣΤΗΜΙΟ
ΚΡΗΤΗΣ

UNIVERSITY
OF CRETE

Experiments using laboratory animals

- Well designed
- Efficiently executed
- Correctly analyzed
- Clearly presented
- Correctly interpreted

Experiment:
scientific procedure undertaken
- to test a hypothesis (confirmatory research) or
- to provide material for the generation of new hypotheses (exploratory research)

Ethical consideration

Principles of the 3Rs (Russel & Burch, 1959)

- Replacement
  - Replace animals by less sentient ones or in vitro methods whenever possible
- Refinement
  - Experimental protocols should be refined to minimize adverse effects
- Reduction
  - Keep the number of animals to the minimum necessary to achieve the aims of the research
  - BUT NOT so few
    - miss biologically important effects or
    - require unnecessary repetition of experiments or
    - the results become statistically invalid

The # of animals used can be reduced by

- **good exp. design**
- **appropriate statistical procedures**

Research description

Clearly state:

- the objectives of the research and the hypotheses to be tested
- the reason for choosing the particular animal model
  - the species, strain, source, and type of animal used
- the details of each experiment
  - experimental design
  - number of animals
- the statistical methods used for analysis

The experimental unit (EU)

- the physical entity which can be assigned, at random, to a treatment
- any two EUs must be capable of receiving different treatments
- commonly (not always) it is an individual animal
- it is also the unit of statistical analysis

Example- mice in a cage

- 6 cages, 5 mice/cage
- 2 treatments, each applied to 3 cages

What is the EU? A mouse or a cage?

Answer: A cage of animals rather than the individual animal

- because treatments are applied to whole cages
- mice in the cage can not have different treatments

Principles of experimental design

- Replication
- Randomization
- Controls
  - Compare results and establish baseline data
  - Minimize confounding effects
- Blinding
  - Animals, samples and treatments should be coded until the data are analyzed
  - Prevent bias from the experimenter in data collection and analysis

Replication

- The process of applying each treatment to more than one EU

- Replicate
  - independent EU to which the treatment is applied to

- Replication allows
  - estimation of variation in population
  - a more *reliable* estimation of the effect of each treatment

What is true replication in the mice in the cage example?

n=3

m=15 mice (pseudo-replication)

Pseudo-replication

- Pseudo-replicates are not independent, while standard statistical analyses assume independence
- Pseudo-replication leads to an incorrect estimate of between unit variability
- Pseudo-replication can either be:
  - temporal, involving repeated measures over time from the same mice or cage
  - spatial, involving several measurements from the same vicinity or cage
- Solution: use the appropriate statistical tool

Randomization

- The process of randomly allocating EUs to treatment groups
- Each EU has equal probability of receiving a given treatment
- Randomization ensures
  - any bias of the experimenter is avoided
  - estimates of population parameters (mean, treatment effects, etc) are unbiased
  - statistical inference (conclusions from statistical tests) are reliable
- Random numbers (generated from statistical packages) can be used to assign EUs to treatments

Formal experimental designs (1)

- Completely randomized designs (CR)
  - EUs are assigned to treatments at random
- Completely randomized block designs (CRB)
  - EUs ("replicates") are grouped into blocks (uniformity within a block)
  - reduce unexplained variation, without increasing size of experiment

Formal experimental designs (2)

- Repeated measure designs
  - each EU measured repeatedly under different treatments and/or times

- Factorial designs
  - more than one type of treatment (e.g. drug treatment and gender)
  - crossed designs: every level of one factor crossed with every level of a second factor
  - nested designs: different (randomly chosen) levels of a factor B nested in each level of a factor A

Experimental Size

- Small: may under-detect the effect of interest in your experiment
- Large: may lead to unnecessary wasting of resources and animals

How many animals do I need for my experiment?

Goal: We strive to have enough samples to reasonably detect an effect, if it really is there, without wasting limited resources on too many samples.

- **Power analysis** is the most common way of determining sample size

Power analysis

Can be used

- Design an experiment
  - determine the sample size required to detect an effect of a given size with a given degree of confidence

- Make *a posteriori* assessment of the usefulness of an experiment
  - determine the probability of detecting an effect of a given size with a given level of confidence, under sample size constraints

Statistical Analysis

General aim is to extract all useful information present in the data

- **Descriptive statistics**
  - statistical methods to summarize, describe or explore a collection of data
  - enable us to present the data in a more meaningful way, allowing a simpler interpretation of the data
- **Inferential statistics**
  - The process of drawing conclusions from sample data about the process or population being studied
  - techniques allowing the use of samples to make generalizations about populations

Descriptive Statistics

Graphical exploration of data

- graphical methods summarize the data in a diagrammatic way
- qualitative and involve a degree of subjectivity

Non-graphical method

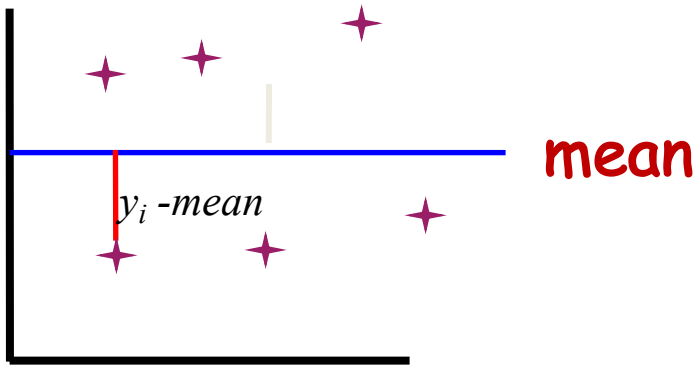
- summary statistics
- quantitative and objective,
  they do not give a full picture of the data

Non-graphical and graphical methods complement each other

Non-graphical methods - Examples

Mean : describes the location of a distribution

Variance : captures its scale or degree of being spread out



$$mean: \bar{y} = \frac{\sum y_i}{n}$$

$$variance: s^2 = \frac{\sum (y_i - \bar{y})^2}{n-1}$$

$$deviance = SS = \sum (y_i - mean)^2$$

$$standard\ deviation: s = \sqrt{\frac{\sum (y_i - \bar{y})^2}{n-1}}$$

Inferential Statistics

- Parateter estimation
- Hypothesis testing

Parameter Estimation

- Parameters of statistical model (unkown)
  - mean, variance, regression slope etc.
- Use sample data to estimate those parameters
  - sample statistic: function of the data drawn from the population
  - sample statistics estimate population parameters

Point estimate: single value estimate of parameter

e.g. $\bar{y}$ (sample mean): point estimate of pop mean μ

s^2 (sample variance): point estimate of pop variance σ^2

Interval estimate: range within which parameter lies, with some degree of confidence

Hypothesis testing (1)

- Specify research hypothesis and null hypothesis

  e.g.

  Question: Do female and male mice have the same metabolic rate?

  H_0 : there is no difference between the sexes in the mean metabolic rate

- Choose test statistic

  – *t* for mean, *F* for variances etc.

- Collect sample data

- Calculate test statistic relevant to hypothesis

  – *t*, *F*-ratio, etc.

Common error: Sampling before
- constructing the hypothesis
- choosing the test analysis

Hypothesis testing (2)

- **Determine the p-value**

  Probability that any given experiment will produce *a value of the chosen test statistic* that is equal to the one observed in our actual experiment or something more extreme, assuming that the *null hypothesis* is true

  – assumptions of the test are (reasonably well) met

- **Interpretation of p-value**

  – measures the "strength of evidence" against H_0

  - not the probability that H_0 is true!

  - not the probability of making a mistake by rejecting a true H_0

Decision criterion

Compare p-value to the *a priori* significant level (α)

- if $p<\alpha$, conclude H_0 is "unlikely" to be true and reject it (statistically significant result)
- else, conclude H_0 is "likely" to be true and do not reject it (statistically non- significant result)

Convention sets significant level $\alpha=0.05$ (5%)

Arbitrary: other significant levels might be valid (e.g. 0.01, 0.001)

Decision errors

Statistical hypothesis tests can produce decision errors

Type I error
– rejecting a true H_0
– probability of making this error is set by significance level α

Type II error
– not rejecting a false H_0
– probability of making this error can only be determined if variability and desired detectable difference is known

- The risks of these two errors are inversely related
- Increasing sample size is the best way to minimise both errors

Power of test

- Probability of detecting an effect if it exists
- Probability of rejecting incorrect H_0
- Complement to a Type II error
  - If β is the probability of making a type II error,

    1-β (the power) is the probability of not making a type II error

Statistical power (1)

To calculate the power of a statistical test you need to specify:

- Effect size (ES)
  - differences between treatments
  - large effects easier to detect

- Background variation
  - variation between experimental units (s^2)
  - greater background variability; less likely to detect effects

Statistical power (2)

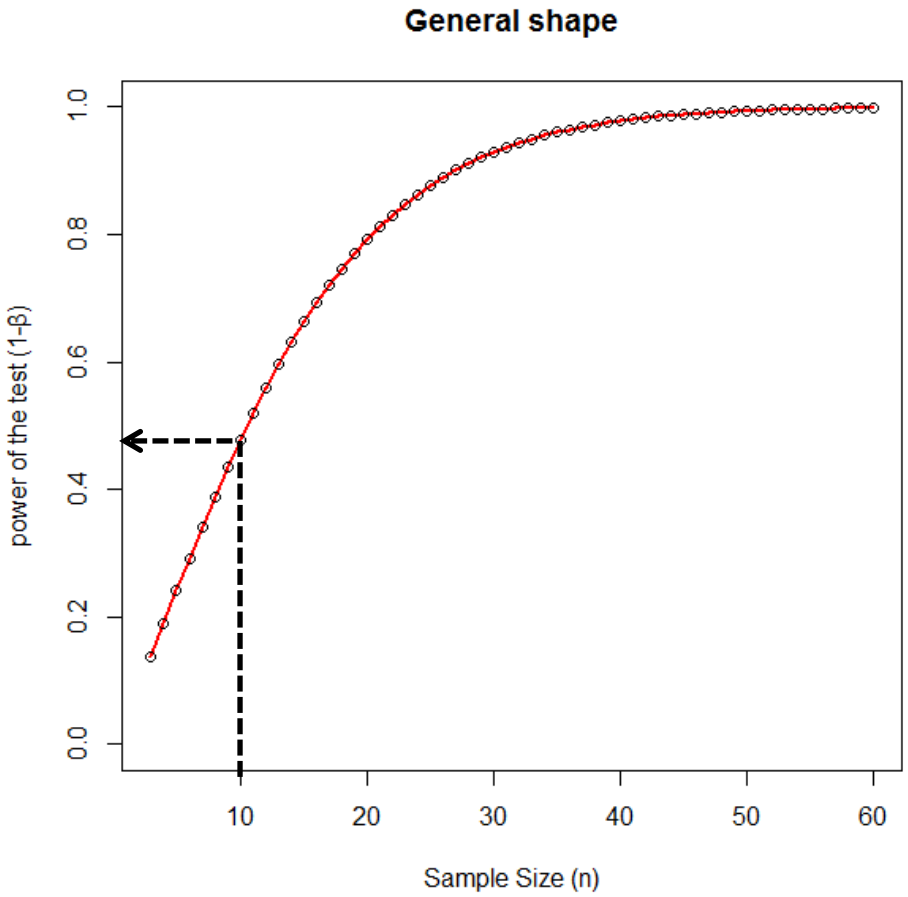
- Sample size (n) for each treatment group
  - increasing sample size makes effects easier to detect
- Significant level (α)
  - Probability of Type I error
  - usually set at 5%, lower values sometimes specified
  - as α decreases, β increases, power (1- β) decreases
- Alternative hypothesis

Power analysis

a posteriori power

For most types of analysis (t-test, ANOVA, regression), the **power** (1-β) is:

$$(1-\beta) \propto \frac{ESa\sqrt{n}}{s}$$



Sample size determination

To determine appropriate sample size we need to:

- solve power equation for n
- know background variation

  (from pilot studies/previous literature)
- know the statistical power (1-β) we want
- know what ES we wish to be able to detect if it occurs

$$\sqrt{\mathrm{n}} \propto \frac{s(1-\beta)}{aES}$$

Statistical Power
Conventions and Decisions

- Acceptable risk of a Type II error is often set at 1 in 5, i.e., a probability of 0.2 (β)

- "adequate" statistical power is therefore set at

  $1-\beta = 1 - 0.2 = 0.8$

Effect Size

- While **power (1-β)** and **Significance level (α)** are set irrespective of the data, the effect size is a property of the sample data
- *ES* formulas depend on statistical test

- Depending on the actual test, the *ES* may be expressed as
  - *d* (difference between two means),
  - *r* (correlation between two variables)
  - *f* (ANOVA test)
  - any other index related to specific test

Cohen, J., (1977). *Statistical power analysis for the behavioural sciences*. San Diego, CA: Academic Press.
Cohen, J., (1992). A Power Primer. *Psychological Bulletin* **112** 155-159.

Calculating Cohen's *d*

Effect size $$d = \frac{\bar{x}_1 - \bar{x}_2}{s_{Pooled}}$$

$$s_{Pooled} = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

$\bar{x}$: mean
s: standard deviation
n: sample size
Subscript refers to the two conditions compared

How Do We Measure Effect Size?

- Use background information in the form of preliminary/trial data to get means and variation, then calculate effect size directly

- Use background information in the form of similar studies to get means and variation, then calculate effect size directly

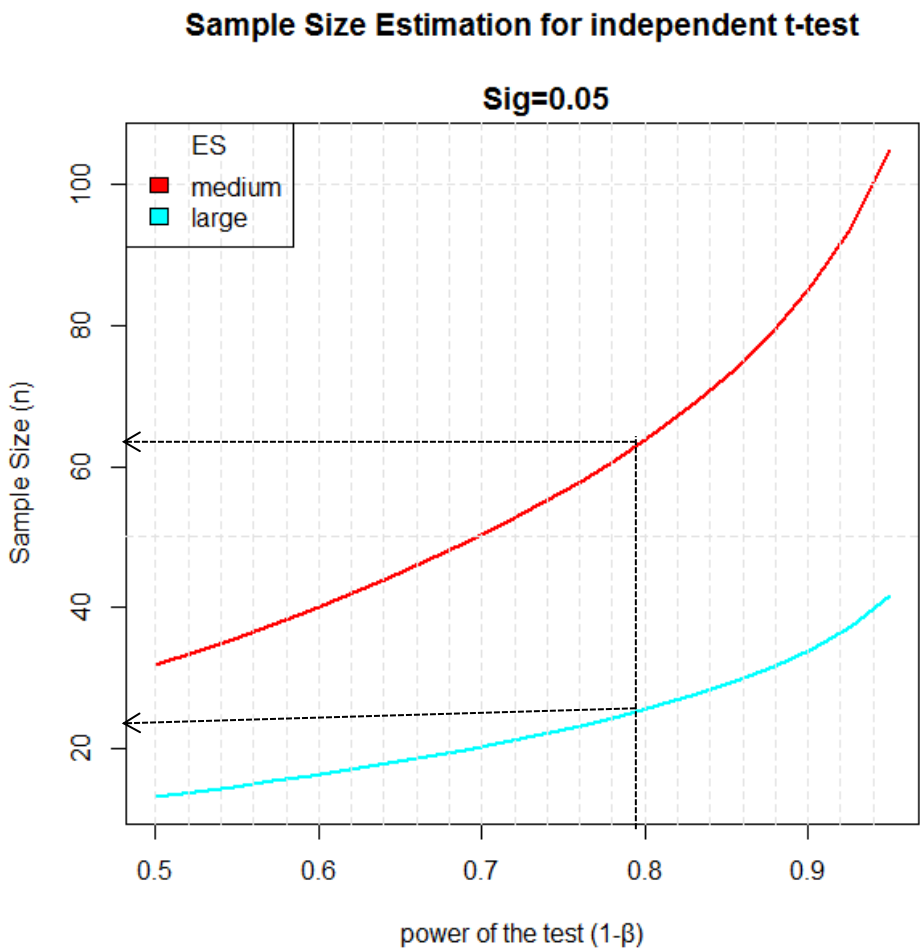
- With no prior information, make an estimated guess on the expected effect size
  - Broad effect sizes categories are small, medium, and large
  - Different statistical tests will have different values of effect size for each category

Cohen's Rules Of Thumb For Effect Size

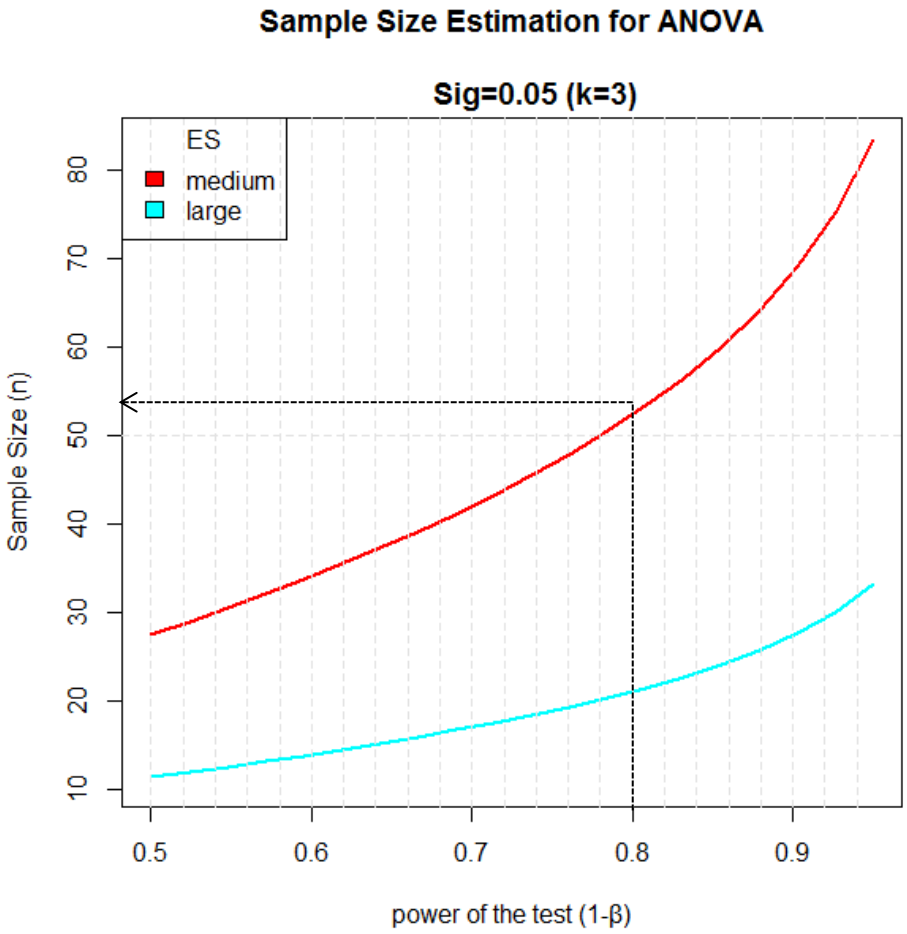
Effect size	Correlation coefficient	t-tests	ANOVA
“Small effect”	r = 0.1	d = 0.2	f=0.1
“Medium effect”	r = 0.3	d = 0.5	f=0.25
“Large effect”	r = 0.5	d = 0.8	f=0.4

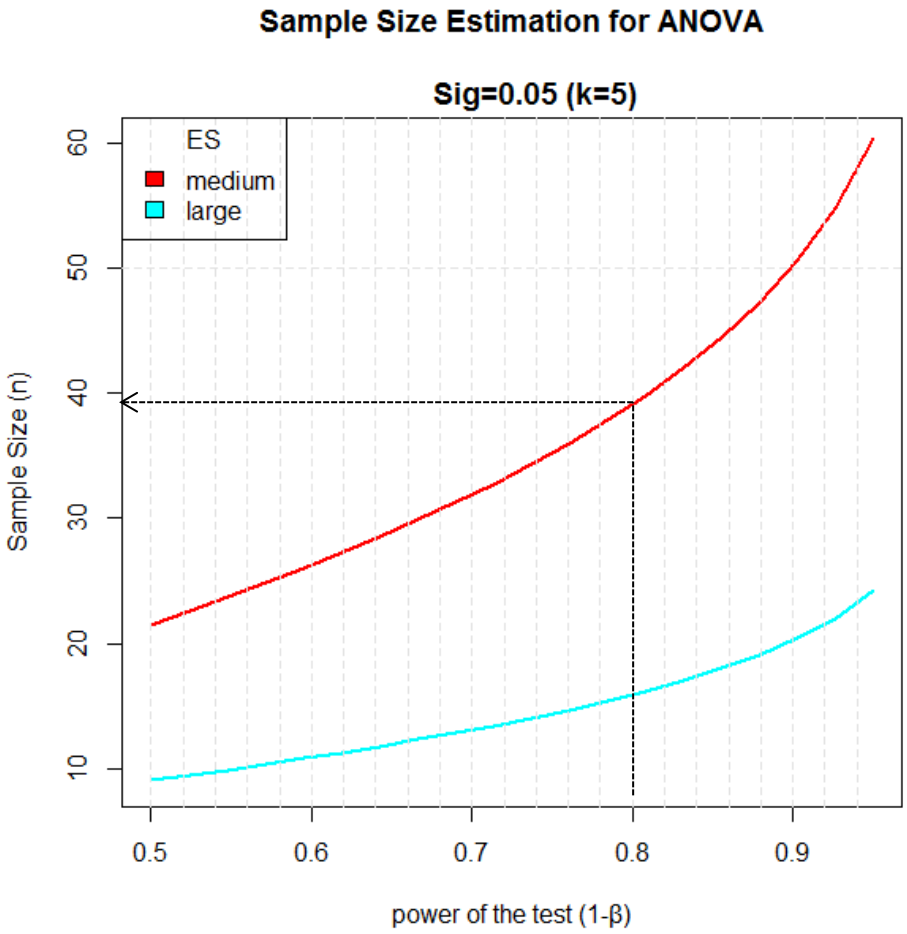
Cohen’s suggestions should be seen as rough guidelines.

Required sample size for t-tests



Required Sample size for one-way ANOVA





Software for power analysis

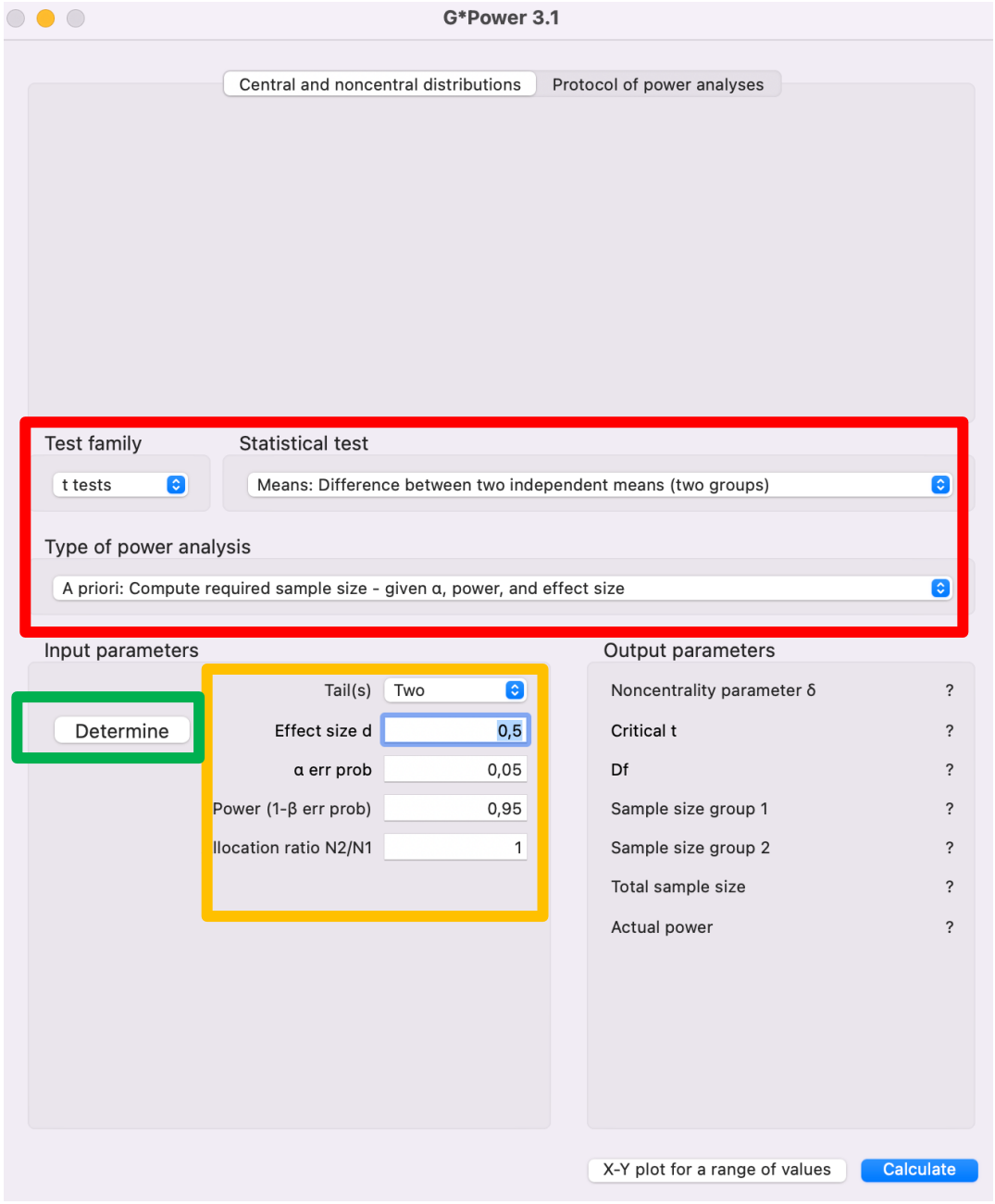
- G*Power
- R (e.g., package ‘pwr’)

G*Power
Main window

[Download GPower]

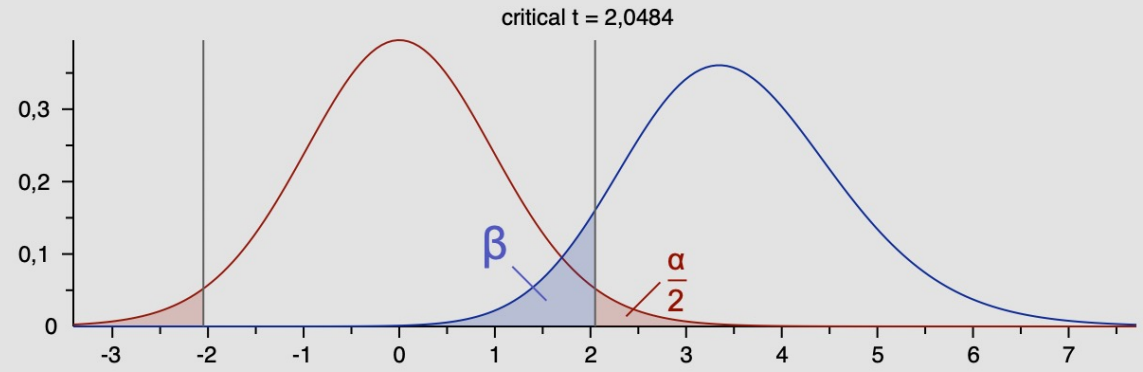
Three basic steps:

- Select appropriate test
- Input parameters
- Determine effect size (can use background info or guess)



Determine effect size (can use background info or guess)

Central and noncentral distributions | Protocol of power analyses



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

Determine

Tail(s)	Two
Effect size d	1,252769
α err prob	0,05
Power (1-β err prob)	0,9
Allocation ratio N2/N1	1

Output parameters

Noncentrality parameter δ	3,4308492
Critical t	2,0484071
Df	28
Sample size group 1	15
Sample size group 2	15
Total sample size	30
Actual power	0,9116579

X-Y plot for a range of values | Calculate

n1 ≠ n2

Mean group 1	0
Mean group 2	1
SD σ within each group	0,5

n1 = n2

Mean group 1	8,7
Mean group 2	11,4
SD σ group 1	2
SD σ group 2	2,3

Calculate | Effect 1,252769

Calculate and transfer to main window

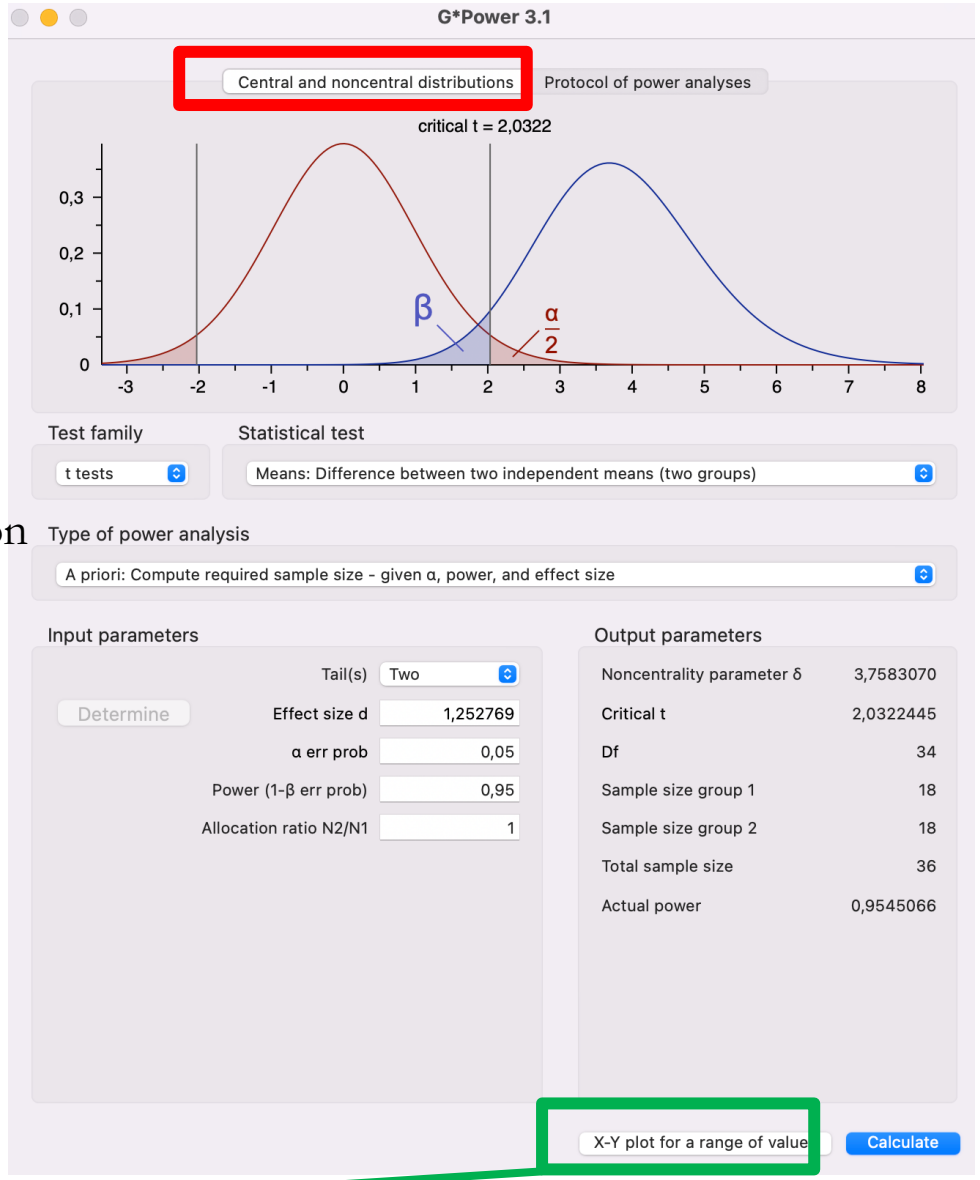
Close effect size drawer

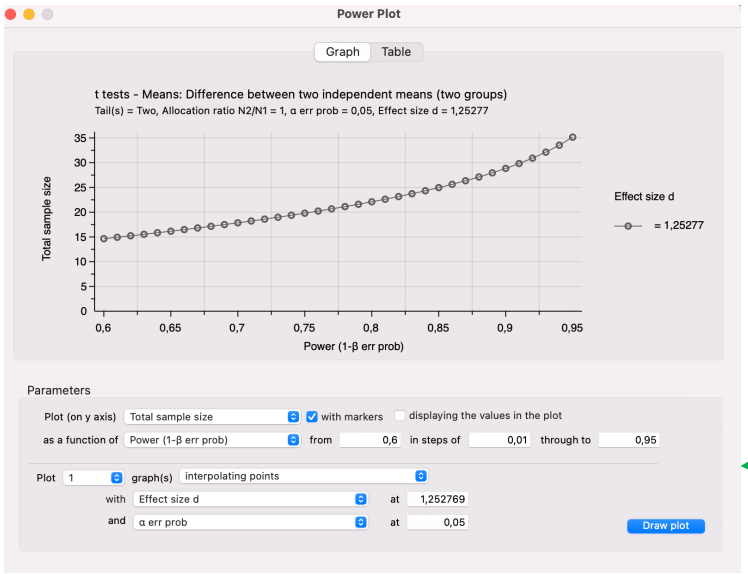
Central and noncentral distributions

Shows the distribution of the null hypothesis (red) and the alternative (blue)

X-Y plot for a range of values

Generates plots of one of the parameters α, effect size, power and sample size, depending on a range of values of the remaining parameters





Exercise 1

The way productive animals are killed concerns the scientific community and society for both bioethical and productive reasons. The aim of the project is the evaluation of used methods of capturing and killing fish.

The evaluation of stress includes measuring cortisol concentrations in sea bass, *Dicentrarhus labrax* between:

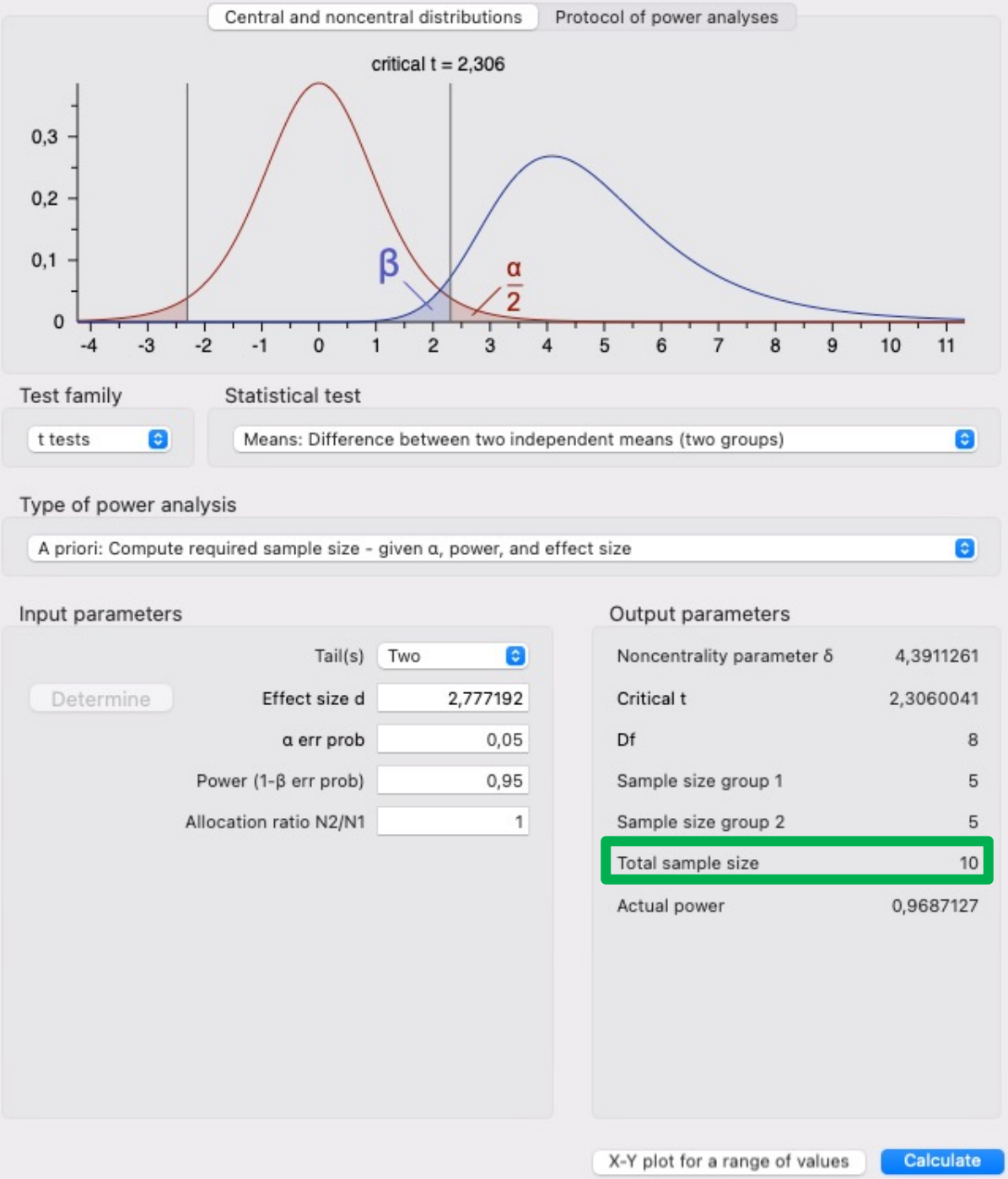
A) Two different ways of capturing the fish

(i) by using hook & line without any other handling (considered to be the fastest method of conception) and

(ii) by the common way of capture (synchronization, netting, exposure to air)

You are interested in determining if the average cortisol concentration differs between treatments. (Determine the sample size you will use)

Cortisol preliminary data: (mean ± SD) Hook & line: 101.5 ± 30.6 ng/ml and Common way: 478.7 ± 95.1 ng/ml



Answer:

A total of 10 fish are needed (5 per group)

Exercise 2

B) Four different ways of killing

(i) by puncturing the spinal cord using Ikigun, (https://www.ikigun.com – considered to be the fastest way to kill) after anaesthesia with benzocaine;

(ii) by immersion in ice slurry (heat shock and anoxia) without prior use of anesthesia (the method used in fish farms);

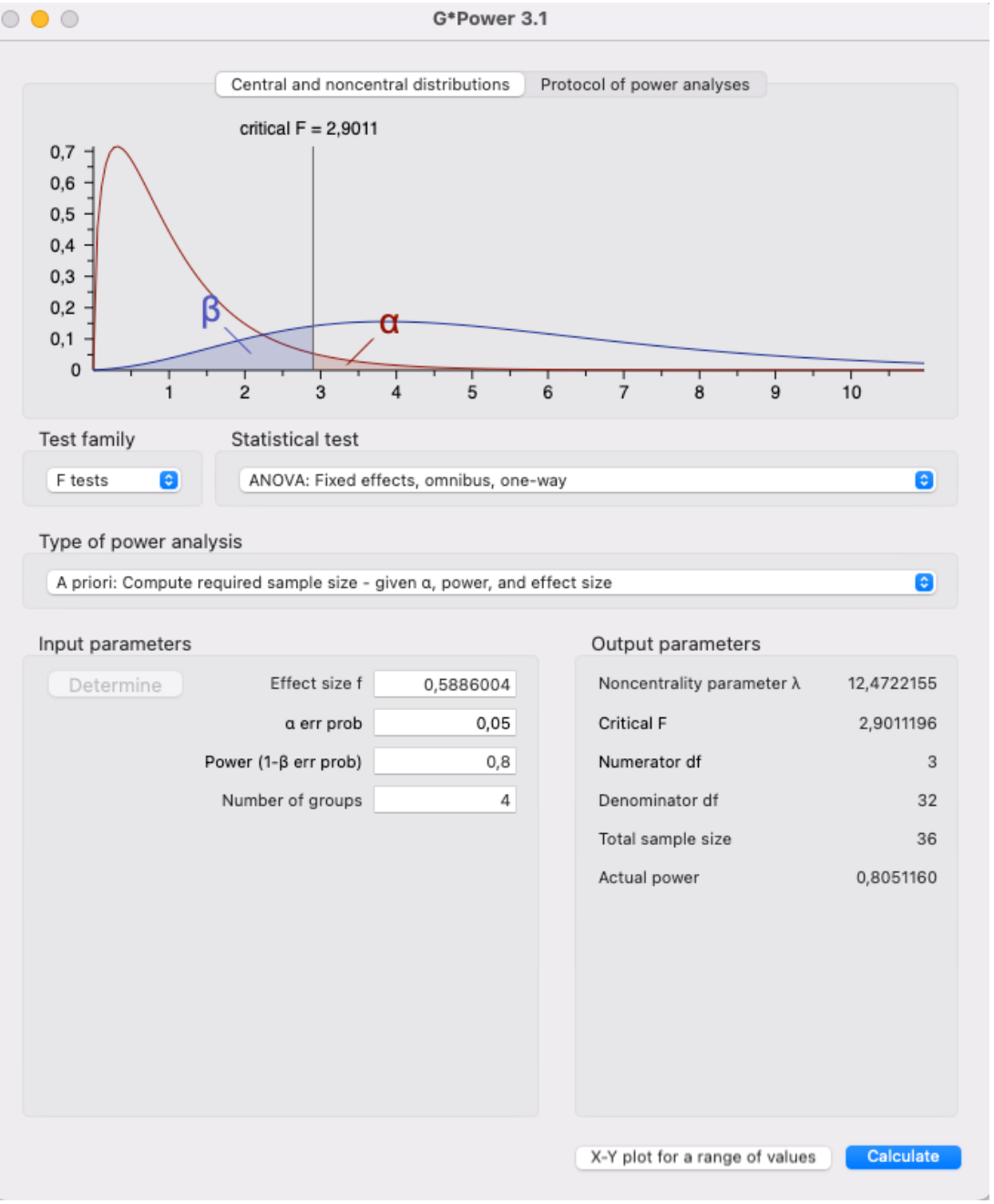
(iii) with prolonged exposure to chemical anesthesia with benzocaine and

(iv) using electroanesthesia followed by immersion in ice water (recommended new humane method of killing in fish farming).

You are interested in determining if the average cortisol concentration differs between treatments. (Determine the sample size you will use)

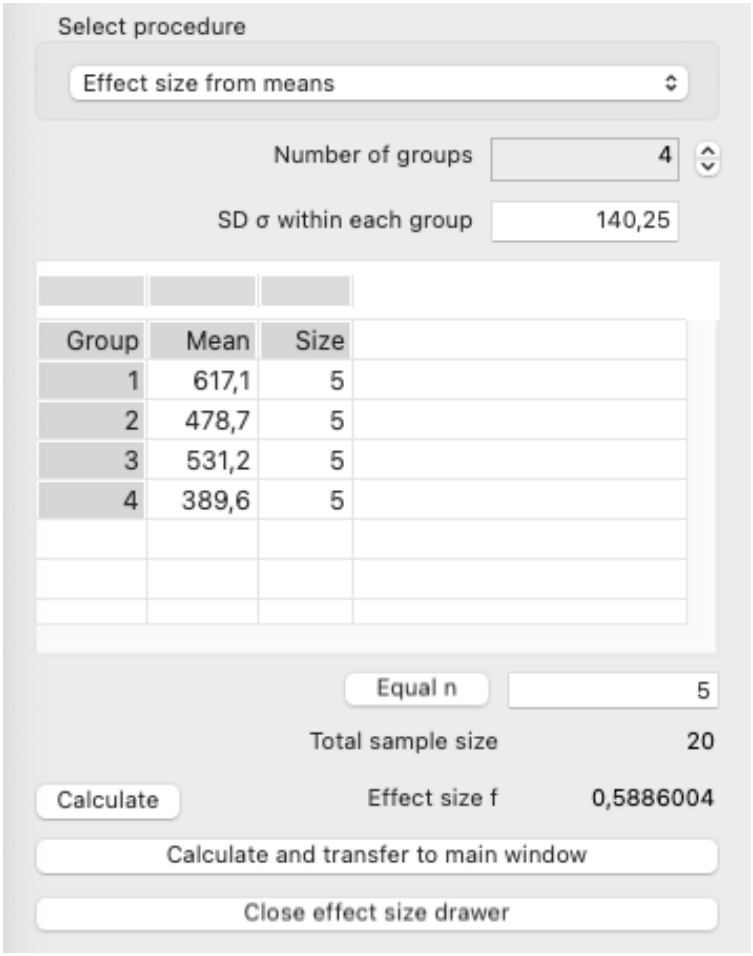
(Cortisol preliminary data: (mean ± SD) ikigun: 617.1 ± 151.2; ice slurry: 478.7 ± 95.1; prolonged anesthesia: 531.2 ± 121.1; electroanesthesia: 389.6 ± 81.4)

Standard deviation within group = 140,25 ng/ml group sizes of 5



Answer:

A total of 36 fish are needed (9 per group)



Exercise 3

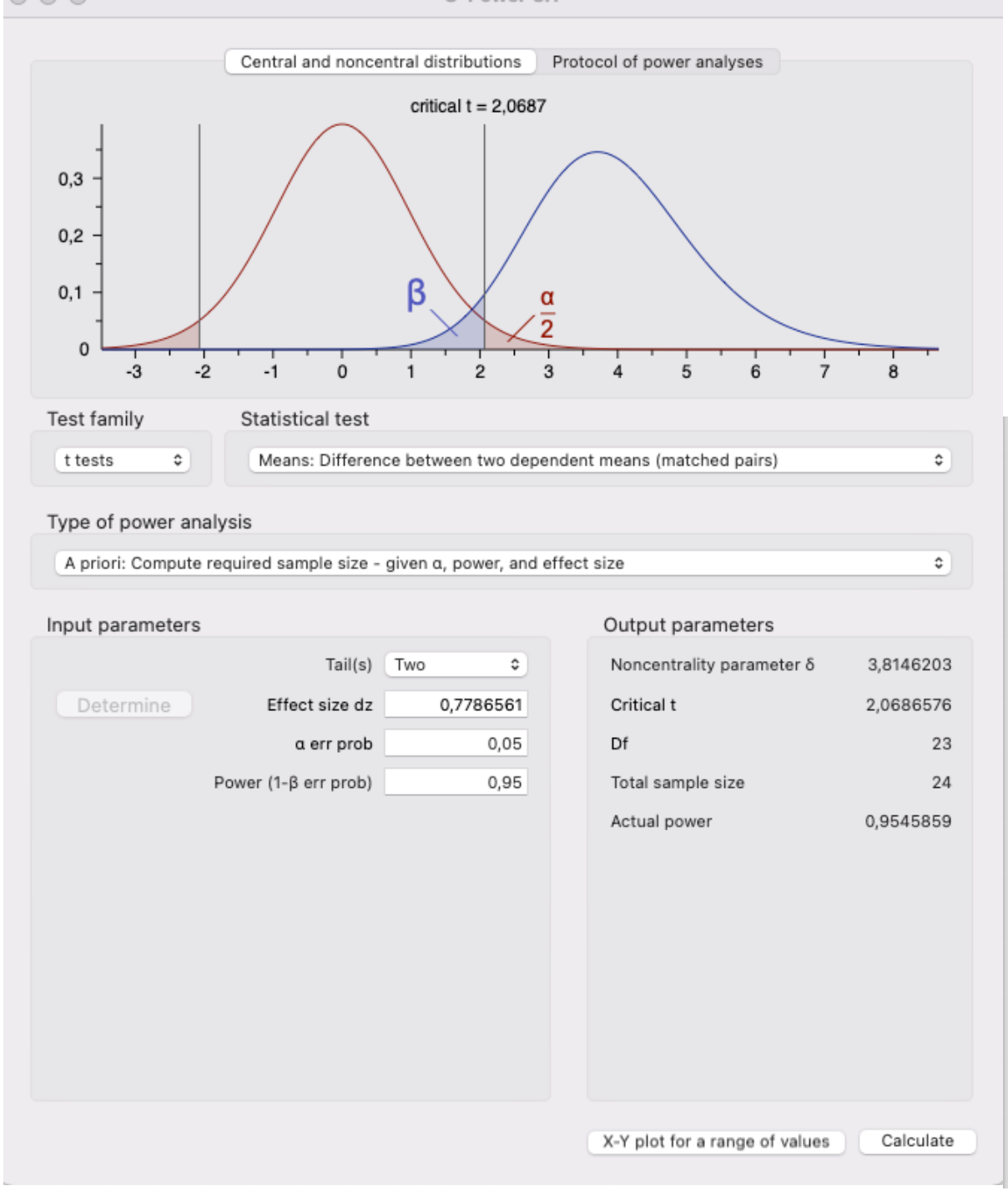
The light-dark test which controls anxiety. Increased latency (in sec) means more anxious.

You are interested in determining

if there is difference between treatments.

(Determine the sample size you will use)

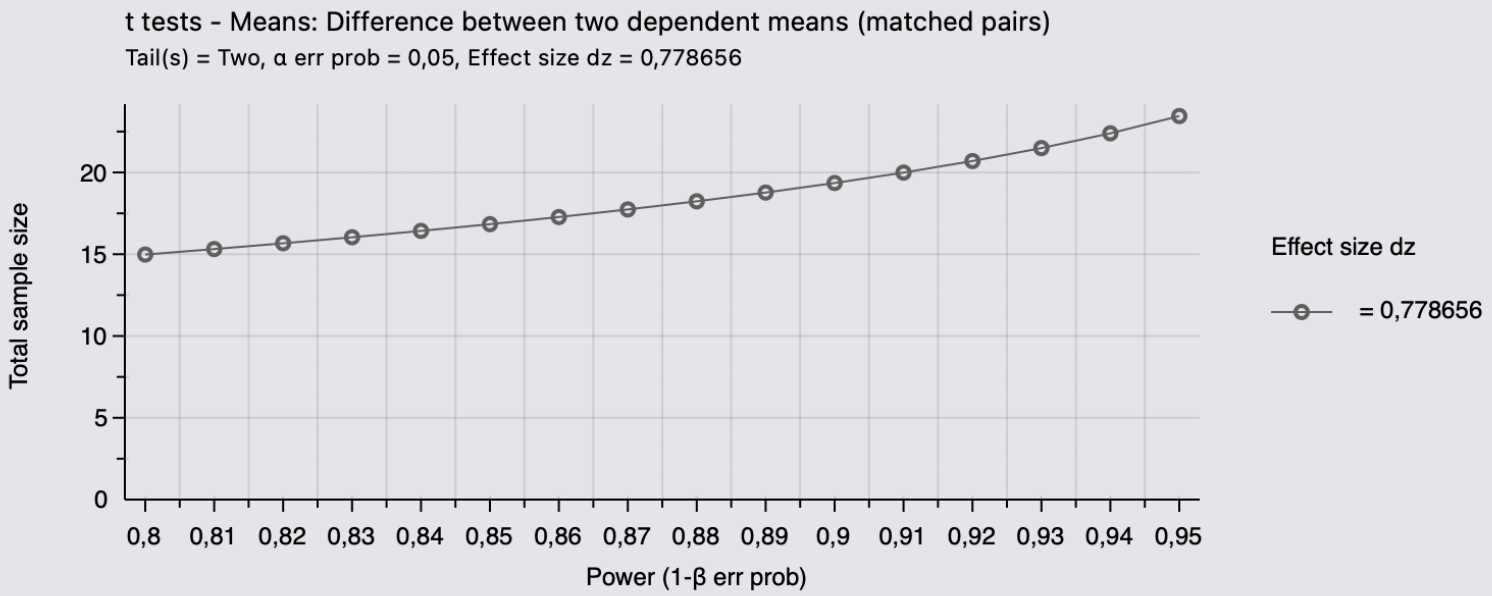
mice	no stress	stress	difference
1	28,00	10,20	17,80
2	64,00	28,00	36,00
3	31,60	23,00	8,60
4	46,00	30,00	16,00
5	19,20	26,51	-7,31
6	68,93	3,83	65,10
7	13,00	17,85	-4,85
8	24,70	8,34	16,36
9	28,81	16,22	12,59
10	29,47	17,52	11,95
11	13,98	12,90	1,08
mean	**33,43**	**17,67**	**15,76**
SD	**18,70**	**8,51**	**20,24**



Answer:
A sample size of 24 mice is needed to detect a difference with prob 0.95

Power Plot

Graph | Table



Parameters

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

as a function of Power (1-β err prob) from 0,8 in steps of 0,01 through to 0,95

Plot 1 graph(s) interpolating points

with Effect size dz at 0,7786561

and α err prob at 0,05

Draw plot

Assumptions

All statistical tests make assumptions about data

- parametric tests (like t-tests and ANOVA) make three main assumptions
  - probability distribution of errors (e.g. normal)
  - homogeneity of variance
  - independence of errors
- assumptions need to be checked before relying on the result of a test

Assumptions not met

- Robust if equal sample sizes
- Transformations of data may be useful
- Nonparametric tests
  - rank transform tests
    - Mann-Whitney test or Wilcoxon rank sum test for comparing two groups (nonparametric equivalent of the t-test)
    - Kruskal-Wallis for comparing several groups (nonparametric equivalent of the one-way ANOVA)
    - Friedman test (nonparametric equivalent of the randomized block ANOVA)

Thank you for your attention!

Dina Lika
lika@uoc.gr

References

- G.P. Quinn and M.J. Keough, Experimental Design and Data Analysis for Biologists. Cambridge (2002)

- J. H. Zar, Biostatistical Analysis, Prentice-Hall International, Inc. (1996)

- C. Dytham, Choosing and using statistics: A Biologist's guide. Blackwell science (1999)

- M.F.W. Festing and D.G. Altman. 2002. Guidelines for the design and statistical analysis of experiments using laboratory animals, *ILAR journal*, 43(4): 244-258