Test Design, Effect Size and Test Power

Matthias RAUTERBERG

Eindhoven University of Technology

2017

Causal Relationships

- ONLY if Independent Variables (IV) are specified and implemented UPFRONT.
- IV's are mainly Nominal Scaled.
- Any other measurements are either a Dependent Variable (DV) or a Co-Variate (CV).
- DV and CV can be of any scale type.

Test Design

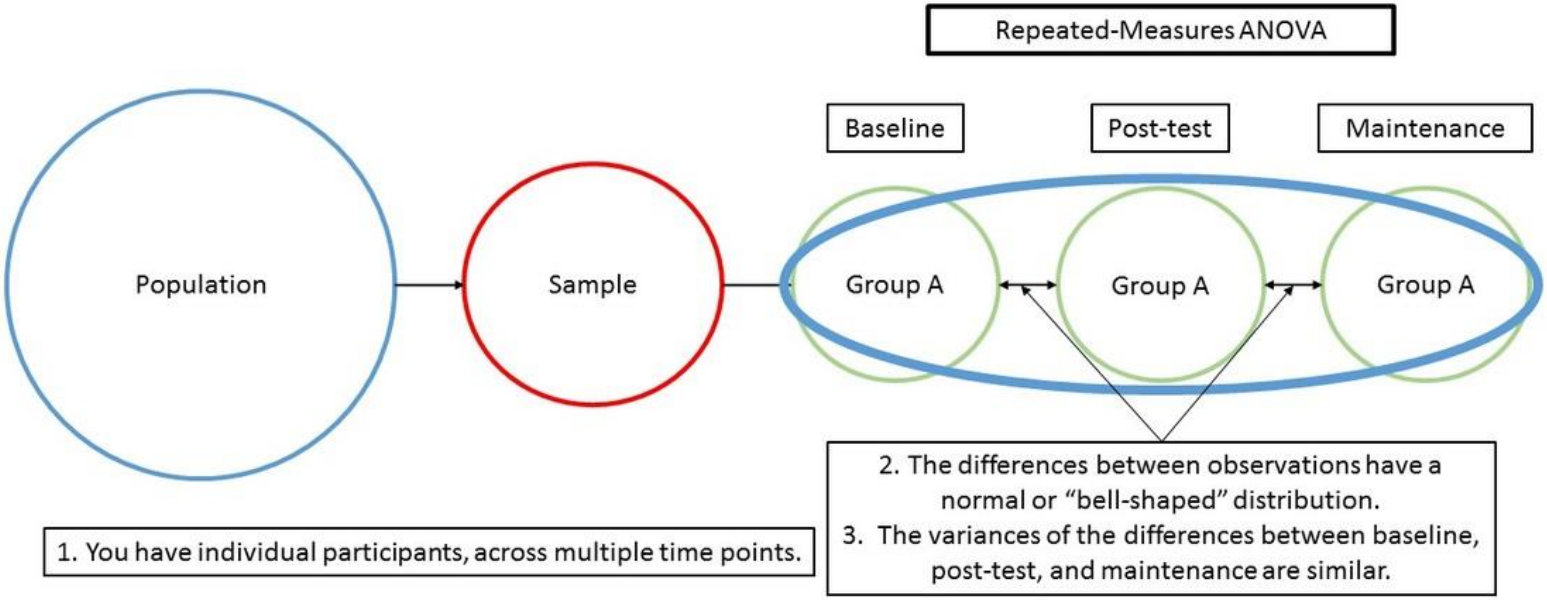
- **Between-subject design**: measurements are taken from different samples.
- **Within-subject design**: measurements are taken from same sample.

How to deal with Learning effects?

- If the main concept under investigation (e.g., a human being) is adaptive and can learn, any measurement might change it's internal state.

- This has to be addressed for within-subject test design with repeated measurements.

Repeated Measurements



Copyright © 2016 Scalë. All Rights Reserved.

Counter Balancing

Assuming we have 4 test conditions: A, B, C, D
Counterbalancing means to test ALL permutations:
[ABCD] [BACD] [BCAD] [BCDA] […]

Total number of possible sequences is 4! = 24

In case of a 2x2 factorial test design, it's called LATIN SQUARE.

We can easily calculate a factorial from the previous one:

n	n!		
1	**1**	1	1
2	2 × **1**	= 2 × **1!**	= 2
3	3 × **2 × 1**	= 3 × **2!**	= 6
4	4 × **3 × 2 × 1**	= 4 × **3!**	= 24
5	5 × **4 × 3 × 2 × 1**	= 5 × **4!**	= 120
6	etc	etc	

Treatment vs Control Group

- How to find an appropriate control test condition?
  - take best solution available (state of art)
  - avoid unfair comparison

Design of Experiments

- In the design of experiments, treatments are applied to experimental units in the **treatment group**(s).
  In *comparative* experiments, members of the complementary group, the **control group**, receive either *no* treatment or a *standard* treatment.

- For the conclusions drawn from the results of an experiment to have validity, it is essential that the items or test subjects are assigned to *treatment and control groups* be representative of the same population.
  In some experiments, such as many in agriculture or psychology, this can be achieved by randomly assigning items from a common population to one of the treatment and control groups. In studies of twins involving just one treatment group and a control group, it is statistically efficient to do this random assignment separately for each pair of twins, so that one is in the treatment group and one in the control group.

- In some medical studies, where it may be unethical not to treat patients who present with symptoms, controls may be given a standard treatment, rather than no treatment at all. Another alternative is to select controls from a wider population, provided that this population is well-defined and that those presenting with symptoms at the clinic are representative of those in the wider population.

Definition of Test Power

- Power (of a test) is the probability of detecting an effect, given that the effect is really there.

- Or likewise, the probability of rejecting the null hypothesis when it is in fact false.

- An example;
  - Power of 0.8 = if we performed a study 1000 times, we would see a statistically significant difference 80% of the time.

Outcomes of a statistical test

	H_0 True (no correlation)	H_1 True (correlation)
Do not reject H_0 (not stat. sig.)	**Correct decision**	**Type II (beta error)**
Reject H_0 (stat. sig.)	**Type I (alpha error)**	**Correct decision**

What you need to know

- Core elements
  - Effect size
  - Power
  - Sample size
  - Significance

- These elements are all inter-related such that;
  - If you know three you can estimate the fourth
  - Manipulating one influences the others

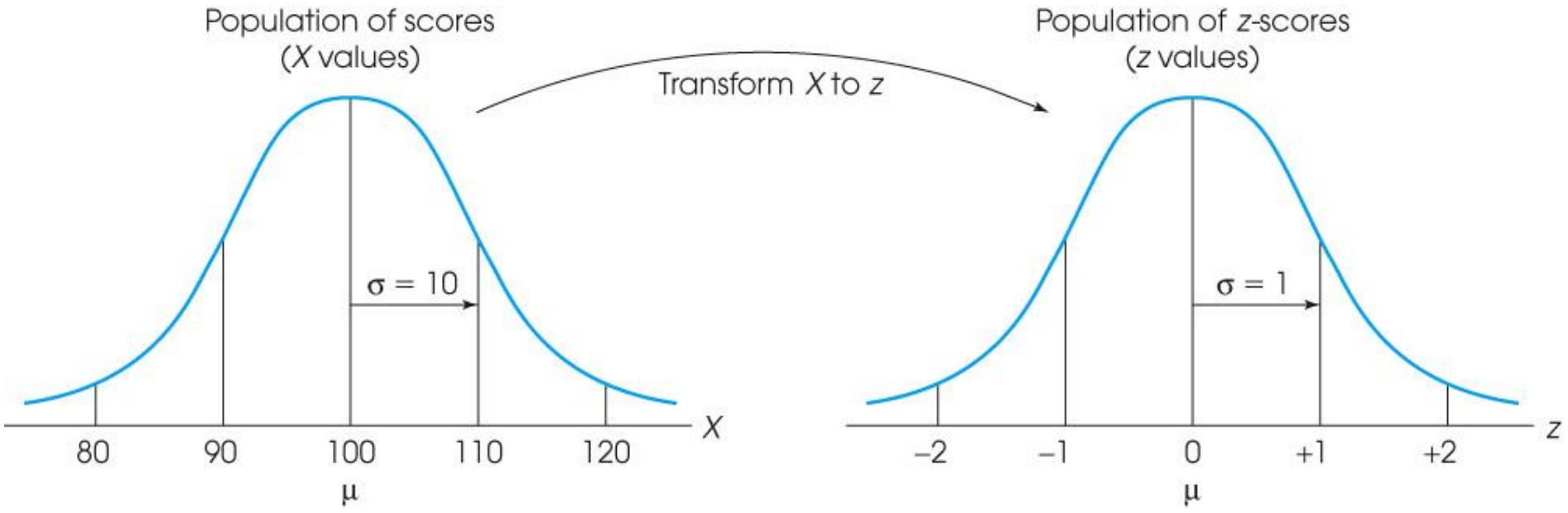
Introduction to sample size and power calculations

How much chance do we have to reject the null hypothesis when the alternative is in fact true?

(what's the probability of detecting a real effect?)

Can we quantify how much power we have for given sample and effect sizes?

The z-transformation of variables



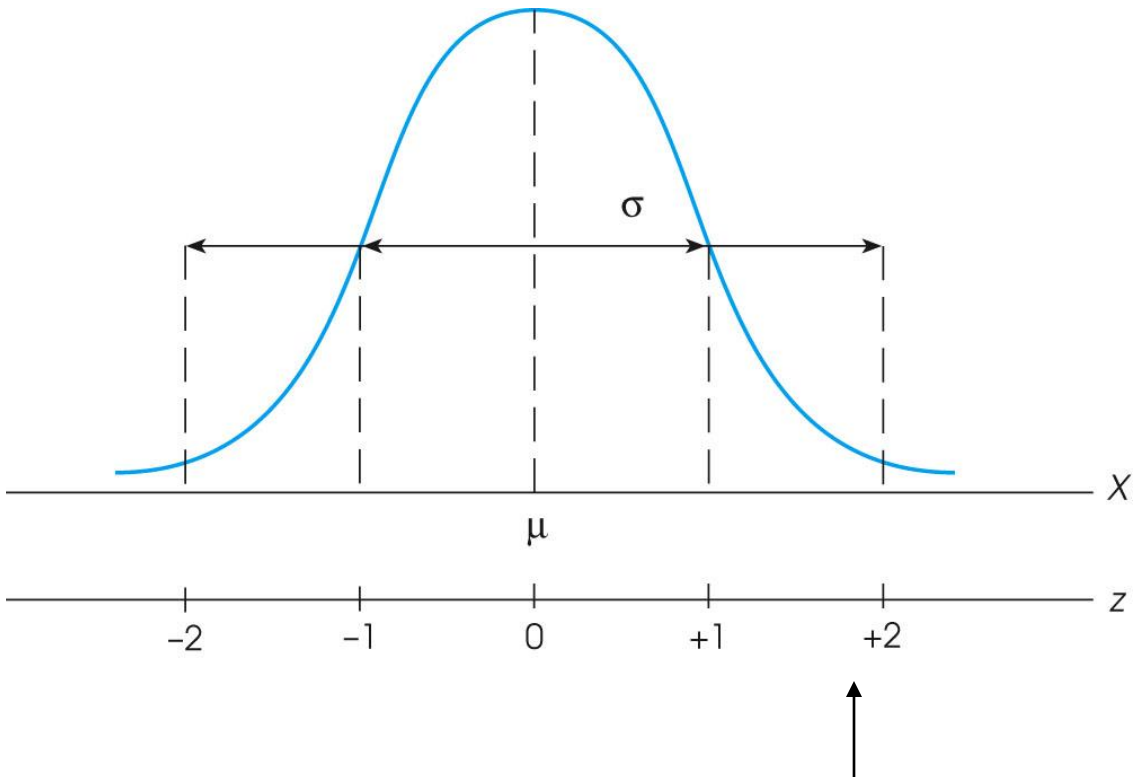
Transforming back and forth between X and z

- The basic z-score definition is usually sufficient to complete most z-score transformations. However, the definition can be written in mathematical notation to create a formula for computing the z-score for any value of X.
-

$$z = \frac{X - \mu}{\sigma}$$ for population

$$z = \frac{X - M}{s}$$ for sample

The z-distribution



Z=1.96 for p=0.05 (two sided)

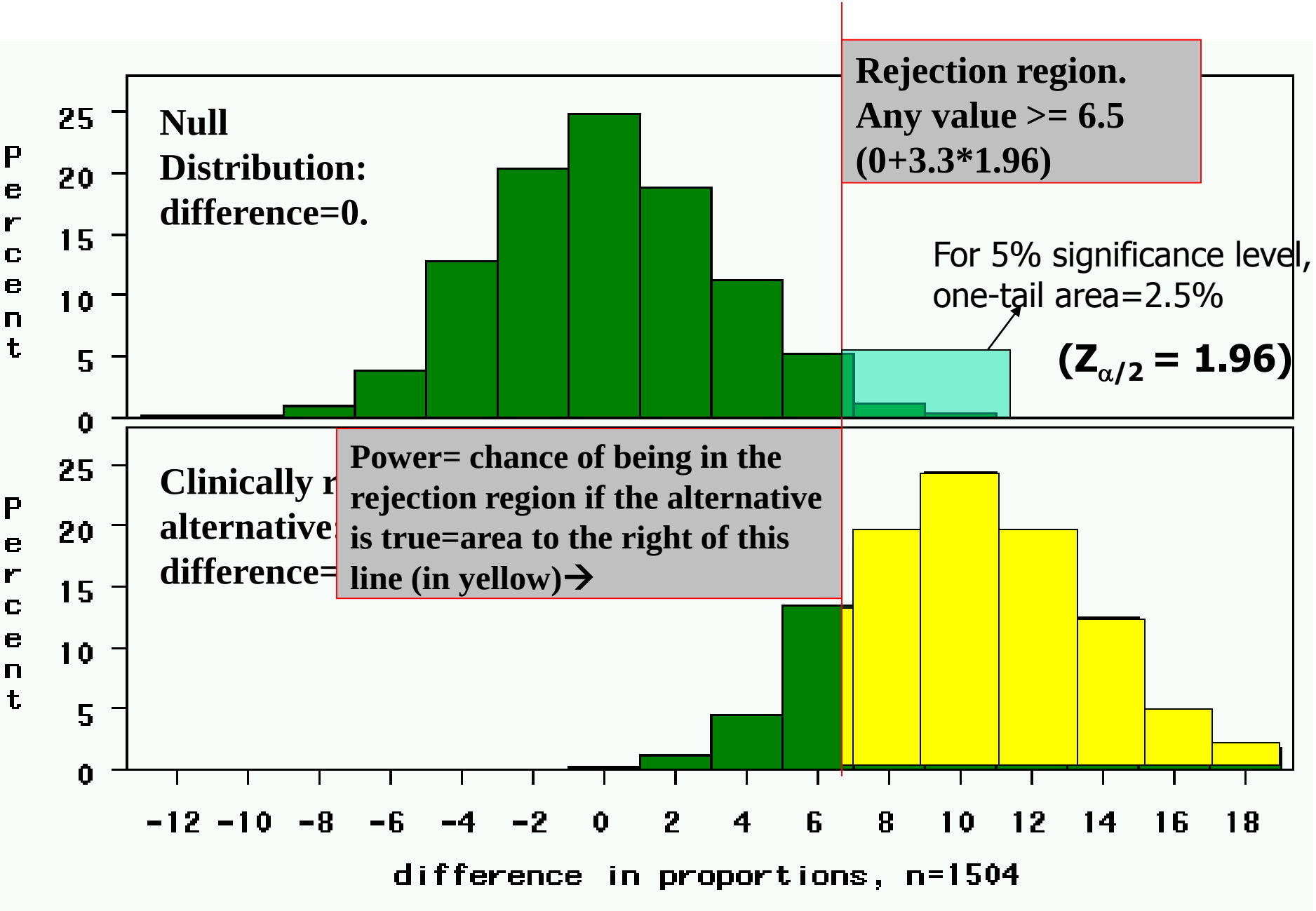
Deciding on levels of α and β

- **Power** (sensitivity) [1-β]
  - Probability of finding a true effect when one does exist
  - Type 2 error [β]: incorrectly accepting the null hypothesis (false negative)
  - Aim to minimise the risk of failing to detect a real effect
  - Typical values for power are 80%, 90% and 95%

- **Significance** (p-value) [α]
  - Probability that an effect occurred by chance alone
  - Type 1 error [α]: incorrectly rejecting the null hypothesis (false positive)
  - Aim to minimise the risk of detecting a non-real/spurious effect
  - Typical values are 0.05, 0.01

- **Reducing the risk of type 1 error ↔ increased risk of type 2 error (i.e. reduced power)**

study 1: 263 cases, 1241 controls



study 1: 263 cases, 1241 controls

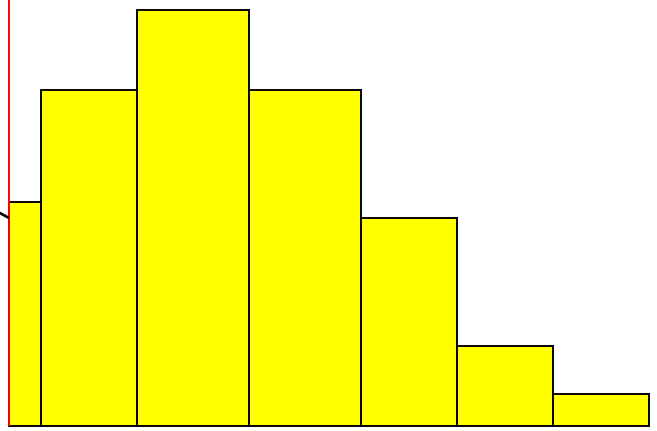
Rejection region. Any value >= 6.5 (0+3.3*1.96)

Power= chance of being in the rejection region if the alternative is true=area to the right of this line (in yellow)

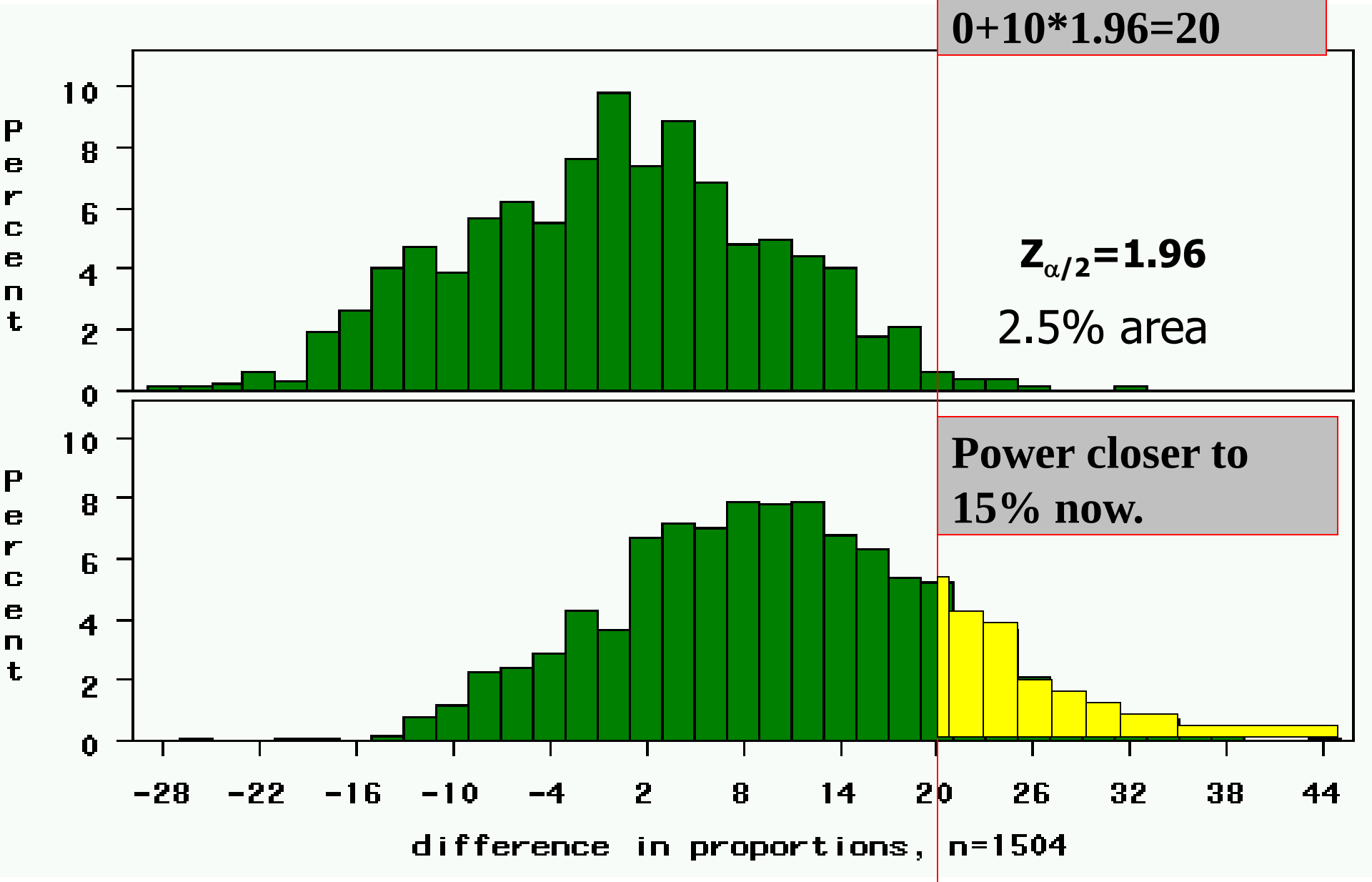
Power here:

$$\mathbf{P(Z > \frac{6.5 - 10}{3.3}) =}$$

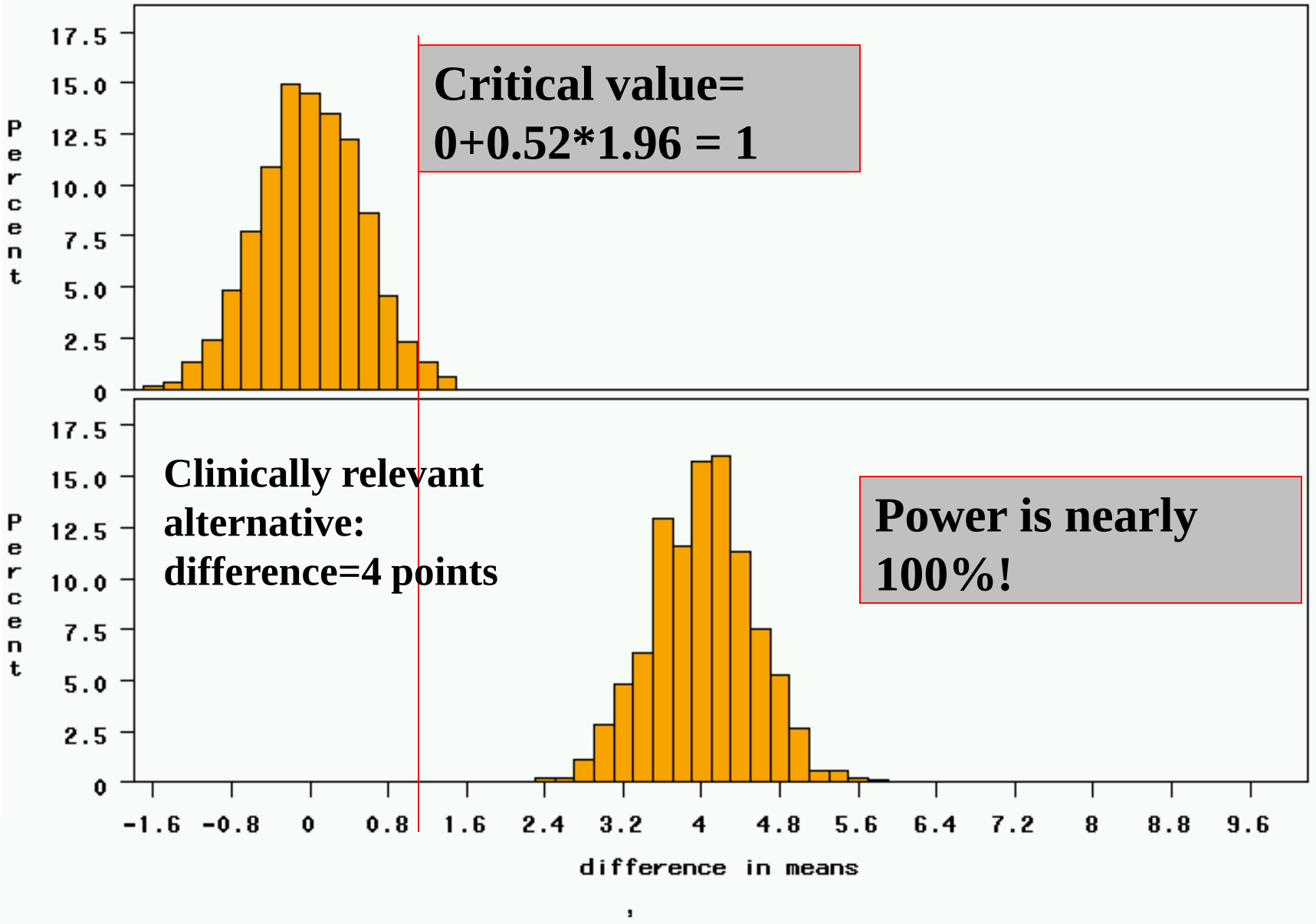
$$\mathbf{P(Z > -1.06) = 85\%}$$



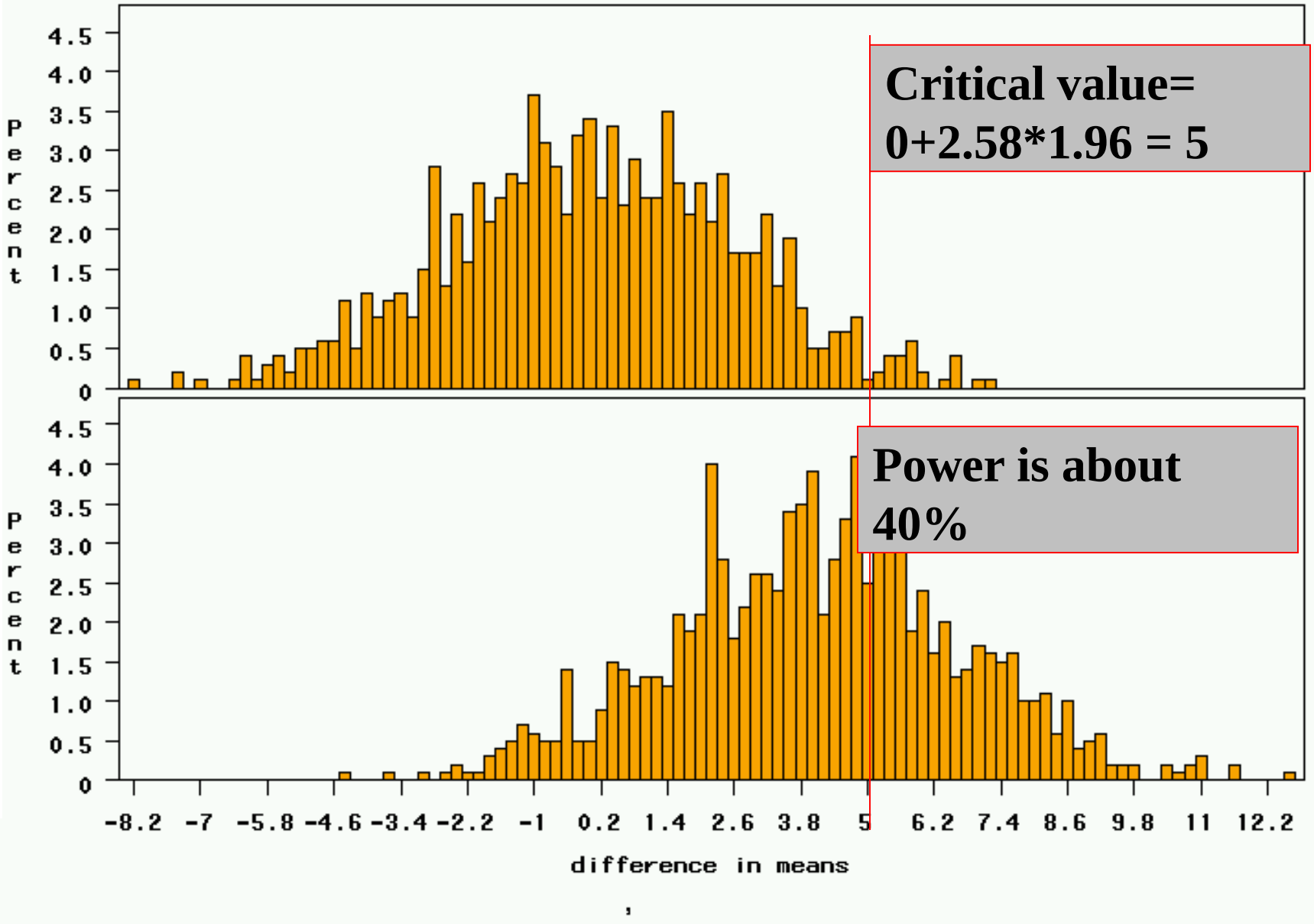
study 1: 50 cases, 50 controls



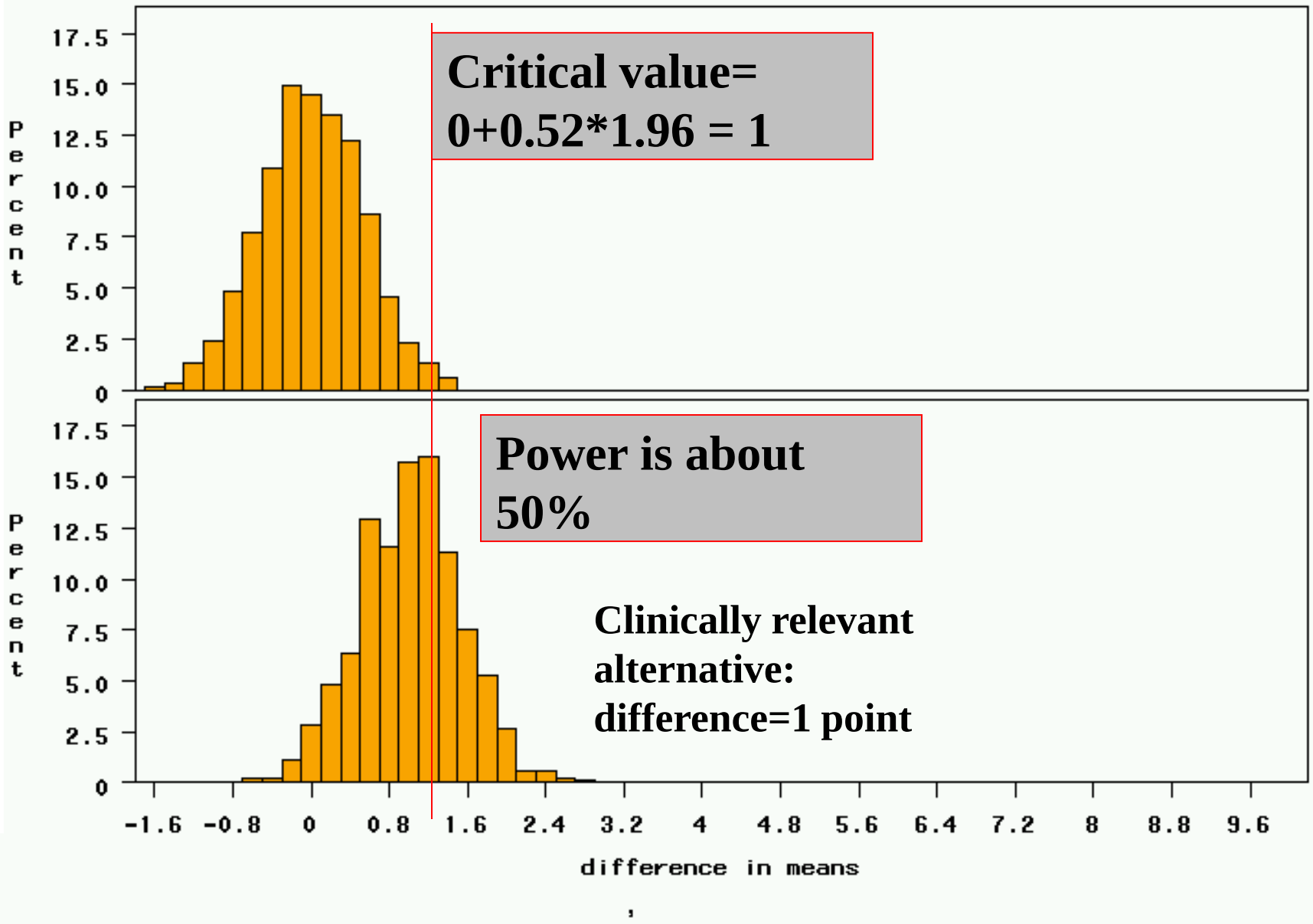
Study 2: 18 treated, 72 controls, STD DEV = 2



Study 2: 18 treated, 72 controls, STD DEV=10



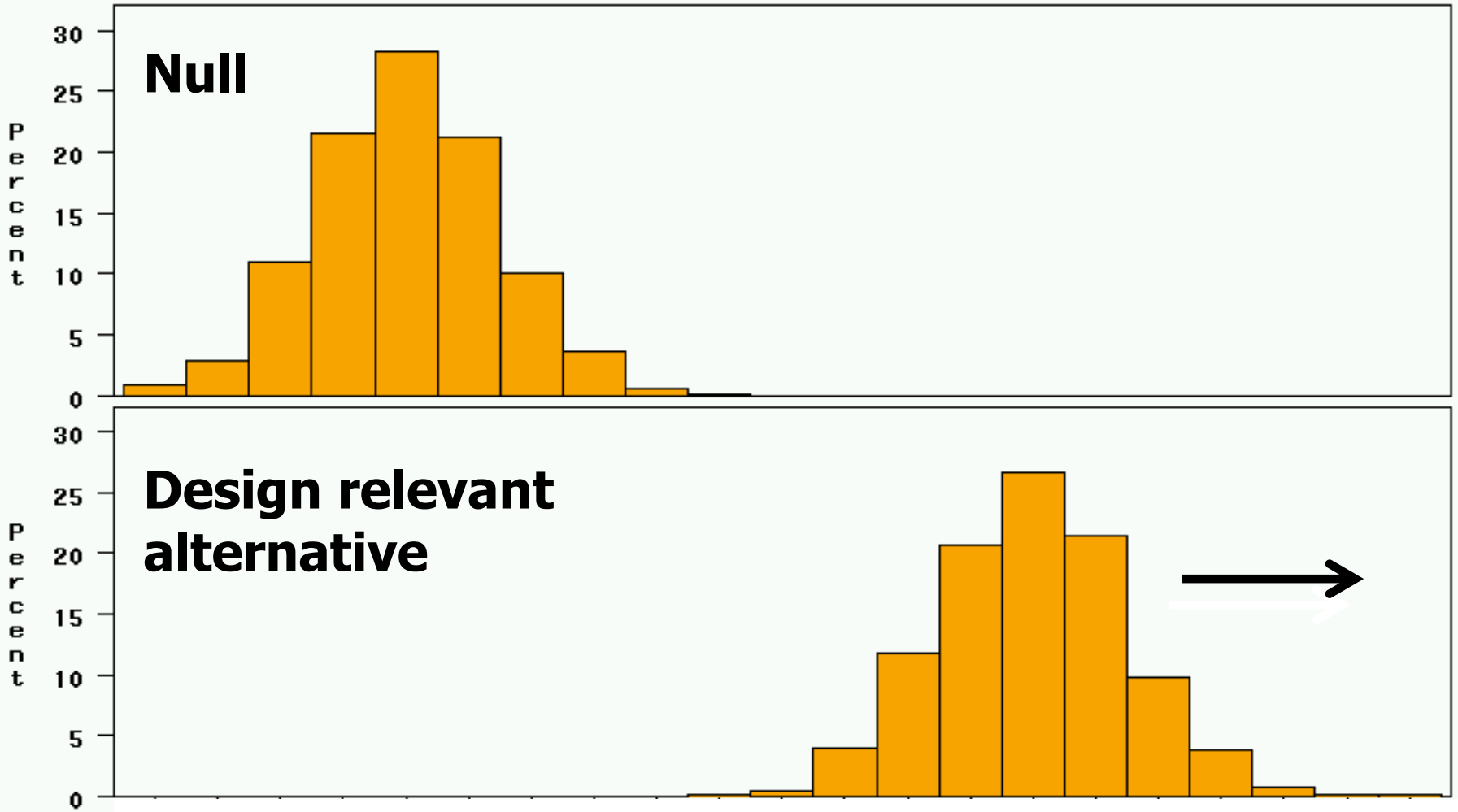
Study 2: 18 treated, 72 controls, **effect size=1.0**



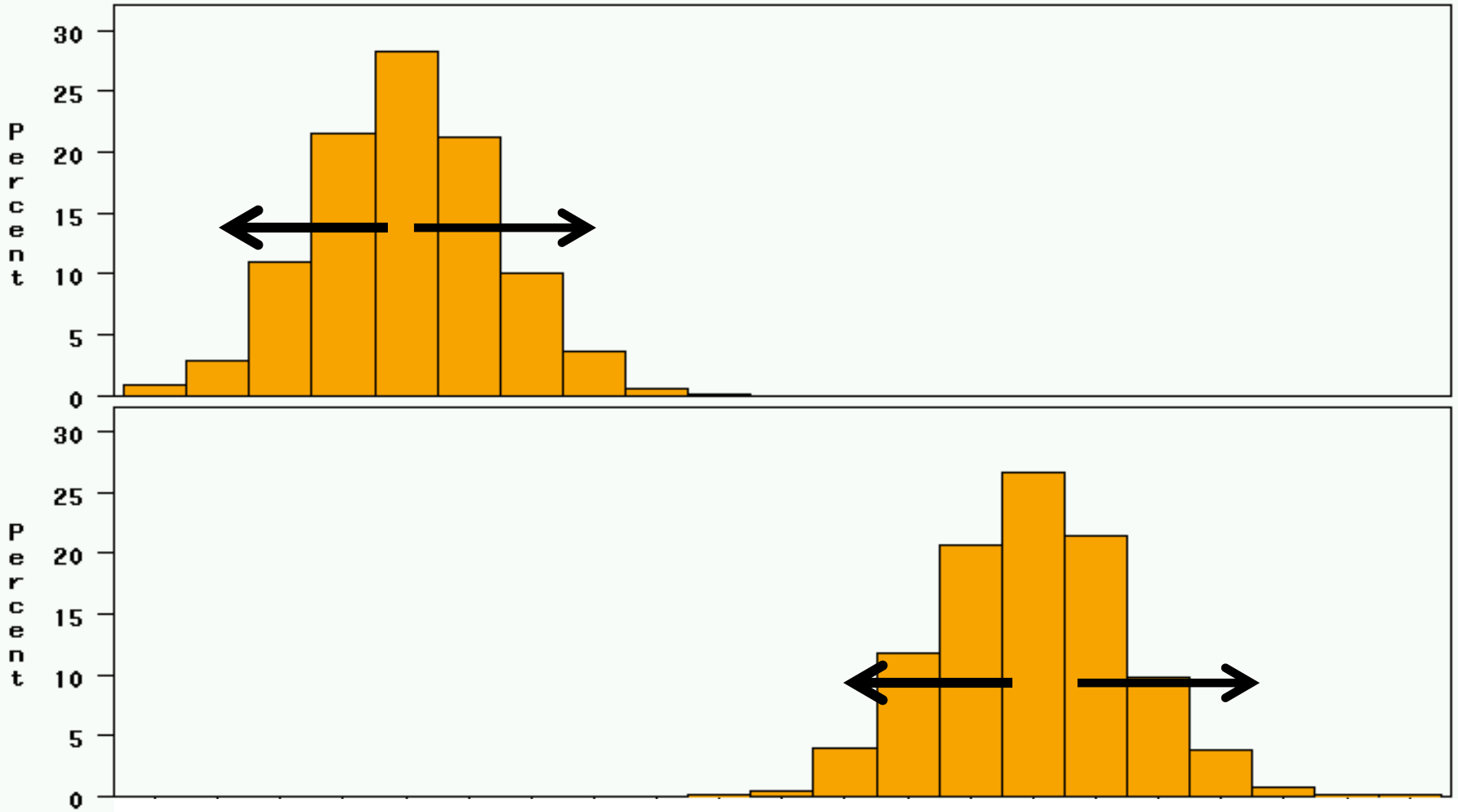
Factors Affecting Power

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

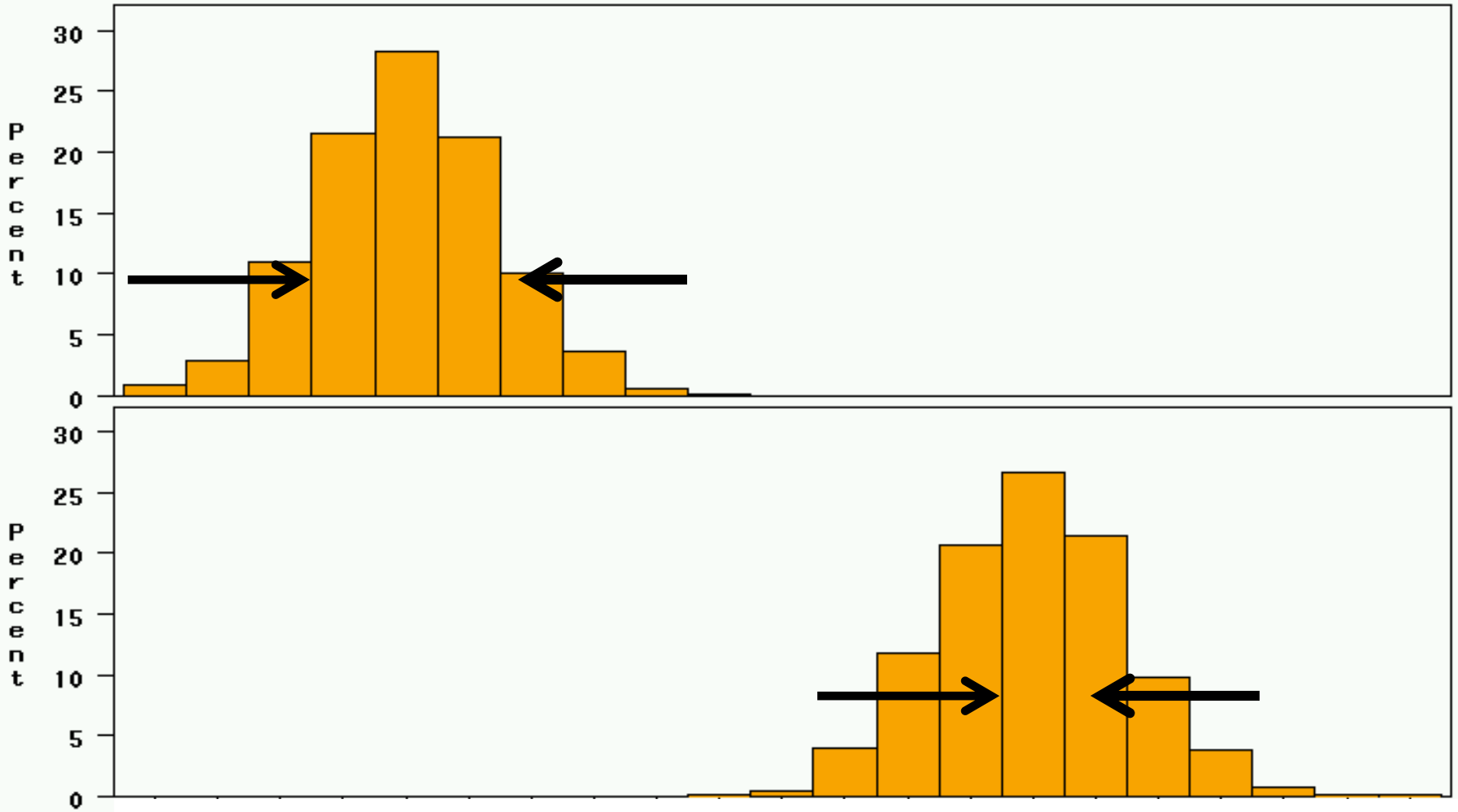
1. Bigger difference from the null mean



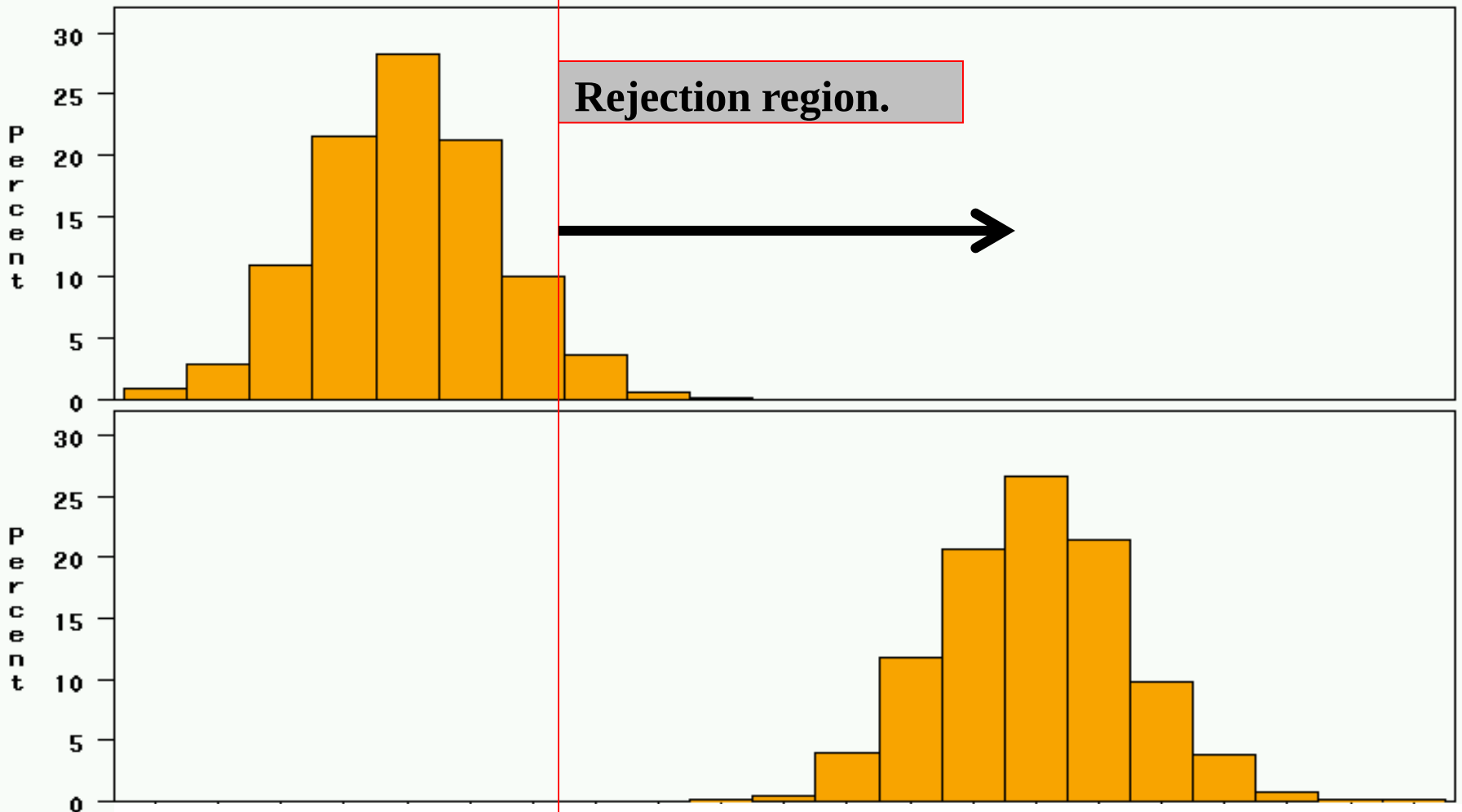
2. Bigger standard deviation



3. Bigger Sample Size



4. Higher significance level



Effect Size

- This is an indication of the size of the treatment effect, its meaningfulness.
- With a large effect size, it will be easy to detect differences and statistical power will be high.
- But, if the treatment effect is small, it will be difficult to detect differences and power will be low.

Effect Size

- Numerous authors have indicated the need to estimate the magnitude of differences between groups as well as to report the significance of the effects
- One way to describe the strength of a treatment effect, or meaningfulness of the findings, is the computation of "effect size" (ES)

$$ES = \frac{M_1 - M_2}{SD}$$

Note: SD represents the standard deviation of the control group or the pooled standard deviation if there is no control group

A note on Effect Size

- There are many ways to define and calculate effect size
  - Difference in means
  - Variance explained
  - Odds ratio

- Standardised vs. unstandardised measures
  - If possible use unstandardized measures
    - Raw difference between group means
    - Raw regression coefficients
- Use standardised effect sizes as a last resort
  - Standardised difference (d): difference in means/SD
  - Pearson's correlation coefficent (r)

Cohen's recommendations

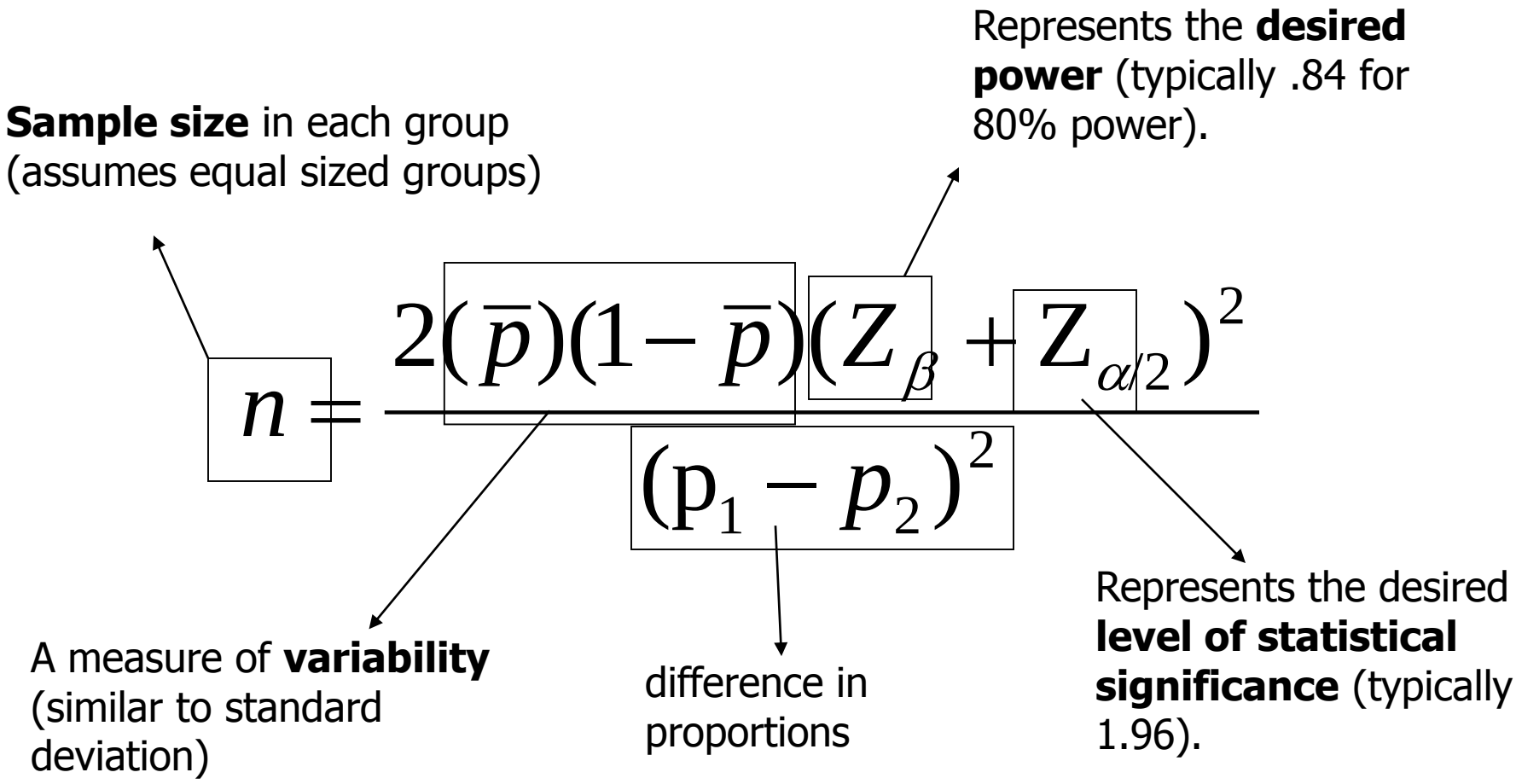
Effect	d	r
Small	≥0.2	≥0.1
Medium	≥0.5	≥0.3
Large	≥0.8	≥0.5

Sample size calculations

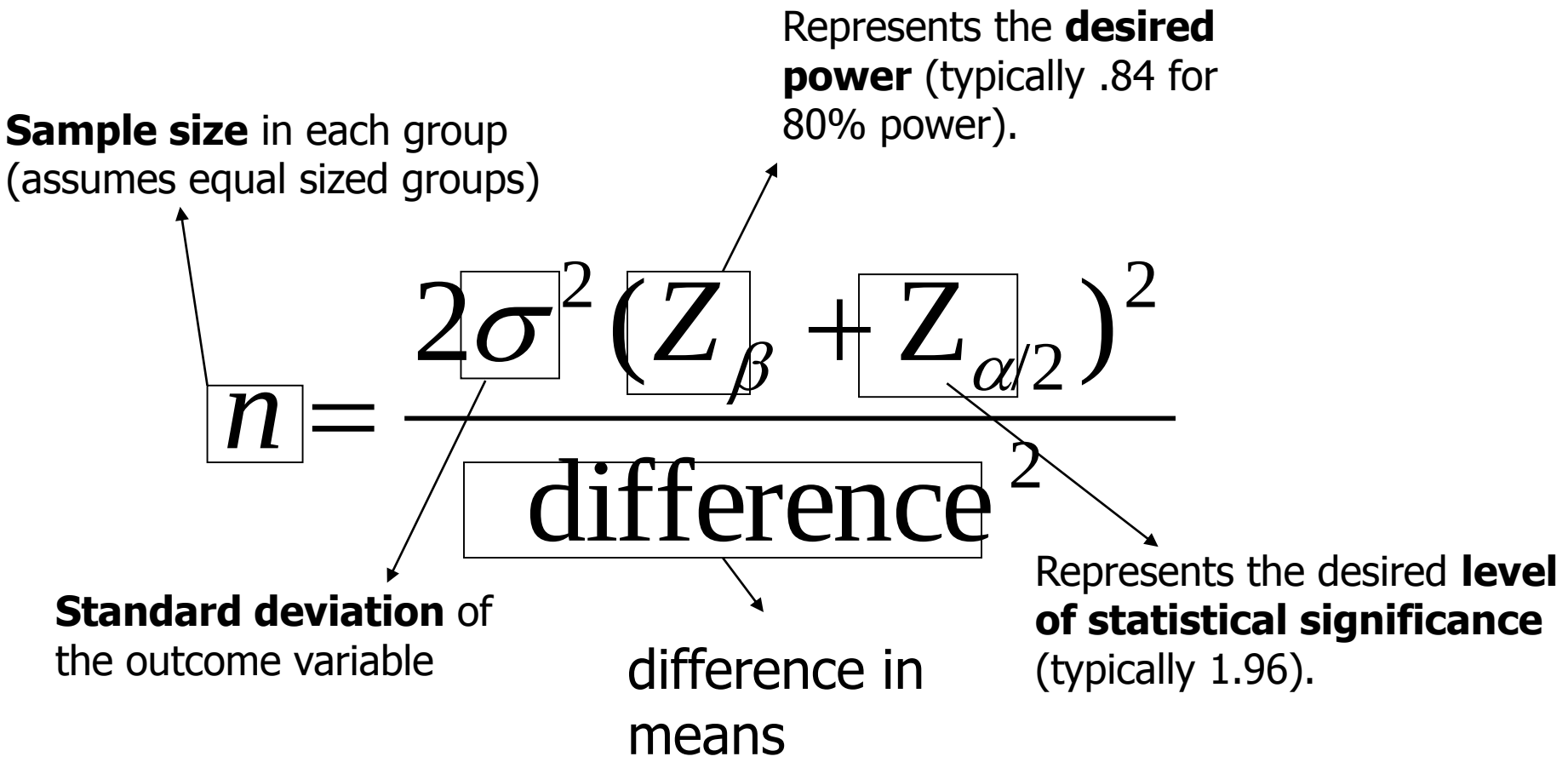
- Based on these elements, you can write a formal mathematical equation that relates power, sample size, effect size, standard deviation, and significance level…

The higher the amplifying factor (=sample size) the smaller the structures to be seen!

Formula for difference in *Proportions*



Formula for difference in *Means*

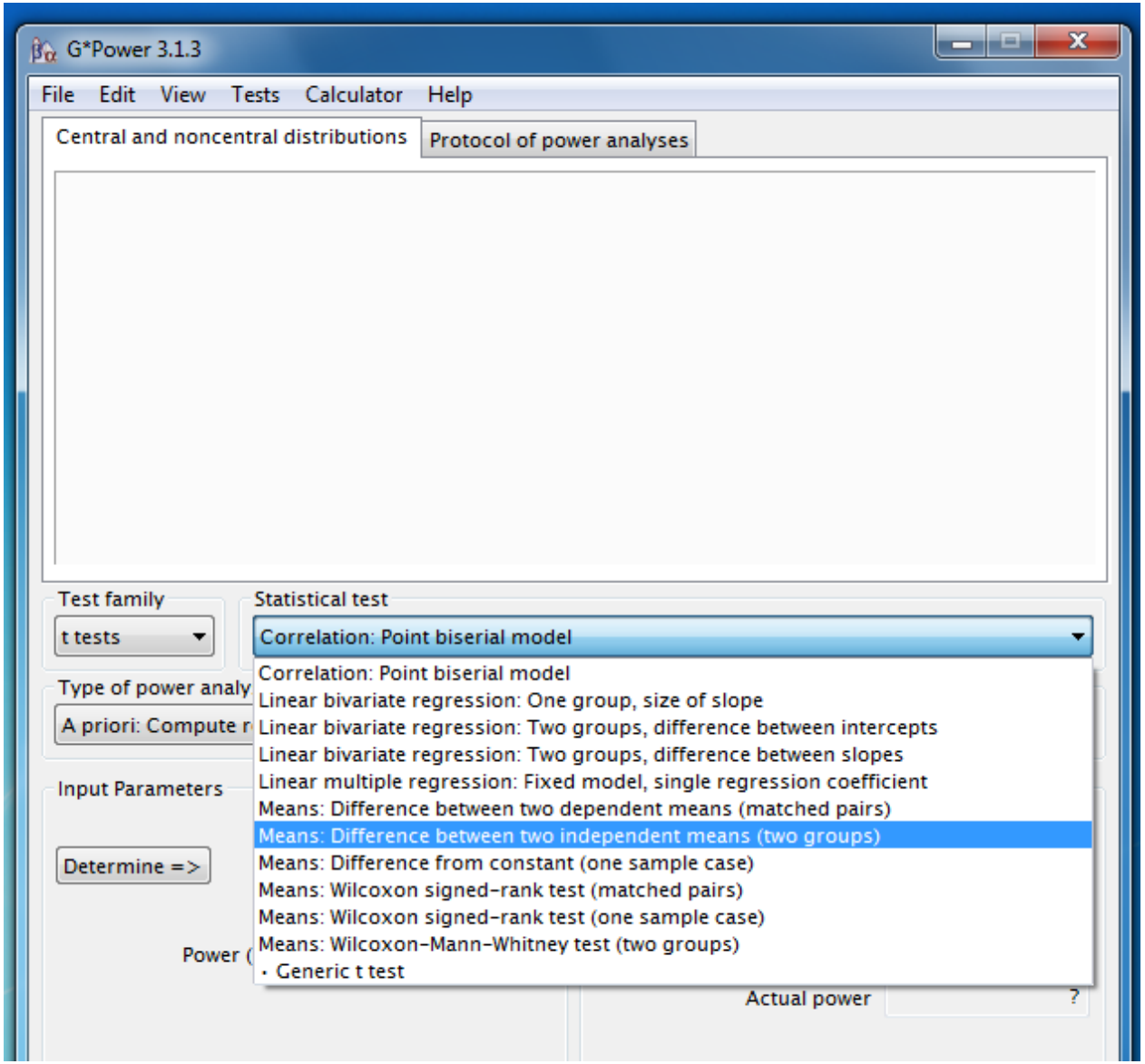


Question

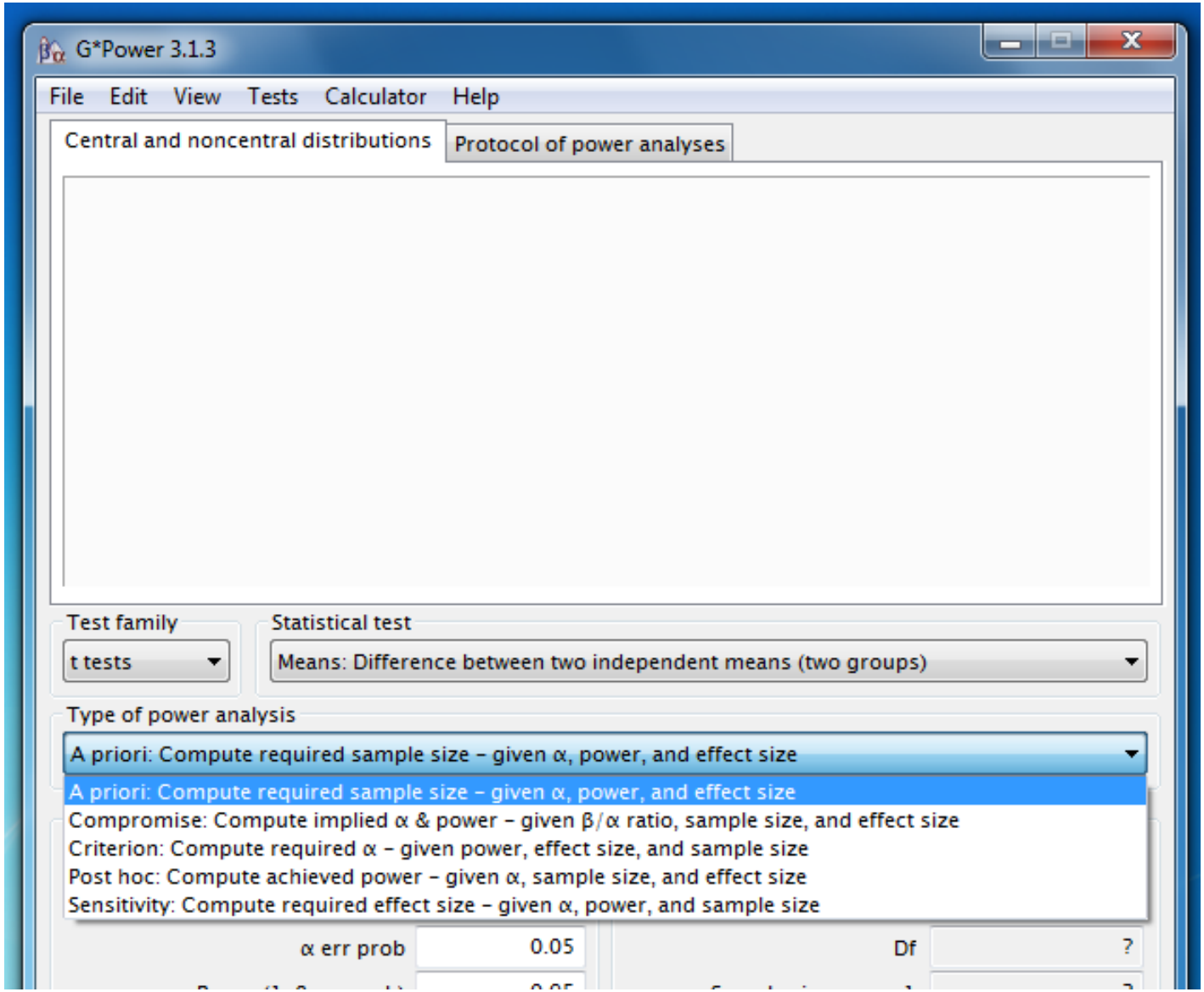
- How many test subjects would we need to sample if we want to have 80% power to detect an average increase in MCAT biology score of 1 point, if the average change without instruction (just due to chance) is plus or minus 3 points (=standard deviation of change)?

- Freely available software
  - G*Power www.psycho.uni-duesseldorf.de/abteilungen/aap/gpower3/
  - Quanto http://hydra.usc.edu/gxe/

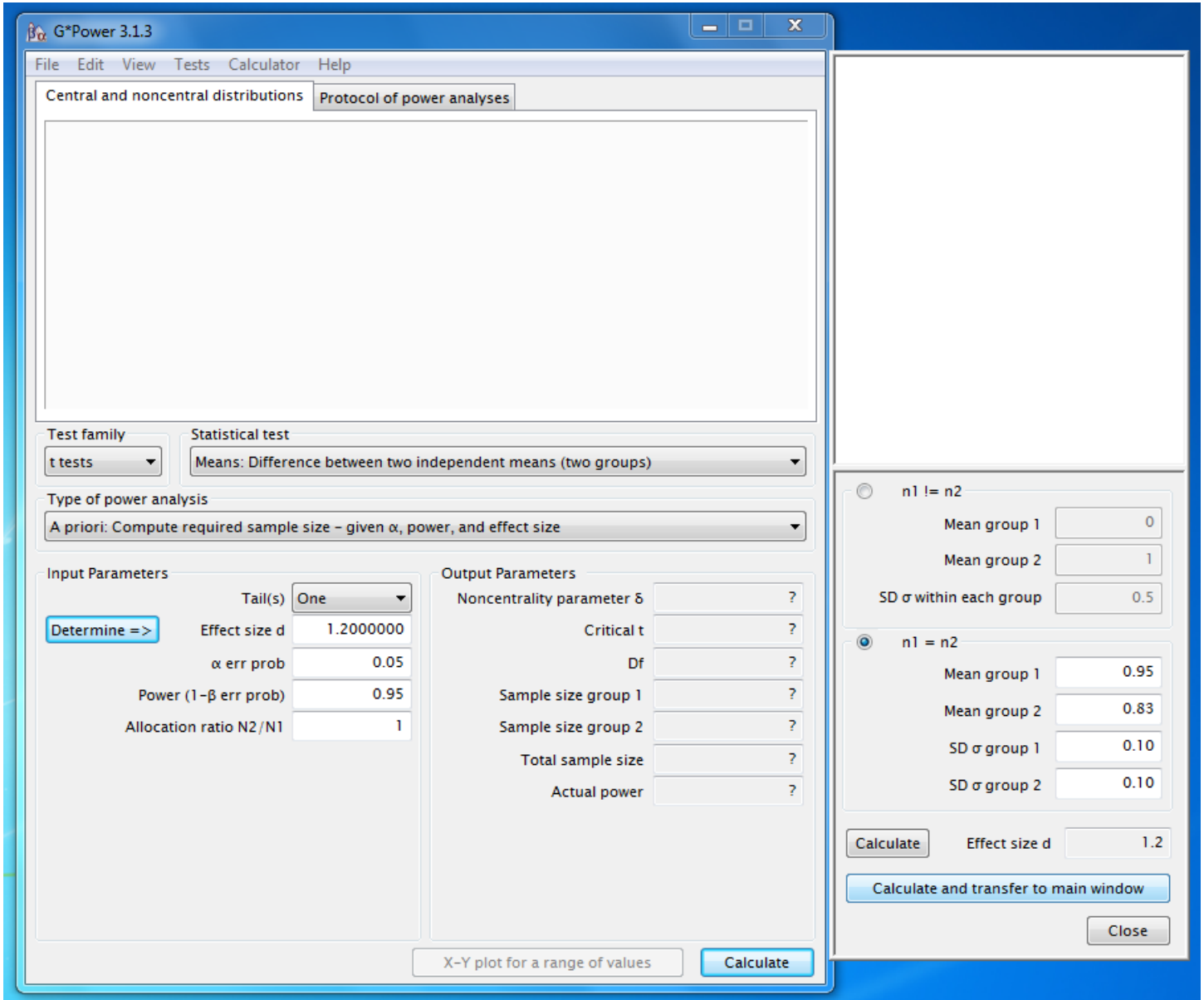
1. Select the statistical test



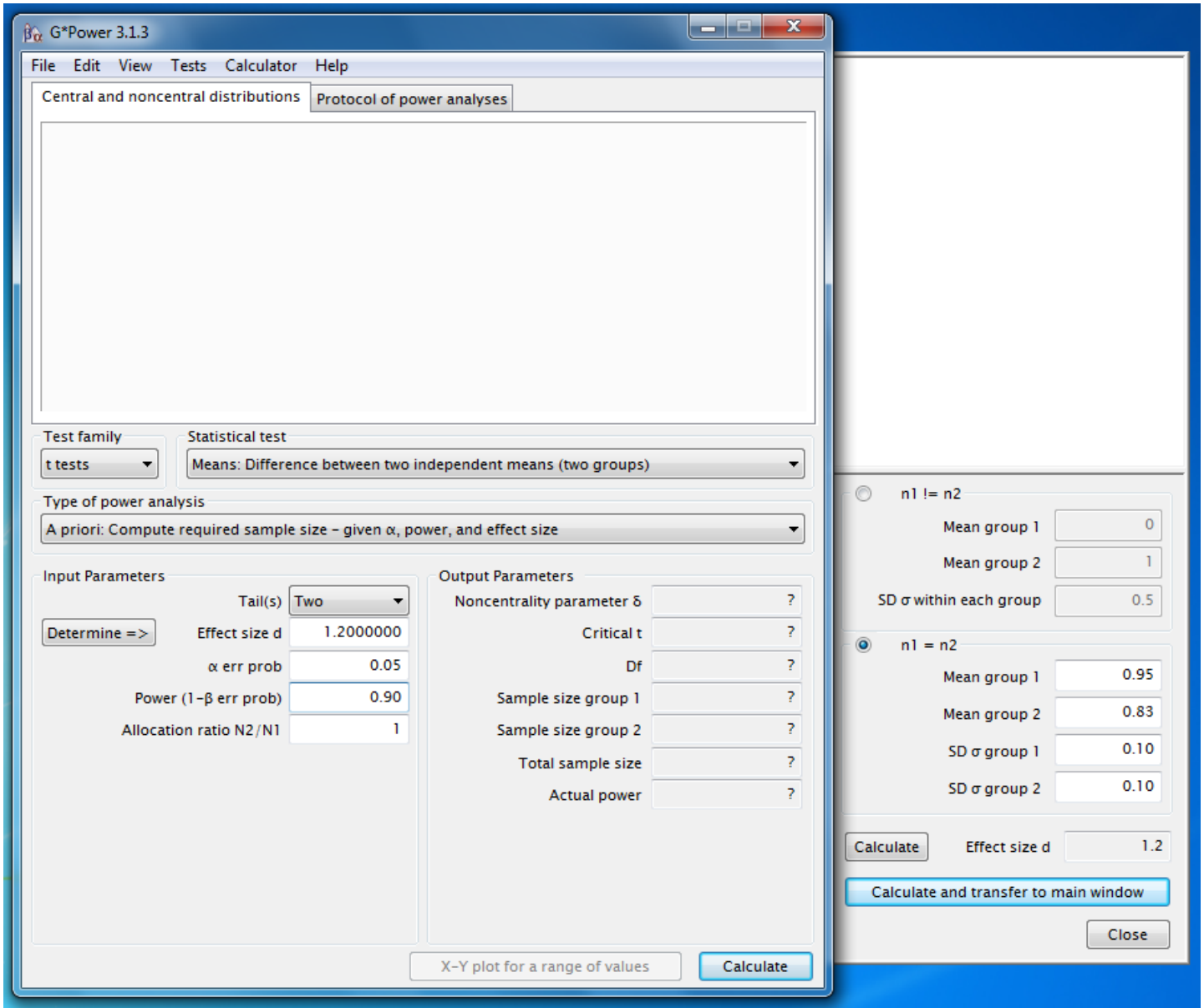
2. Select the type of power analysis



3. Input the data characteristics to determine the effect size



4. Input power parameters



5. Draw plot for a range of values