MSc Applied Information and Data Science

Lucerne University of Applied Sciences and Arts

Data Analytics for Data Scientists

Design of Experiments (DoE)

Lecture 06: Effect size & Power analysis

2025

Prof. Dr. Jürg Schwarz

Program: 16:15 until 17:55

16:15	**Begin of the lesson**
	Lecture: Jürg Schwarz ◦ Introduction ◦ Effect size ◦ Power analysis ◦ Preview of Lecture 07
	Tutorial: Students / Jürg Schwarz / Assistants ◦ Working on the exercise ◦ Support by Jürg Schwarz / Assistants
17:55	**End of the lesson**

Detailed table of contents → Slide 24

Introduction

Research question: Body height of women

Definition of the population (N = 2,600,000)

Women between the ages of 20 and 65 in Switzerland in 2024 (last completed statistical year)

Body height is measured

Mean μ_0 = 166 cm / Standard deviation σ_0 = 12.3 cm

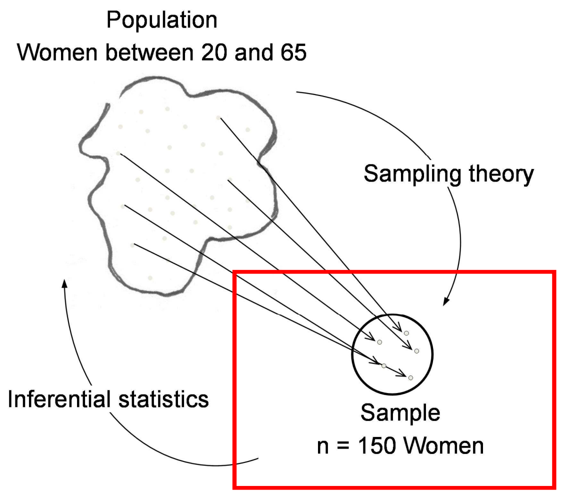
Sample with n = 150 women → data set *height.csv*

	key	height
1	1	177.5823
2	2	166.6227
3	3	186.8885
4	4	153.0198
5	5	173.1756
6	6	162.0006

Sample values

Mean	$\overline{x}$ = 165
Standard deviation	s = 12.7

Remember Lecture 05 "Sampling"



Note
For details on the t-test refer to the module "Classical and Bayesian Statistics"

t-test for a sample with data set *height* and *Collection of R code*

Sample of **n = 150** women with body height: $\bar{x} = 165$, $s = 12.7$

Hypotheses

H_0: Mean body height of the women in the sample is the same as in the population.

H_A: Mean body height of the women in the sample is not the same as in the population.

Significance level $\alpha = 5\%$

```
library(readr)
weight <- read_csv("height.csv")

t.test(height$height, alternative = "two.sided", mu = 166, conf.level = 0.95)
```

```
  One Sample t-test

data:  height$height
t = 1.4182, df = 149, p-value = 0.1582
alternative hypothesis: true mean is not equal to 161
…
```

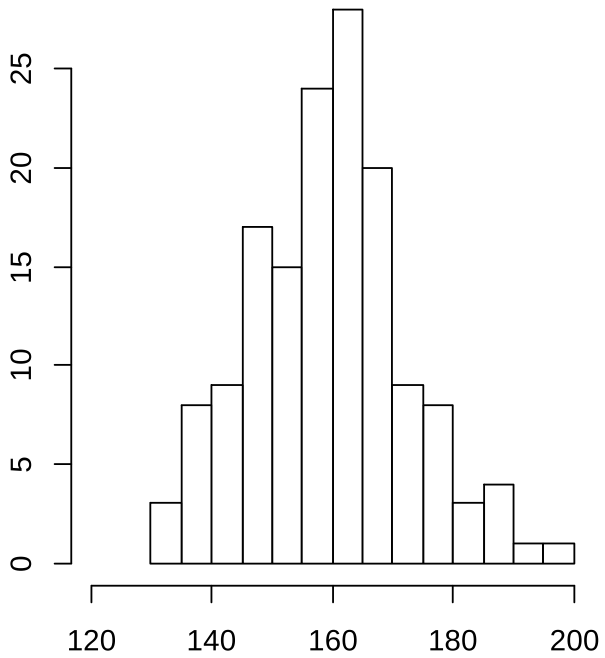
p-value $0.1582 > 0.05 \rightarrow H_0$ is accepted.

The mean of the body height of the women in the sample does **not differ significantly** from the mean in the population.

Experiment

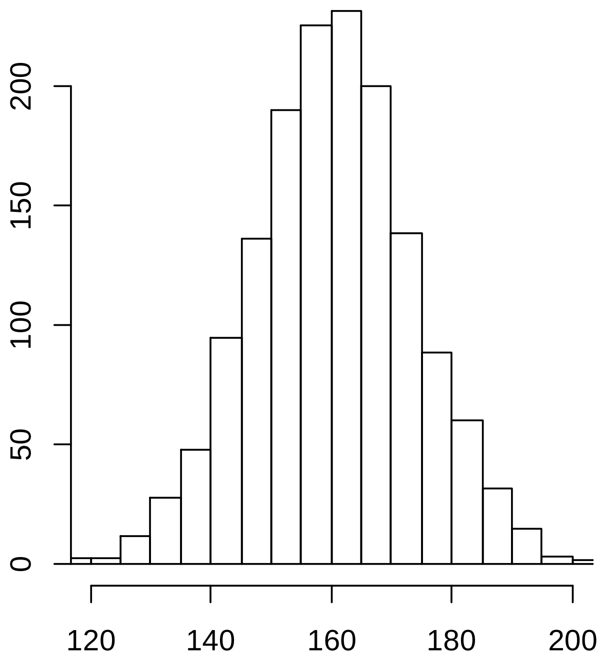
From the same population, two further samples are drawn

n = 150



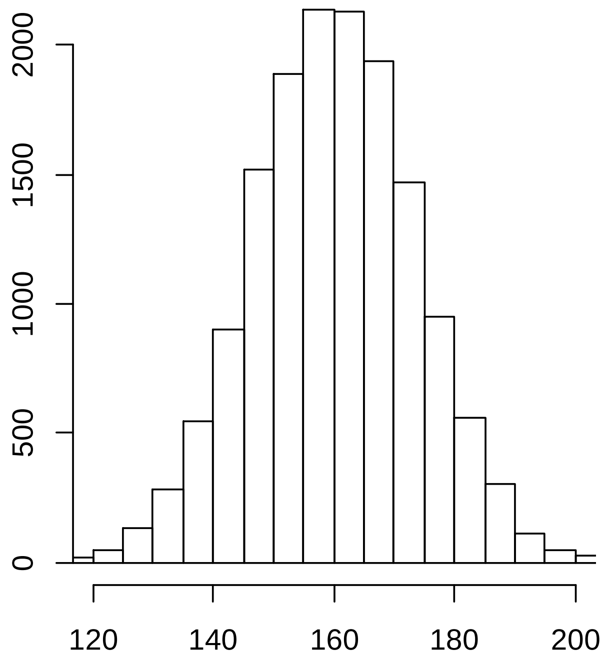
$\bar{x}$ = 165
s = 12.7

n = 1,500



$\bar{x}$ = 167
s = 13.1

n = 15,000



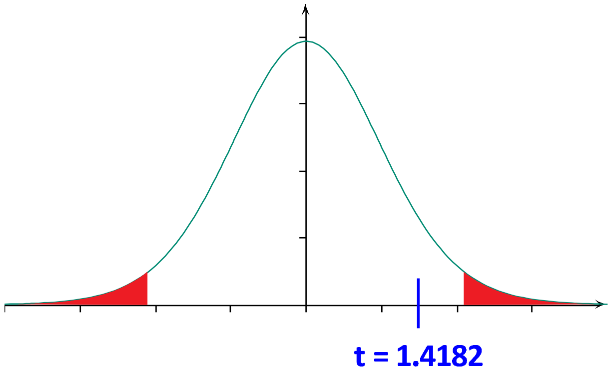
$\bar{x}$ = 164
s = 13.5

t-test

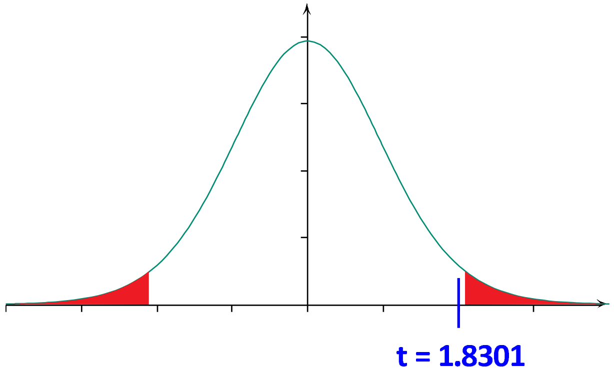
n = 150	n = 1,500	n = 15,000
`t-value = 1.4182`	`t-value = 1.8301`	`t-value = 8.2422`
`p-value = 0.1582`	`p-value = 0.06743`	`p-value = 0.0000`
H_0 is accepted.	H_0 is accepted.	H_0 is rejected.

Note
For details on the normal distribution refer to the module "Classical and Bayesian Statistics"

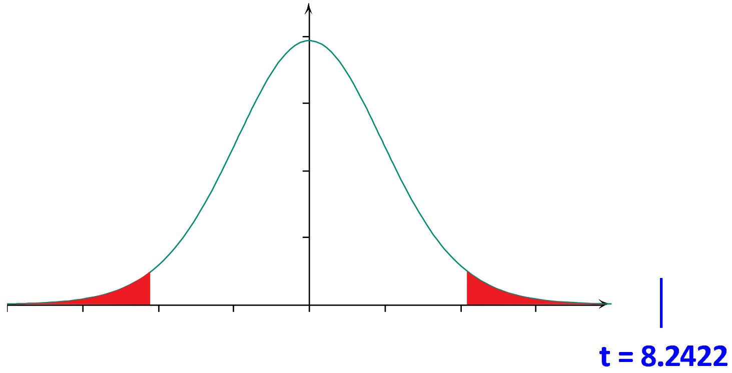
Carrying out the hypothesis test



H_0 is accepted



H_0 is accepted



H_0 is rejected

What happened?

The means and the standard deviations of the three samples are comparable.

Why does the t-test not result in acceptance of the null hypothesis for all samples?

The larger the sample size n, the larger the t-value becomes.

$$|t| = \frac{|\bar{x} - \mu_0|}{\frac{s}{\sqrt{n}}} = \frac{\sqrt{n}}{s} \cdot |\bar{x} - \mu_0|$$

The larger the t-value, the more likely it is that it lies outside the critical value.

Statistical significance and importance of an effect / strength of an effect

The chance of having a significant result in a hypothesis test is …

- larger if sample size **n** increases
- smaller if standard deviation **s** increases

$$|t| = \frac{\sqrt{n}}{s} \cdot |\bar{x} - \mu_0|$$

If sample n is large, …

- effects can become significant that in fact have no importance / strength

If sample size n is small, …

- effects may not become significant that in fact have importance / strength

What is an important / a strong effect in a difference or relationship?

This question can only be answered in the context of an actual study.

An effect in the population can be specified by an **effect measurement**.

This measurement results in what is referred to as effect size.

The name *effect size* (ES) comes from Cohen (1992).

This is the majority!
So the hypothesis must be true!

Huber (2017)

Effect size

Study on the effect of a new medicine that helps to lower blood pressure

Fact sheet

Research question

How does a newly developed medicine affect blood pressure?

Study design

Randomized control trial (RCT)

Dependent variable (**DV**) and independent variable (**IV**)

- Blood pressure [mmHg] (DV)
- Group [0/1] (IV): 0 = current medicine / 1 = new medicine

Population / Inclusion criteria

University Hospital of Zurich / Persons (age > 55) with high blood pressure (> 140 mmHg)

Net sample size

128 based on a power analysis with G*Power*, by assuming a medium effect.

Statistical analysis method

Independent samples t-test

*The G*Power tool will be introduced starting → Slide 19

Findings

Result [mmHg]	**Hypothesis test**	**Effect size**	**Impact**
Current 146.4	$p < .001$	d = 0.51	Target is ≤ 120 mmHg
New 137.9	Difference is significant	Effect size is medium	Is 137.9 sufficient?

Result

What knowledge can be extracted from a numerical result?

→ The difference is 8.5 mmHg, otherwise almost none.

Hypothesis test

What does a significant hypothesis test say (↔ null hypothesis is rejected)?

→ It is restricted to the context of the inferential statistics as a subset of research methods. That a hypothesis test becomes significant also **depends on the sample size**.

Effect size

What does the measure of effect size mean?

→ It indicates the **importance** of the study results.

Impact

How can the **clinical effect** be assessed?

→ It lies partly out of the context of empiricism.

Effect size according to Cohen (1992) – Example: Cohen's d

Examples of results from studies with new vs. current medicine / method / etc.

- New medicine reduces blood pressure by 8.5 units
  $|\bar{x}_{new} - \bar{x}_{current}| = 8.5$ mmHg
- New teaching method increases exam result by 5 points (scale 0 to 100 points)
  $|\bar{x}_{new} - \bar{x}_{current}| = 5$ points
- New production method reduces scrap by 20%
  $|\bar{x}_{new} - \bar{x}_{current}| = 0.2$

The differences can only be meaningfully assessed when standardized with a reference value.

For differences in mean values, this is the pooled standard deviation.

$$\hat{d} = \frac{\left|\bar{x}_{new} - \bar{x}_{current}\right|}{s}$$

$\hat{d} =$ Estimation for Cohen's d (Effect size after Cohen)

$\bar{x}_{new}, \bar{x}_{current} =$ Means of independent samples

$$s = \text{Pooled standard deviation} = \sqrt{\frac{(n_{new} - 1)s^2_{new} + (n_{current} - 1)s^2_{current}}{n_{new} + n_{current} - 2}}$$

Table from the article *A Power Primer* by Cohen (1992)

Test	ES index	Effect size Small	Effect size Medium	Effect size Large
1. m_A vs. m_B for independent means	$d = \frac{m_A - m_B}{\sigma}$	.20	.50	.80
2. Significance of product–moment r	r			
3. r_A vs. r_B for independent rs	$q = z_A - z_B$ where z = Fisher's z			
4. $P = .5$ and the sign test	$g = P - .50$			
5. P_A vs. P_B for independent proportions	$h = \phi_A - \phi_B$ where ϕ = arcsine transformation			
6. Chi-square for goodness of fit and contingency	$w = \sqrt{\sum_{i=1}^{k} \frac{(P_{1i} - P_{0i})^2}{P_{0i}}}$			
7. One-way analysis of variance	$f = \frac{\sigma_m}{\sigma}$			
8. Multiple and multiple partial correlation	$f^2 = \frac{R^2}{1 - R^2}$			

Cohen divides the effect size into three levels

ES index is valid for the population

$$d = \frac{|\mu_{new} - \mu_{current}|}{\sigma}$$

In empirical research, the value is estimated from the sample

$$\hat{d} = \frac{|\bar{x}_{new} - \bar{x}_{current}|}{s}$$

Also see note on the APA style on → Slide 14

Summary: Example of the new medicine for lowering blood pressure

Features of the study

- Control group (n_C = 64) / Treatment group (n_T = 64): Values of blood pressure

  Mean $\overline{x}_C$ = 146.4 mmHg, standard deviation s_C = 15.6 mmHg

  Mean $\overline{x}_T$ = 137.9 mmHg, standard deviation s_T = 17.7 mmHg

t-test with R (see details in the Appendix on Slide 22)

```
Two Sample t-test

data:  pressure by group
t = 2.9011, df = 126, p-value = 0.00439
```

→ t-test is significant at the 5% level (p-value < .050)

Estimation of effect size according to Cohen (1992)

Simplification: Since $n_C \approx n_T$ und $s_C \approx s_T$ → follows $s \approx (s_C + s_T) / 2$

$$\hat{d} = \frac{\left|\overline{x}_{new} - \overline{x}_{current}\right|}{s} = \frac{\left|146.4 - 137.9\right|}{16.7} = \frac{8.5}{16.7} = 0.51$$

Interpretation

The blood pressure values of the groups differ significantly, $t(126) = 2.901$, $p = .004$.
Effect size $d = 0.51$ (Cohen 1992).

*The three levels of d are found in "Table from the article A Power Primer by Cohen (1992)" → Slide 12

Note on the APA style

Numbers with values strictly between 0 and 1 are written
without leading zero according to APA (2020).

Example: p-values are strictly between 0 and 1, because it is a probability.
Therefore they are written without leading zero → p = .013

Numbers with values between 0 and ∞ are written
with leading zero according to APA (2020).

Example: In principle, standard deviations can lie between 0 and ∞ and are therefore written with leading zero → SD = 0.89

Effect sizes by Cohen (1992) can take values between 0 and ∞.

Nevertheless, Cohen himself described it without leading zero → d = .50

Power analysis / Determining the sample size

An example from Jones et al. (2003)

Among other things, a power analysis serves to determine the **sample size**.

Again: Study on a new medicine to lower blood pressure

Features of the research project

- Randomized control trial: New medicine compared to the current one
- Analysis with t-test
- Significance level $\alpha = 5\%$
- After Cohen level of $\beta = 20\% \rightarrow$ Power $= 1 - \beta = 80\%$*

These quantities are determined.

Clinically important difference δ – importance / strength

- Which blood pressure reduction is considered important?
  A reduction of 10 mmHg is considered clinically important

These quantities are determined by practical experience.

Other data

- It is known from validated studies that the standard deviation of blood pressure in the population is $\sigma_0 = 20$ mmHg.

*Thought as lower limit → Slide 17

Effect size

- The effect size in the population relevant for the study is

$$d = \frac{\text{Difference } \delta}{\sigma_0} = \frac{10 \text{ mmHg}}{20 \text{ mmHg}} = 0.5$$

Question concerning power analysis

- How large does the total sample have to be for

    significance level α = 5%

    power = 1 – β = 80%

    effect size d = 0.5 in the population

  to obtain a significant t-test for independent samples
  if the change in blood pressure is in fact at least 10 mmHg?

Answer

- Use …

    tables

    nomograms

    software – e.g. G*Power

Answer in Jones et al. (2003) – by using a table

Table 3 How power changes with standardised difference

	Power level ($p\beta$)			
Sdiff	0.99	0.95	0.90	0.80
0.10	3676	2600	2103	1571
0.20	920	651	527	394
0.30	410	290	235	176
0.40	231	164	133	100
0.50	148	105	86	64
0.60	104	74	60	45
0.70	76	54	44	33
0.80	59	42	34	26
0.90	47	34	27	21
1.00	38	27	22	17
1.10	32	23	19	14
1.20	27	20	16	12
1.30	23	17	14	11
1.40	20	15	12	9
1.50	18	13	11	8

Sdiff, standardised difference.

Power level corresponds to
Power = 1 - β
Sdiff corresponds to
effect size d = 0.5

The values in the table refer to one group.

Therefore, the sample has to be doubled: 2 x 64 = 128

With a sample size of n = 128, a blood pressure reduction of at least 10 mmHg can be tested significantly by using a t-test with α = 5% and a power of 80%.

Note: As shown in table 3, a value of 80% for the power is generally a lower limit (Cohen 1992).

Answer in Jones et al. (2003) – by using a nomogram

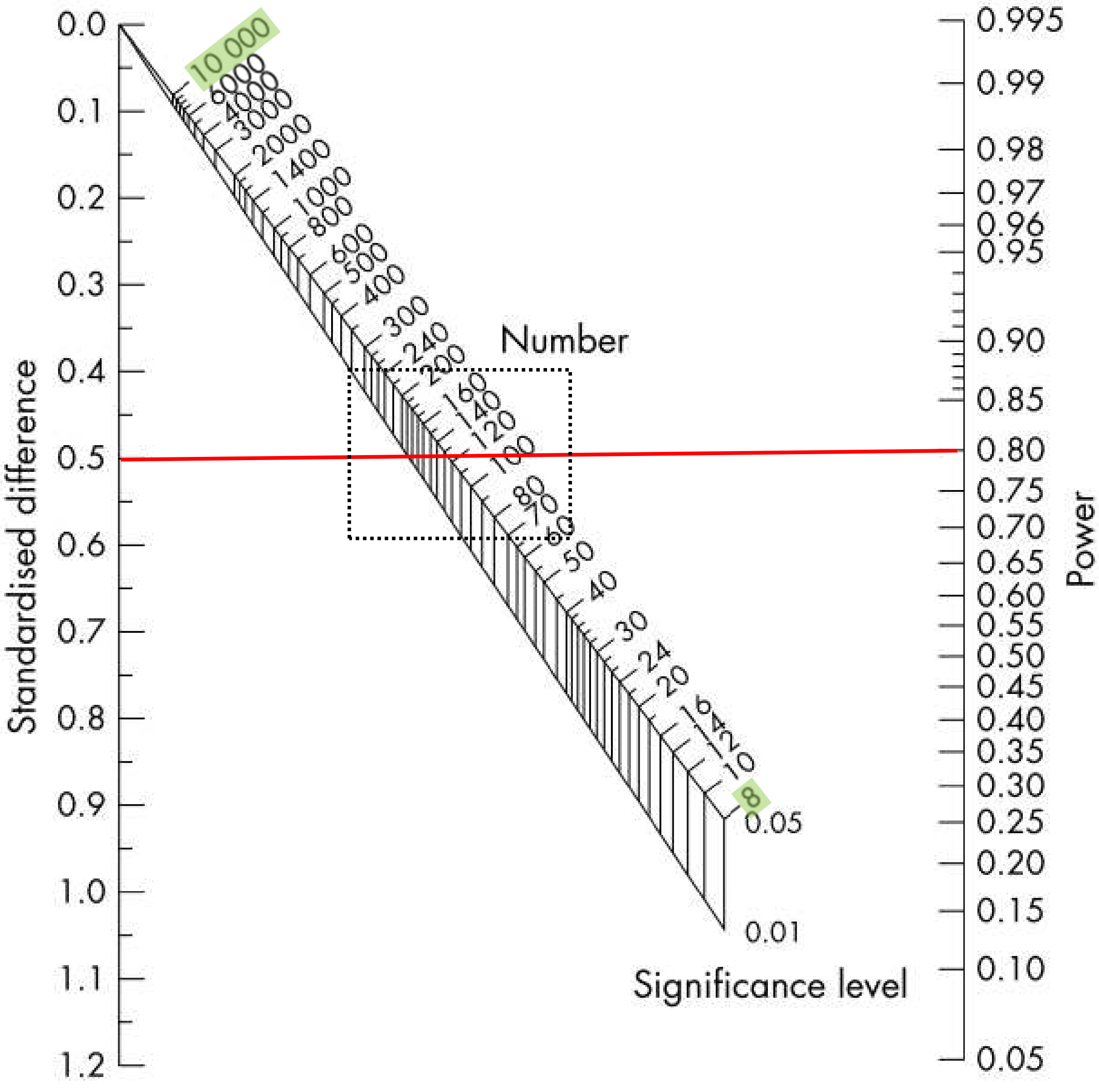
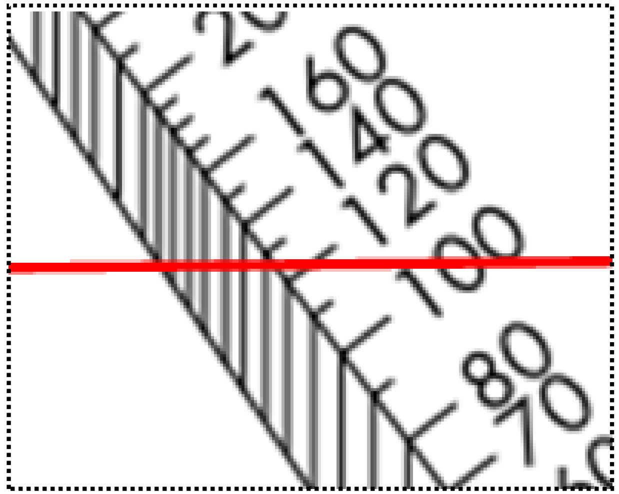


Figure 3 Nomogram for the calculation of sample size.



n is somewhat
less than 130

→ n = 128 (see above)

Determining the sample size with G*Power

Example independent samples t-test

Study on a new medicine to lower blood pressure

Features of the research project

- New medicine compared to the current one → analysis with t-test
- Dependent variable: Blood pressure measured in mmHg
- Grouping variable: Treatment group and control group
- Significance level α = 5% / Power = 1 – β = 80%

Clinically important difference δ – importance / strength

- A reduction of 10 mmHg is clinically important

Other data

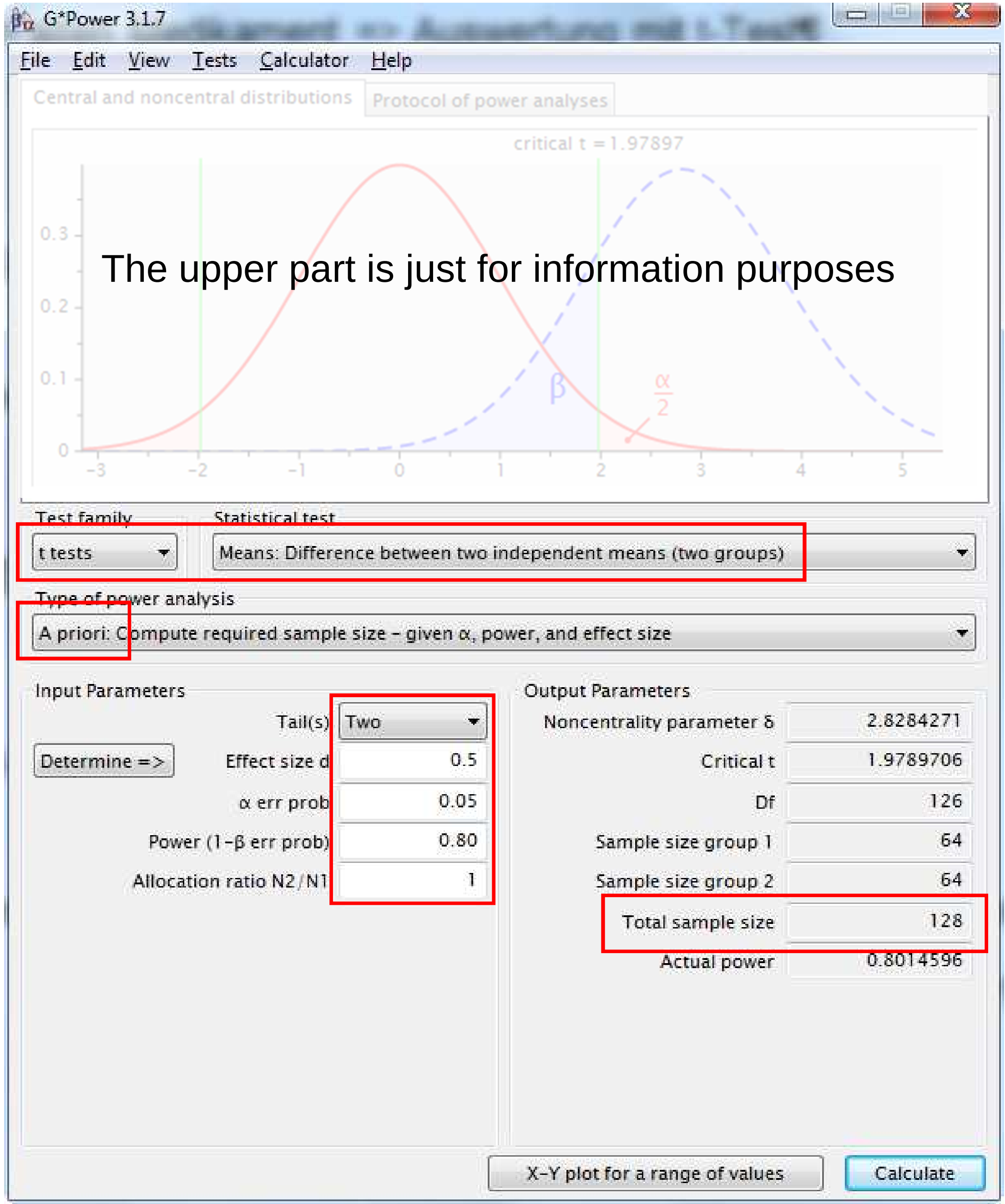
- Standard deviation of blood pressure in the population: σ_0 = 20 mmHg

Effect size

- $$d = \frac{\text{Distance } \delta}{\sigma_0} = \frac{10 \text{ mmHg}}{20 \text{ mmHg}} = 0.5$$

Question concerning the power analysis

- Sample size for significant t-test for independent samples?



A priori …

Determining the sample size before the study

Allocation ratio N2/N1 = 1

Assumption: Sample sizes N2 and N1 are equal in size

Result: n = 128

Determining the sample size with R

Example t-test for independent samples using *Collection of R code*

```
library(pwr)

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
```

number in each group = 63.76561 ≈ 64 → **sample size n = 2 x 64 = 128**

With a sample size of 64 probands in the control group and a sample size of 64 probands in the treatment group, a change in blood pressure of at least 10 mmHg can be tested significantly with a t-test for independent samples.

Preview of Lecture 07

What has happened so far

The result of a hypothesis test depends on the **sample size**, among other things.

The **effect size** according to Cohen (1992) introduces a new empirical concept.

The variables around the hypothesis test (sample size, significance level α, Power 1- β, effect size) are related to each other by the **power analysis**.

The **power analysis** makes it possible to determine the sample size required for a study.

What follows in Lecture 07

So far, the subject matter is based on the paradigm of "classical" **quantitative research**.

What is the role of **data driven research** and of **«Big Data»**?

The **paradigms** are considered and related to each other.

Instead of a tutorial, there will be an introduction to analysis of variance.

Analysis of variance is an important statistical analysis procedure for Design of Experiments.

Appendix

t-test with R in *Collection of R code*

```
library(readxl)

blood <- read_excel("blood.xlsx")

t.test(pressure ~ group, data = blood, var.equal = TRUE)
```

```
  Two Sample t-test

data:  pressure by group
t = 2.9011, df = 126, p-value = 0.00439
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  2.40266 12.71515
sample estimates:
mean in group 0 mean in group 1
       146.1947        138.6358
```

Table of contents

Introduction 3
Research question: Body height of women 3
Effect size 9
Study on the effect of a new medicine that helps to lower blood pressure 9
Effect size according to Cohen (1992) – Example: Cohen's d 11
Power analysis / Determining the sample size 15
Determining the sample size with G*Power 19
Determining the sample size with R 21
Preview of Lecture 07 22
What has happened so far 22
What follows in Lecture 07 22
Appendix 23