Statistical Testing and Sample Size Calculation Basics Part I

06.03.2024

Christina Abele
Center for Thrombosis and Hemostasis
University Medical Center, Mainz
christina.abele@uni-mainz.de

Agenda

- Research question
- Statistical testing
- Confidence intervals
- Sample size planning

Agenda

- **Research question**
- Statistical testing
- Confidence intervals
- Sample size planning

Research question

Imagine you have developed a new drug against pulmonary hypertension. There are many possible research questions:

- Efficacy compared to the standard treatment:
  - Is the new drug more effective than standard treatment?
  - Is the new drug different from standard treatment with respect to efficacy?
  - Is the new drug at least almost as effective as standard treatment?
- Meaning of „more effective“:
  - Patients live longer.
  - It takes longer to reach a certain level of disease progression.
  - Clinical parameters (6MWT distance, mPAP, ...) or patient reported outcomes (QoL, dyspnea, ...) are better on average.
  - Improvement reaches a certain level in more patients.
- Target population

Research question

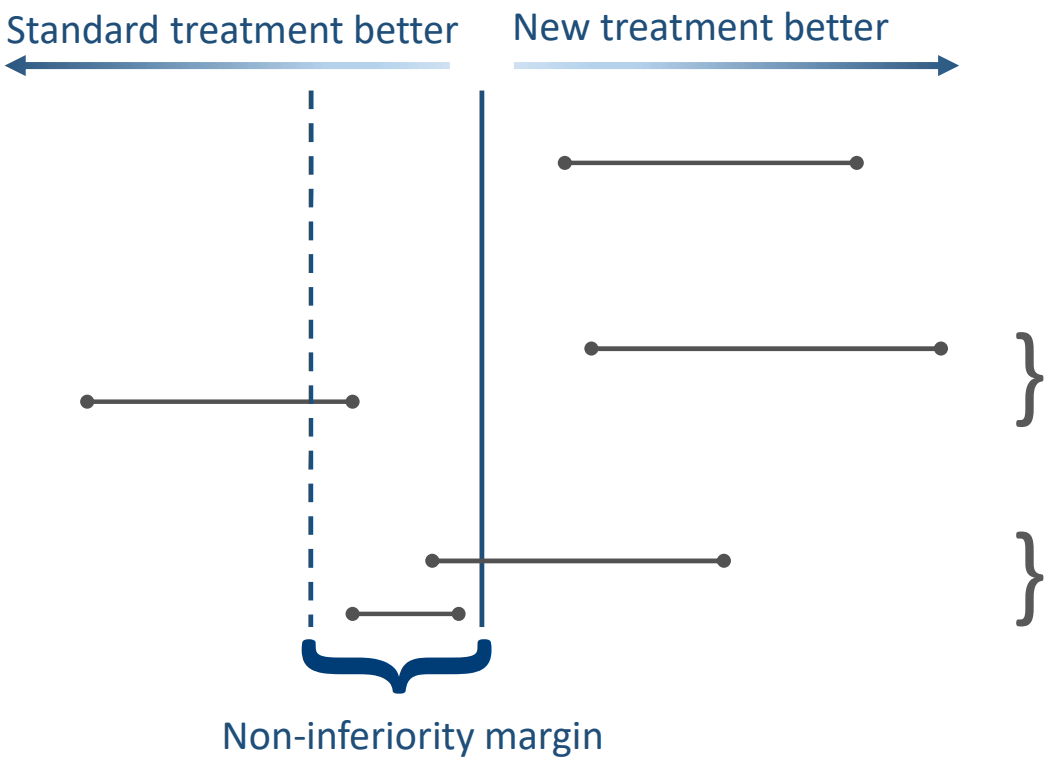
Imagine you have developed a new drug against pulmonary hypertension. There are many possible research questions:

- Efficacy compared to the standard treatment:
  - Is the new drug more effective than standard treatment?
  - Is the new d… …to efficacy?
  - Is the new d… …nt?
- Meaning of …
  - Patients live longer.
  - It takes longer to reach a certain level of disease progression.
  - Clinical parameters (6MWT distance, mPAP, ...) or patient reported outcomes (QoL, dyspnea, ...) are better on average.
  - Improvement reaches a certain level in more patients.
- Target population

This is the most crucial step in the whole process.

Research question

Testing scenarios:



In this course:
Focus on testing for difference.

„More effective“
Test for **superiority** ✓

„Different with respect to efficacy“
Test for **difference** ✓

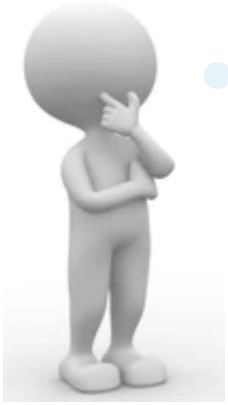
„At least almost as effective“
Test for **non-inferiority** ✓

Agenda

- Research question
- **Statistical testing**
- Confidence intervals
- Sample size planning

Statistical testing

Does the new drug lead to a different average mPAP after 1 month compared to standard treatment?

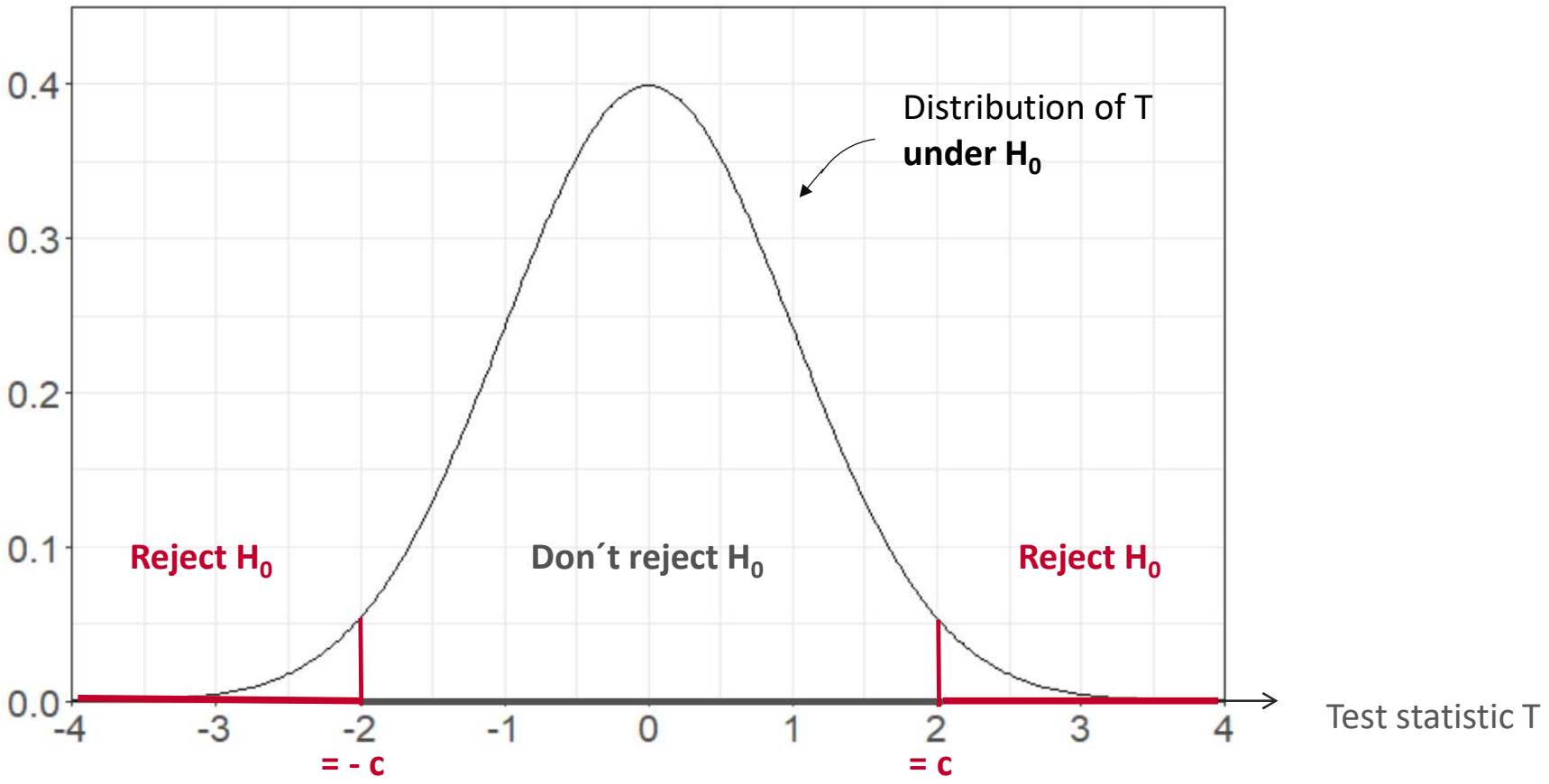


What is our null hypothesis?

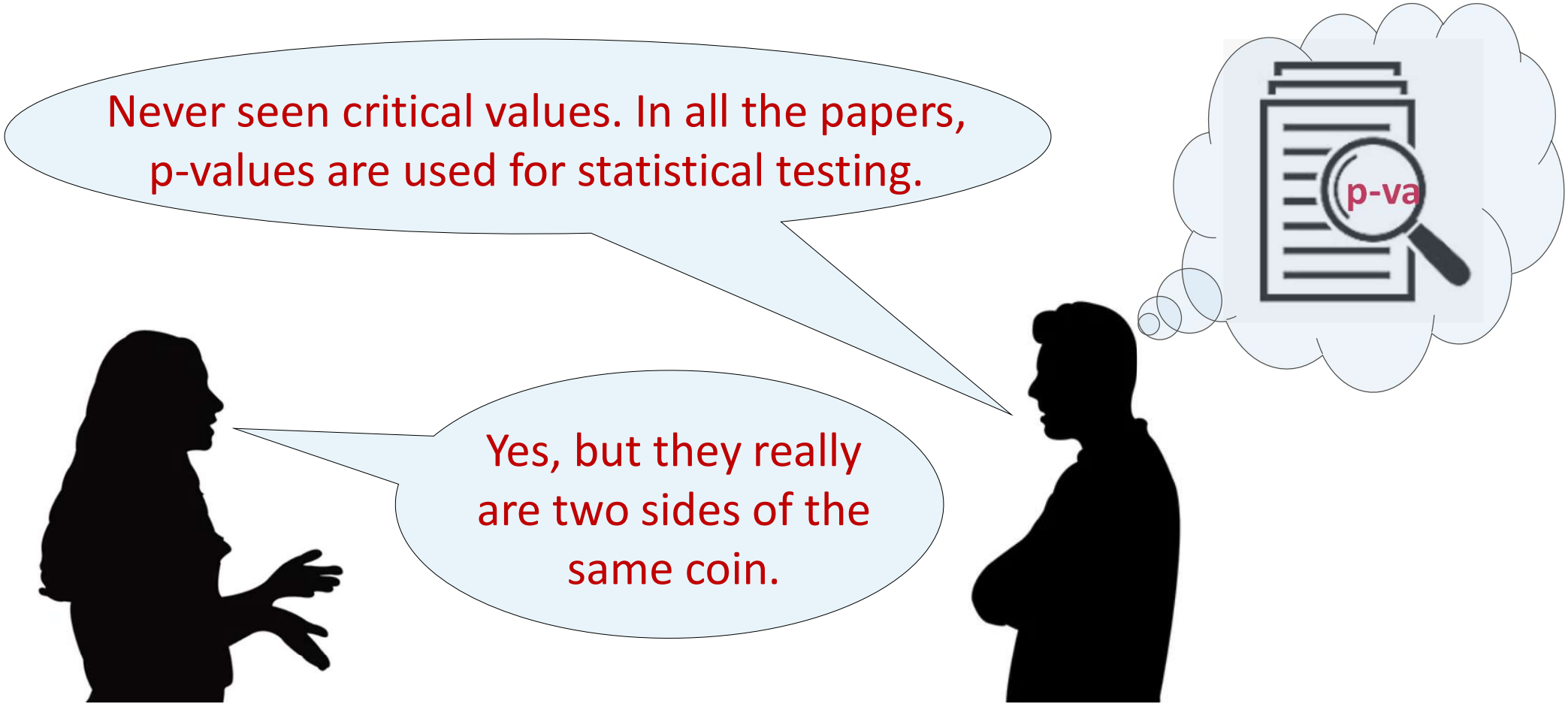
Statistical testing

- Basic principle of statistical testing: **Proof by contradiction**
- Step 1:
  Frame the **null hypothesis** H_0. It always states the **opposite** of what we want to show (this is the **alternative** hypothesis H_1)!
- Step 2:
  Measure outcomes in a sample representative of the target population. Assess how well the data observed agree with H_0. To quantify this, we use the **test statistic** T: The closer T is to zero, the better the data agree with H_0.
- Step 3:
  If the distance between T and zero exceeds the **critical value** c, the data agree so badly with H_0 that we are sure H_0 must be false and reject it.
  $\Rightarrow$ „The result of the test was significant.“ / „H_1 could be confirmed.“

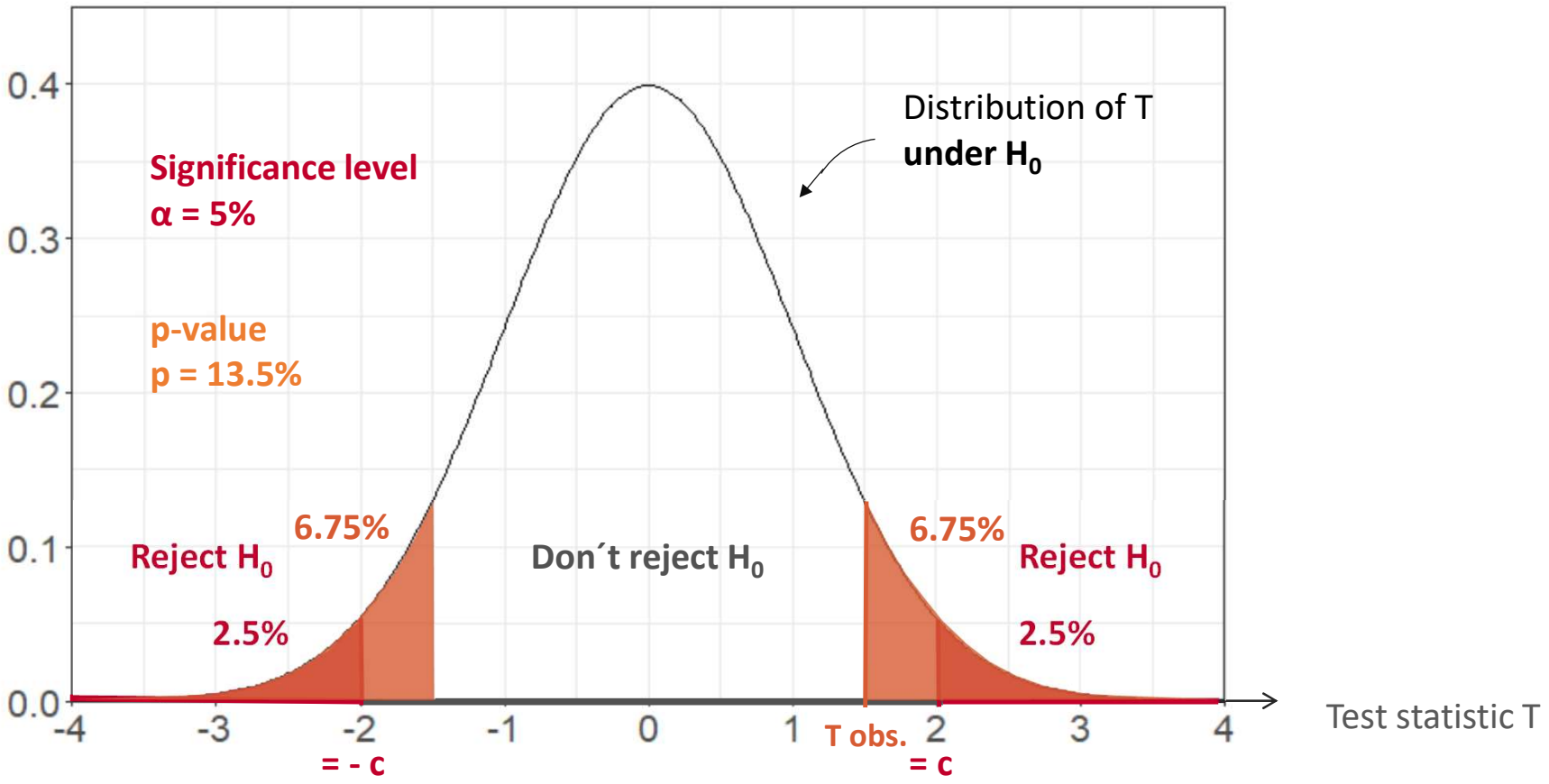
Statistical testing



Statistical testing



Statistical testing



Statistical testing revisited

- Step 1:
  Frame the **null hypothesis** H_0.
- Step 2:
  Answer the following question:
  „**Assuming H_0 is true**, how likely is it to obtain data whose corresponding test statistic is at least as far away from zero as with the observed data?“
  This likelihood is the **p-value**. If the p-value is small, the observed data agree badly with H_0.
- Step 3:
  The limit of this likelihood below which we assume H_0 is false and reject it is the **significance level** α of the test. The lower it is, the stricter is the test.
- The p-value equals α just if the test statistic equals the critical value.

Statistical testing

What can go wrong?

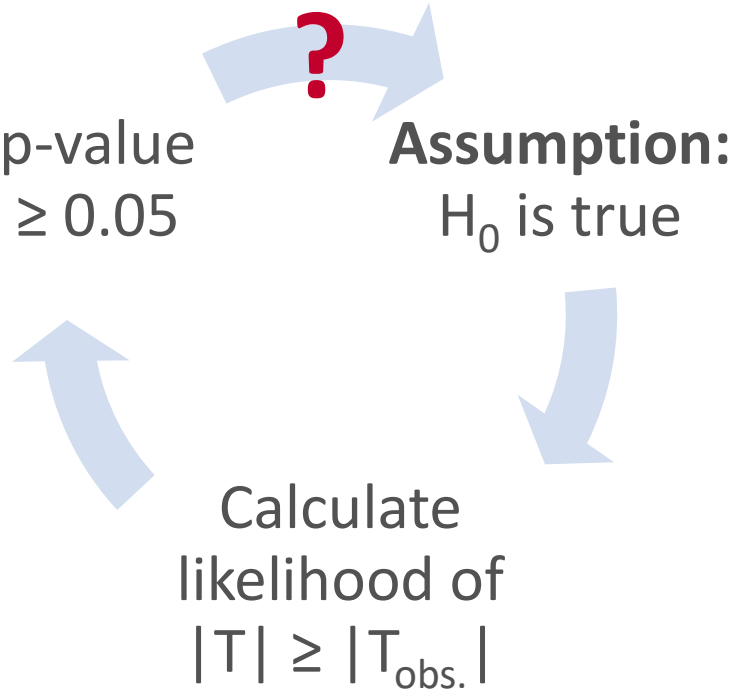
	H_0 is rejected H_1 is confirmed	H_0 is not rejected H_1 can not be confirmed
H_0 is true **H_1 is false**	**Type I error** (false positive rate, α error): $\Rightarrow$ Fix **level of significance** in order to control for it	OK
H_0 is false **H_1 is true**	OK	**Type II error** (false negative rate, β error): $\Rightarrow$ Use **sufficiently large samples** to get the power $1 - \beta$

Statistical testing

$p < \alpha$ is evidence for the alternative H_1. Is $p \geq \alpha$ evidence for H_0?

NO! Logically, this would be circular reasoning:

In this case we can only state that we can´t reject H_0 / can´t confirm H_1
$\Rightarrow$ **Further research is necessary!**



Agenda

- Research question
- Statistical testing
- **Confidence intervals**
- Sample size planning

Confidence intervals

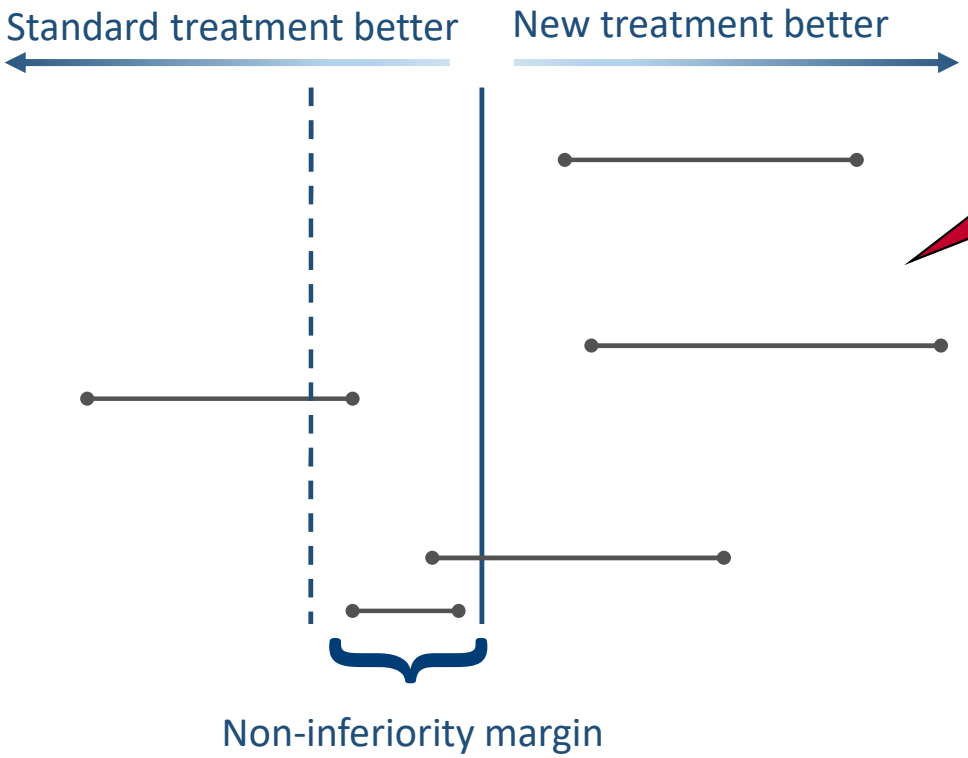
- Observed data permit not only hypothesis testing, but also estimation of quantities of interest (e. g. difference of means, odds ratio, ...). The estimated value is called **point estimate**.
- Due to random variation, different samples from the same population will lead to different point estimates.
- So how is the true value related to its point estimate? Or: How precise is the estimation?
- This question is answered by the concept of the **confidence interval** (CI). It is a range around the point estimate, computed at a specified **confidence level** (e. g. 95%).
- The confidence level is the chance of obtaining a CI containing the true value.

Confidence intervals

- Often, there is a mathematical relationship between test statistics and point estimates. In such cases, CIs can be constructed by inverting hypothesis tests if the distributions under H_1 are known or can be estimated.
- A test at significance level α corresponds to an interval at confidence level 1-α.
- This opens up a **third way of significance testing**:
  - Step 1:
    Compute the (1-α)-CI for the quantity of interest.
  - Step 2:
    Check whether the CI contains the value the quantity of interest would have under the null hypothesis (0 for differences, 1 for ratios).
  - Step 3:
    Reject the null hypothesis if the CI doesn´t contain this value.

Research question

Testing scenarios:



The „ranges“ on slide 7 were nothing but confidence intervals.

…ore effective“
Test for **superiority** ✓

„Different with respect to efficacy“
Test for **difference** ✓

„Almost as effective“
Test for **non-inferiority** ✓

Agenda

- Research question
- Statistical testing
- Confidence intervals
- **Sample size planning**

Sample size planning

- Decide on **relevant alternative H_1**.
  (What is „relevant“? Not a statistical, but a medical/biological question!)
- Fix **power** = probability of getting a significant result under H_1.
- For all practically relevant test statistics T, the distribution of T depends on sample size N. Power usually grows with increasing N. Find the value of N at which the desired power is obtained.
- Simple trial designs: Sample size formulas
- More complex trial designs: Sample size planning by simulation (Monte Carlo method)