Three-sided Hypothesis Testing

Simultaneous Testing of Superiority, Equivalence and Inferiority

Jelle Goeman Aldo Solari Theo Stijnen

Medical Statistics & Bioinformatics
Leiden University Medical Center

Ribno, 2010-09-21

Multiple testing in one-parameter models

Jelle Goeman Aldo Solari Theo Stijnen

Medical Statistics & Bioinformatics
Leiden University Medical Center

Ribno, 2010-09-21

Outline

Outline

A simple testing problem?

Null hypothesis

$H_0 : \mu = 0$

Alternative hypothesis

$H_A : \mu \neq 0$ (two-sided)

A simple testing problem?

Null hypothesis

$H_0 : \mu = 0$

Alternative hypothesis

$H_A : \mu \neq 0$ (two-sided)

The test result

p-value $< \alpha$

Estimate: $\hat{\mu} > 0$

A simple testing problem?

Null hypothesis

$H_0 : \mu = 0$

Alternative hypothesis

$H_A : \mu \neq 0$ (two-sided)

The test result

p-value $< \alpha$
Estimate: $\hat{\mu} > 0$

Our conclusion?

- We conclude $\mu \neq 0$?
- We conclude $\mu > 0$?

Classical point of view

Classical Neyman-Pearson theory

- We should conclude: Reject $H_0 : \mu = 0$
- Concluding $\mu > 0$ is post hoc $\rightarrow$ may inflate error level?

Classical point of view

Classical Neyman-Pearson theory

- We should conclude: Reject $H_0 : \mu = 0$
- Concluding $\mu > 0$ is post hoc $\rightarrow$ may inflate error level?

Directional error

Correct rejection of H_0 but false inference of the sign of the parameter

Also known as

Type III errors (Kaiser 1967)

This talk

Conclusion (well-known)

Without inflating error levels we may reject both $\mu = 0$ and $\mu < 0$

This talk

Conclusion (well-known)

Without inflating error levels we may reject both $\mu = 0$ and $\mu < 0$

But additionally

Without inflating error levels
We may sometimes reject $\mu < 0$ if we fail to reject $H_0 : \mu = 0$

This talk

Conclusion (well-known)

Without inflating error levels we may reject both $\mu = 0$ and $\mu < 0$

But additionally

Without inflating error levels
We may sometimes reject $\mu < 0$ if we fail to reject $H_0 : \mu = 0$

How?

By making use of the latest developments in multiple testing

A multiple testing perspective

Multiple inferences

We want to reject not only $\mu = 0$, but also $\mu > 0$ or $\mu < 0$

Type I error

Committed in case of any false inference among all inferences made

Probability of a type I error

Familywise error rate

Outline

Closed testing (Marcus, Peritz, Gabriel, 1976)

Closure

- Create all intersection hypotheses of original hypotheses
- Example: H_1, H_2, $H_3 \rightarrow$
  H_1, H_2, H_3, $H_1 \cap H_2$, $H_1 \cap H_3$, $H_2 \cap H_3$, $H_1 \cap H_2 \cap H_3$
- Test all hypotheses at level α

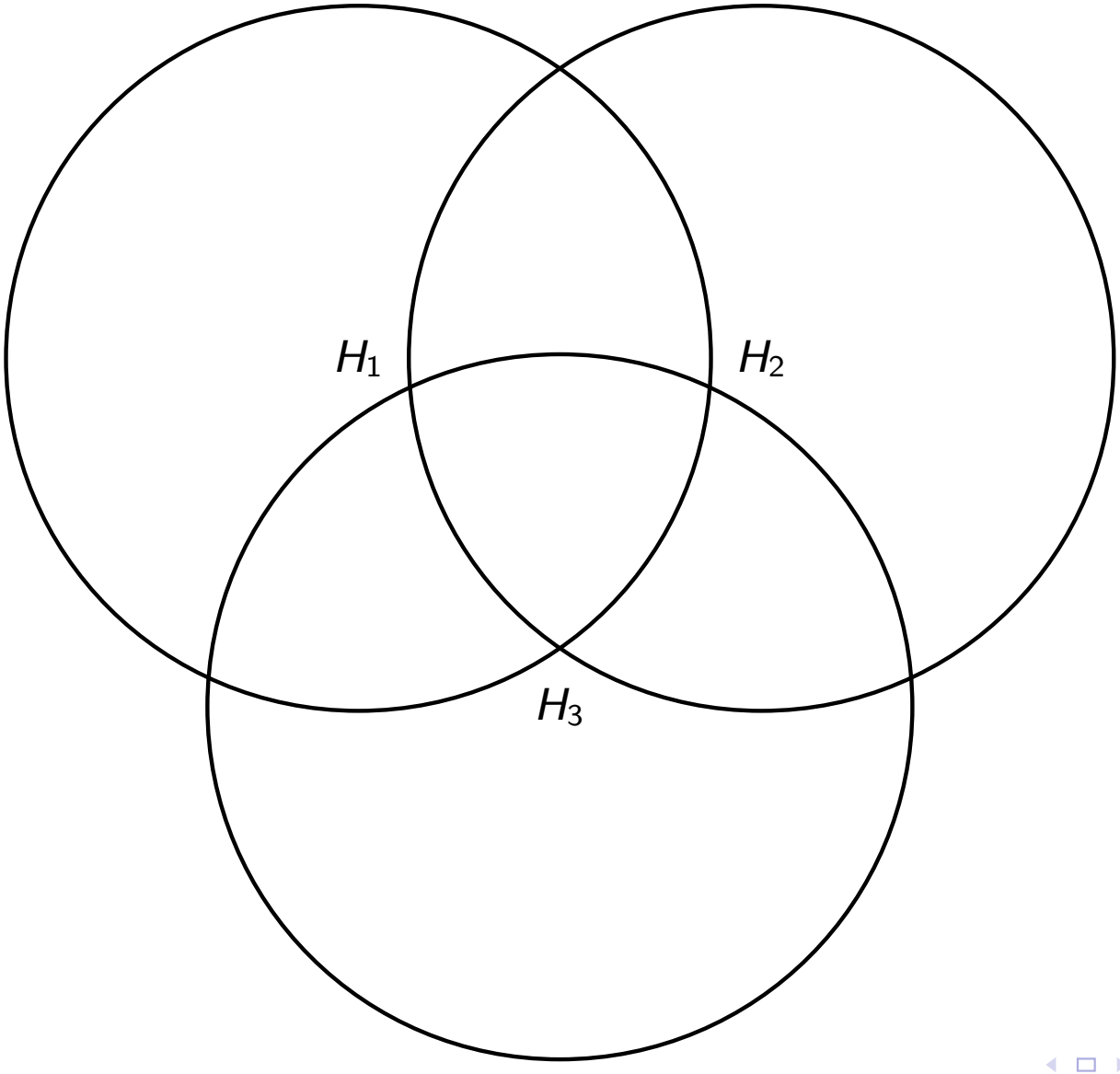
Reject hypothesis H if

All intersection hypotheses $\subseteq H$ are rejected

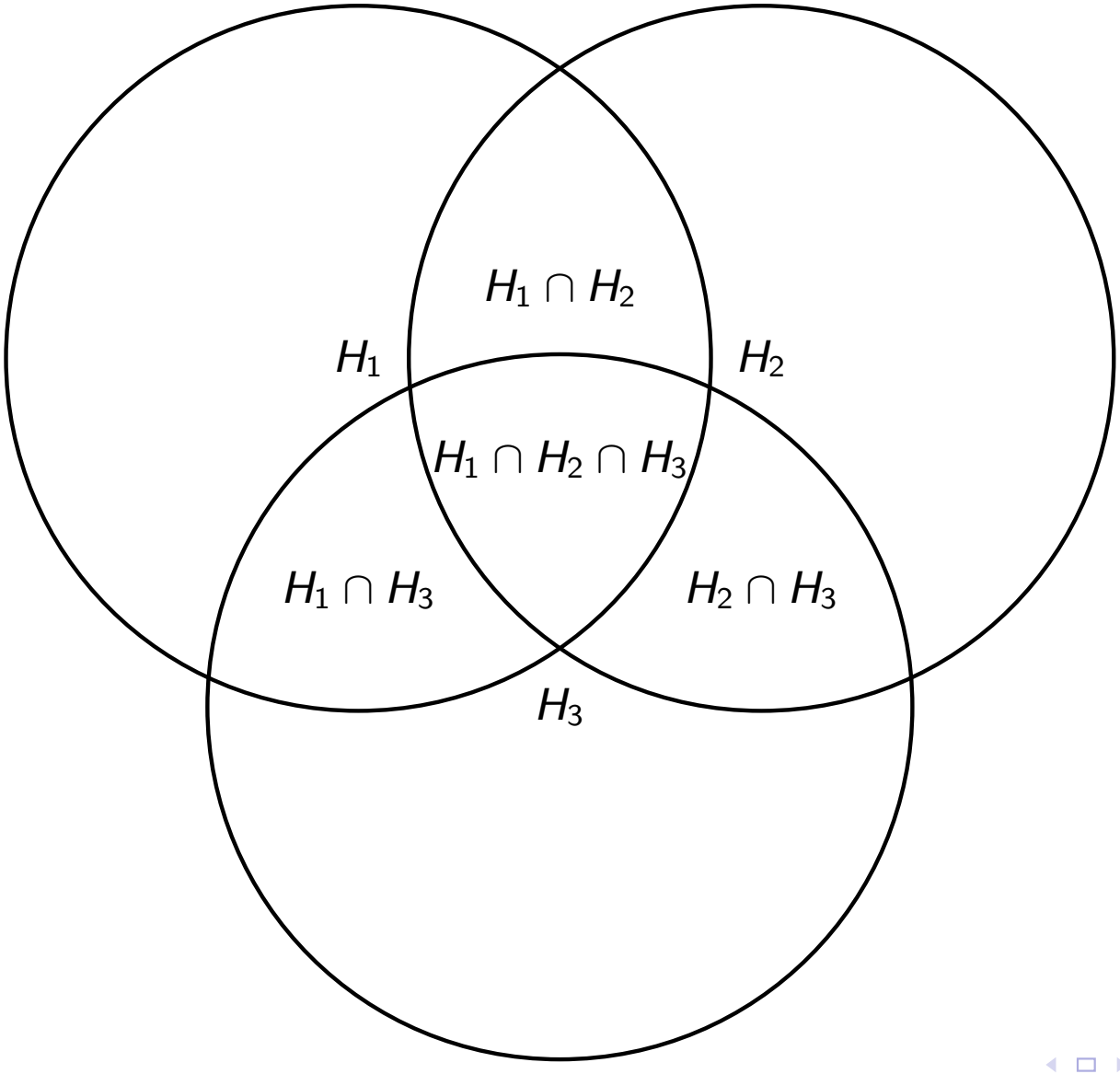
Control

Strong control of FWER at level α

Closed testing (graphically)



Closed testing (graphically)



Directional errors via closed testing

Two hypotheses

$$\begin{aligned} H_{0+} &: \mu \geq 0 \\ H_{0-} &: \mu \leq 0. \end{aligned}$$

Intersection hypotheses

$H_0 : \mu = 0$ is $H_{0+} \cap H_{0-}$

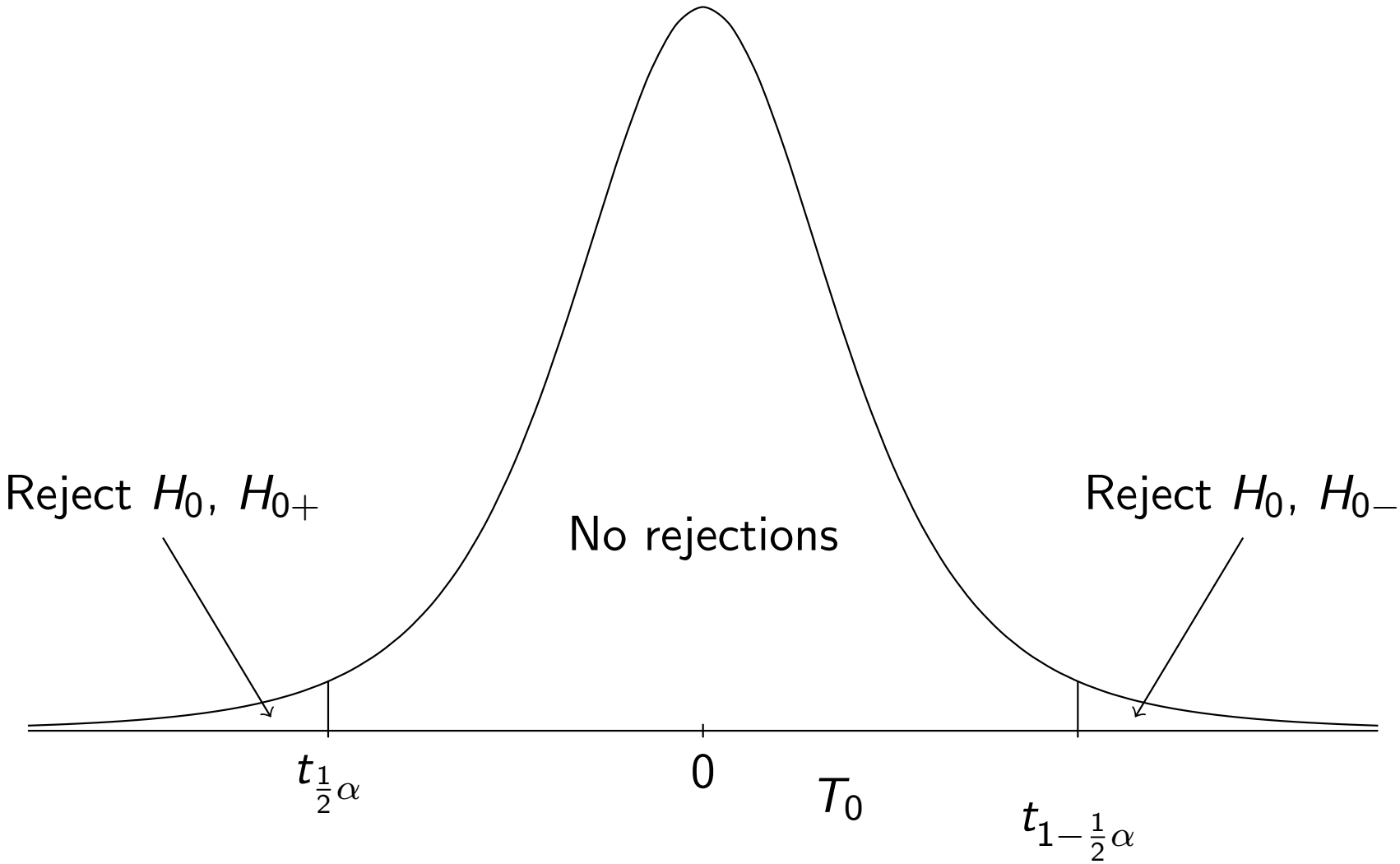
Closed testing

- Test H_0 with a two-sided test
- Test H_{0+} with a one-sided test (left)
- Test H_{0-} with a one-sided test (right)

By closed testing

Start testing H_0. If significant, go on with H_{0+} and H_{0-}

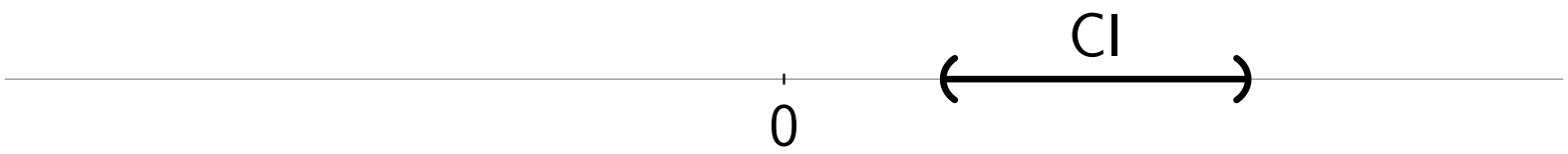
Closed testing results: diagram



Equivalent: Confidence interval based approach

CI based approach

- Make a two-sided confidence interval (l_μ, u_μ) for μ
- If $l_\mu \geq 0$: reject H_0 and H_{0-}
- If $u_\mu \leq 0$: reject H_0 and H_{0+}



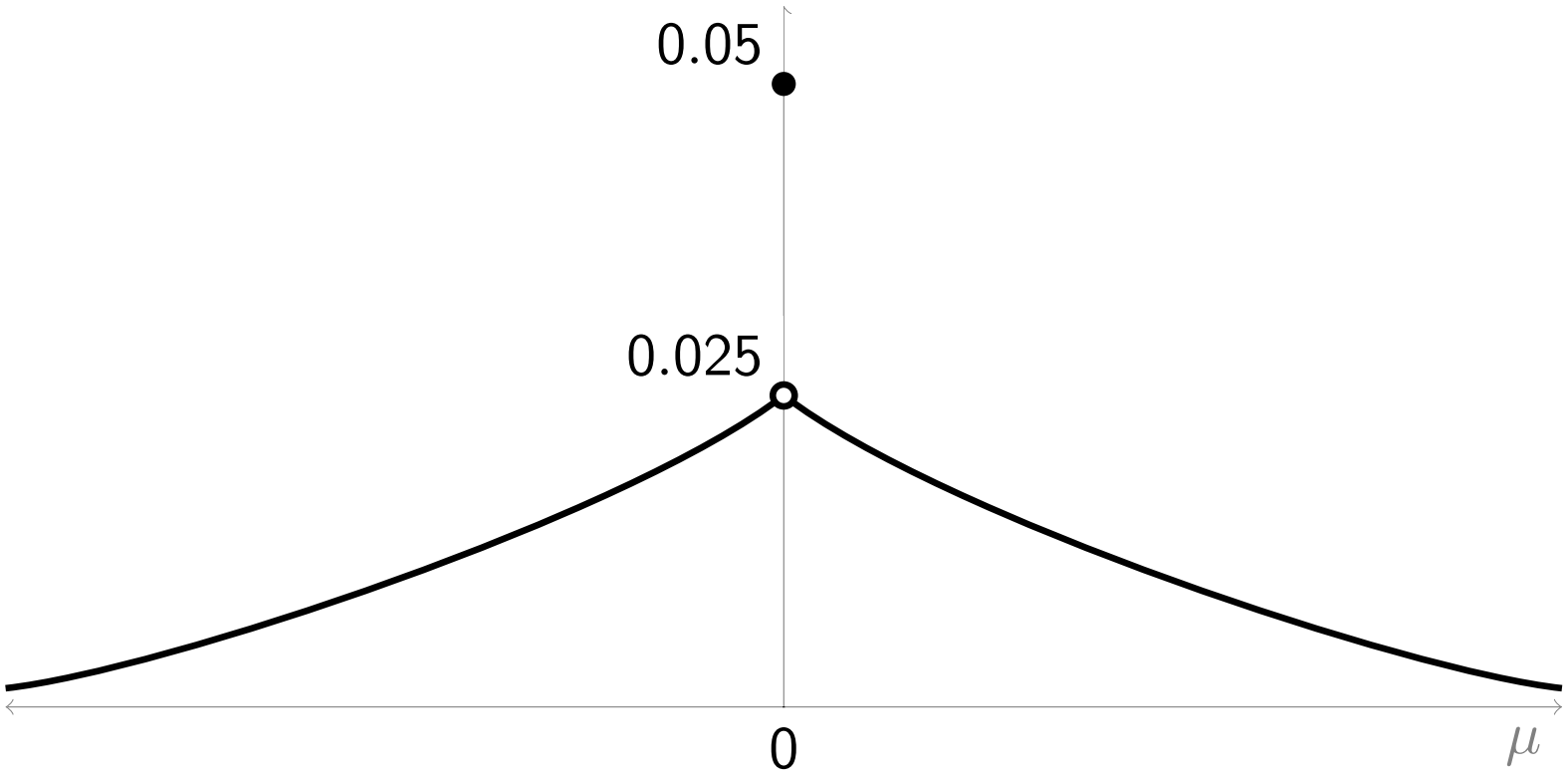
Equivalent

To the results of a closed testing approach

Room for improvement

Probability of a directional error

As a function of true μ



Bonferroni and Shaffer

Set-up

p-values $p_1, \ldots, p_m$ for hypotheses $H_1, \ldots, H_m$

Bonferroni

Reject all H_i for which $p_i \leq \alpha/m$

Bonferroni and Shaffer

Set-up

p-values $p_1, \ldots, p_m$ for hypotheses $H_1, \ldots, H_m$

Bonferroni

Reject all H_i for which $p_i \leq \alpha/m$

Restricted combinations

If no more than $k < m$ hypotheses can be simultaneously true

Bonferroni and Shaffer

Set-up

p-values $p_1, \ldots, p_m$ for hypotheses $H_1, \ldots, H_m$

Bonferroni

Reject all H_i for which $p_i \leq \alpha/m$

Restricted combinations

If no more than $k < m$ hypotheses can be simultaneously true

Shaffer

Reject all H_i for which $p_i \leq \alpha/k$

The partitioning principle

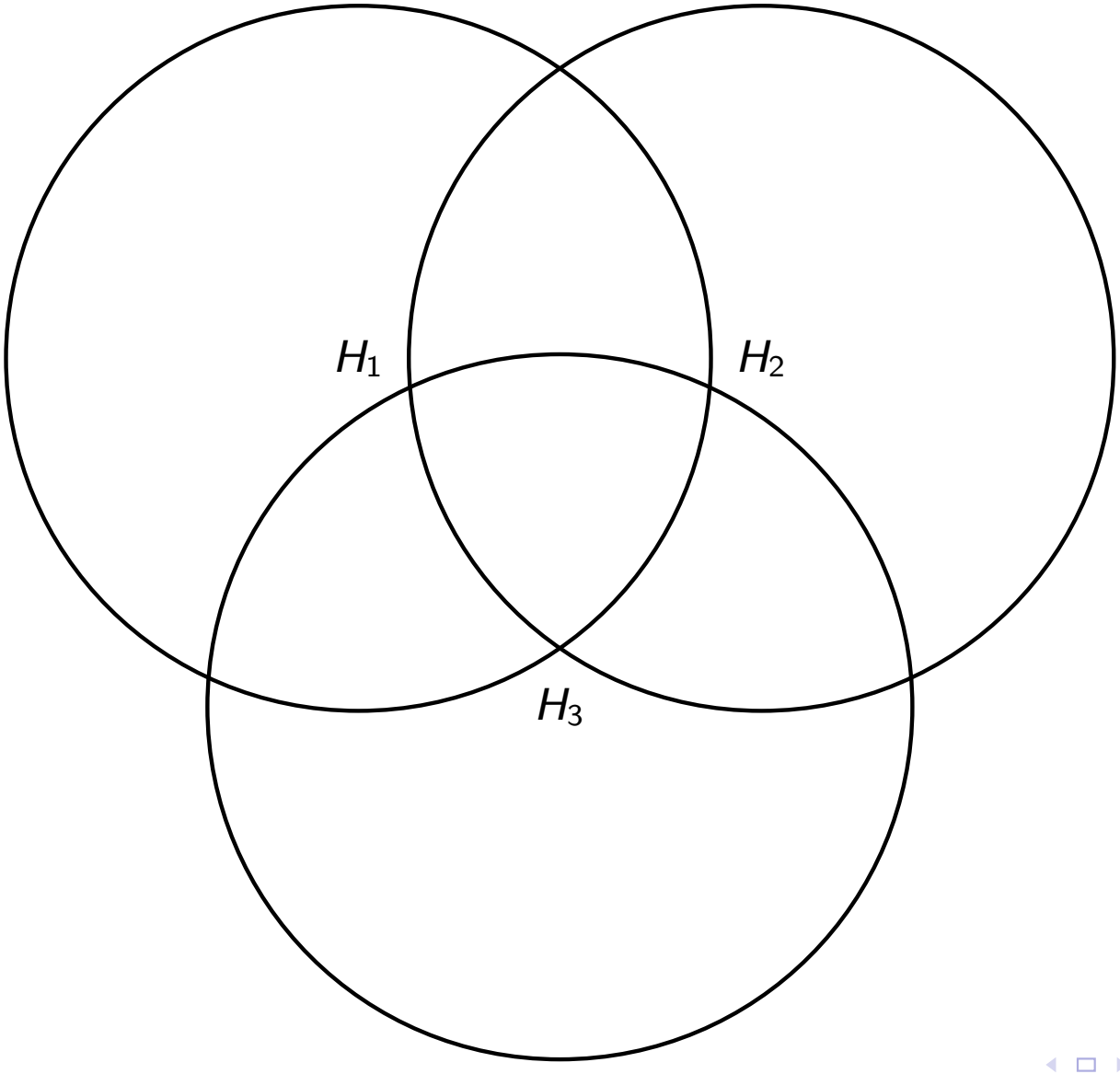
Partitioning principle (Finner and Strassburger, 2002)

- Disjoint hypotheses: no multiple testing correction needed
- Do all tests at level α and still control FWER
- Reason (Shaffer): at most one hypothesis can be true

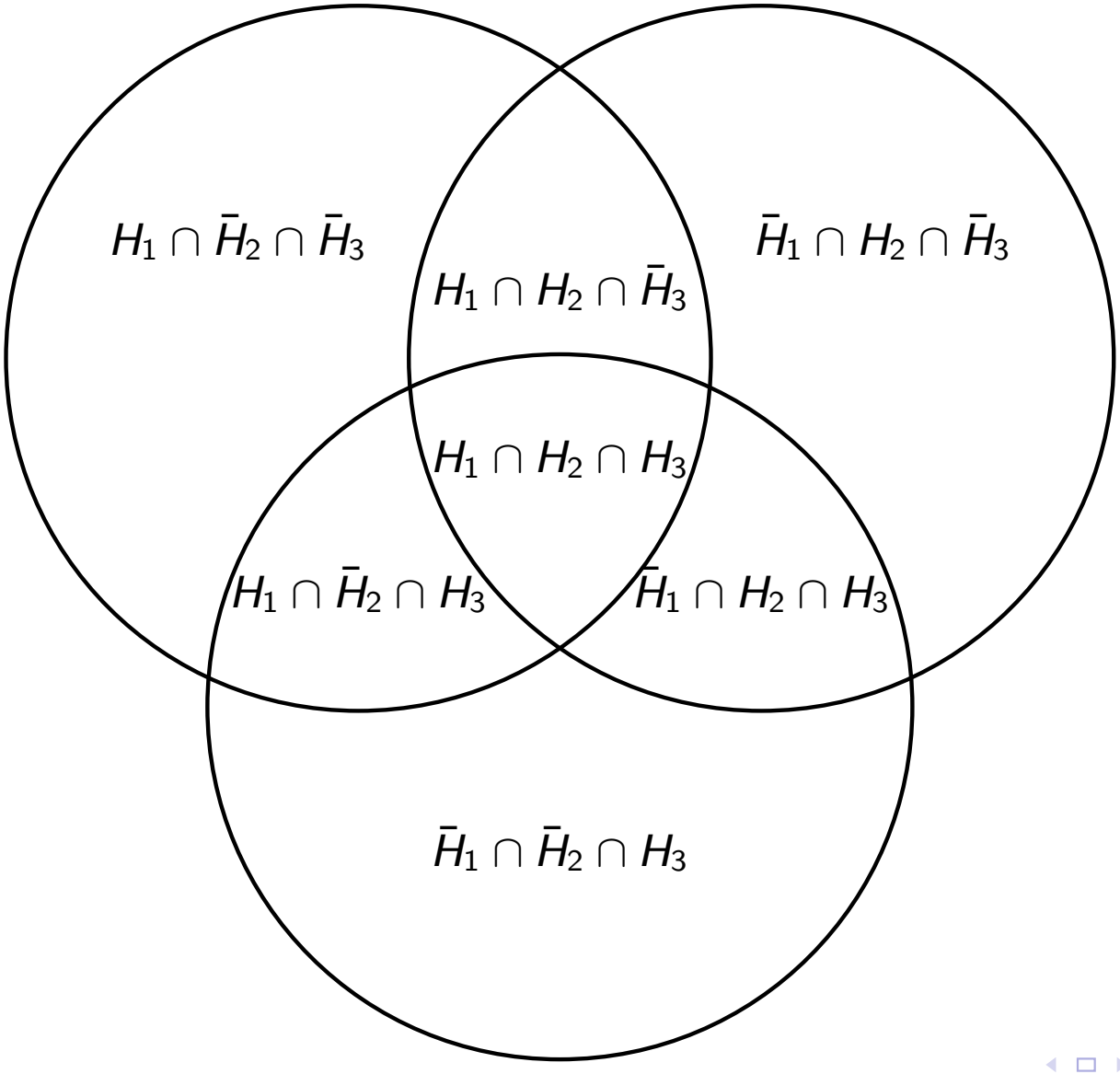
Partitioning: recipe

- Partition parameter space into disjoint subhypotheses
- Test disjoint hypotheses at level α
- Reject original hypotheses if all component parts are rejected

The partitioning principle (graphically)



The partitioning principle (graphically)



Partitioning as a principle

Fundamental

Every known FWER control procedure is a special case of partitioning

Closed testing

Partitioning uniformly improves on closed testing

Disjoint hypotheses

Define three hypotheses

$$
\begin{array}{lcll}
H_0 & : & \mu = 0 & \text{(equivalence)} \\
H_+ & : & \mu > 0 & \text{(superiority)} \\
H_- & : & \mu < 0 & \text{(inferiority).}
\end{array}
$$

Disjoint hypotheses

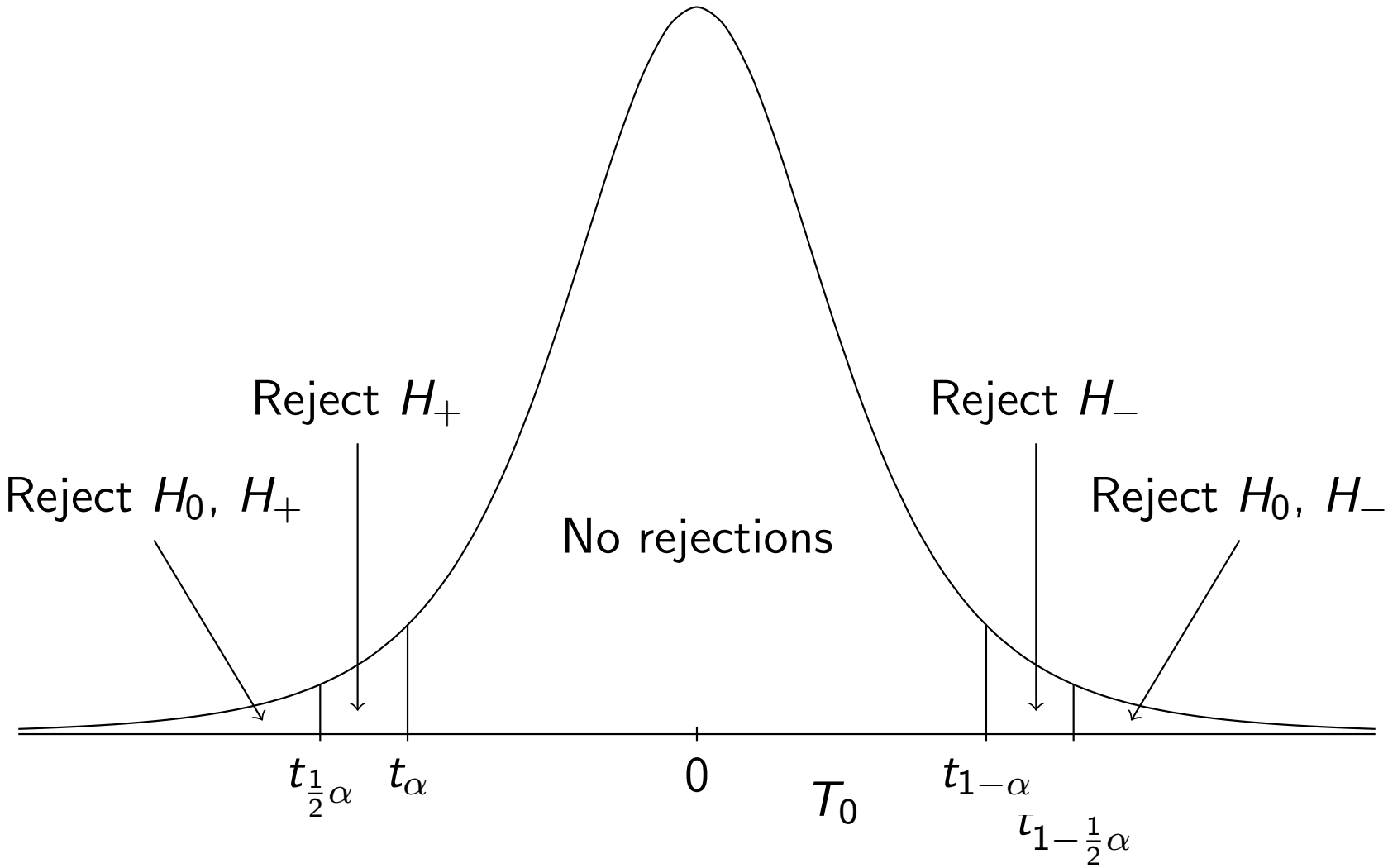
Define three hypotheses

$$\begin{aligned} H_0 &: \mu = 0 && \text{(equivalence)} \\ H_+ &: \mu > 0 && \text{(superiority)} \\ H_- &: \mu < 0 && \text{(inferiority).} \end{aligned}$$

Three-sided testing

- Test H_0 with a two-sided test
- Test H_+ with a one-sided test (left)
- Test H_- with a one-sided test (right)

Three-sided testing: diagram



Three-sided testing

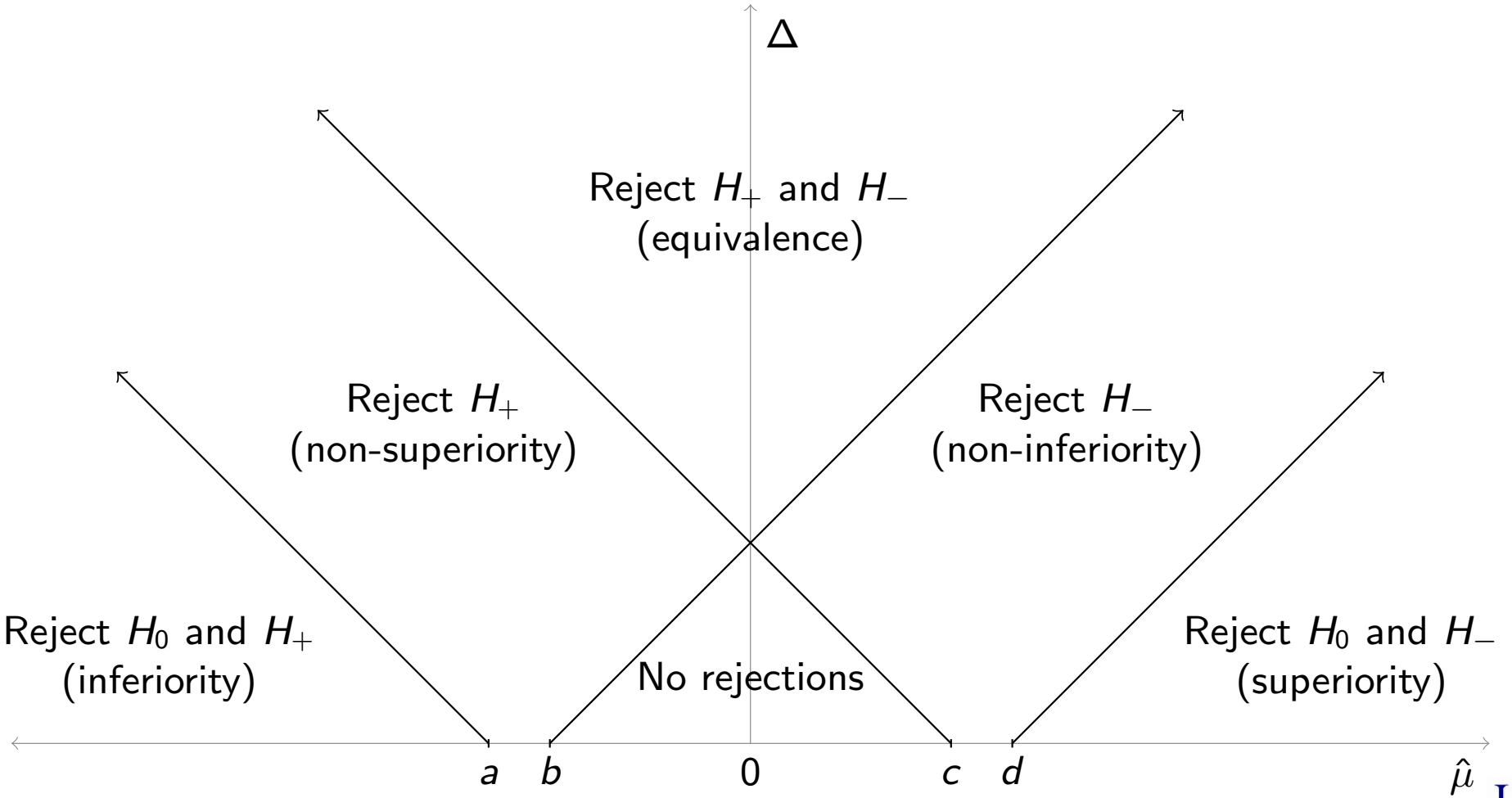
Equivalence margin

$\Delta > 0$

The three hypotheses

$$
\begin{array}{lcll}
H_0 & : & -\Delta \leq \mu \leq \Delta & \text{(equivalence)} \\
H_+ & : & \mu > \Delta & \text{(superiority)} \\
H_- & : & \mu < -\Delta & \text{(inferiority).}
\end{array}
$$

Three-sided testing: diagram



Outline

Free additional inference?

Additional inference

Sometimes H_+ or H_- rejected even if H_0 not rejected

Question

Does the additional inference come at a price?

Free additional inference?

Additional inference

Sometimes H_+ or H_- rejected even if H_0 not rejected

Question

Does the additional inference come at a price?

Answer

Yes: forget about the classical confidence intervals

Reminder: CI as inverted test

What is a confidence interval

- Test $H_x : \mu = x$ for every x
- Record which H_x get rejected
- Confidence interval: $\{x : H_x \text{ not rejected}\}$

Reminder: CI as inverted test

What is a confidence interval

- Test $H_x : \mu = x$ for every x
- Record which H_x get rejected
- Confidence interval: $\{x : H_x \text{ not rejected}\}$

Doing infinitely many tests

Multiple testing correction needed?

Reminder: CI as inverted test

What is a confidence interval

- Test $H_x : \mu = x$ for every x
- Record which H_x get rejected
- Confidence interval: $\{x : H_x \text{ not rejected}\}$

Doing infinitely many tests

Multiple testing correction needed?

Not necessary by the partitioning principle

Because all hypotheses H_x are disjoint

Tests to use for confidence intervals

What test to use

Confidence interval theory does not prescribe a test to use

Tests to use for confidence intervals

What test to use

Confidence interval theory does not prescribe a test to use

Standard confidence interval

- Uses a two-sided test for every $H_x : \mu = x$
- Not consistent with three-sided inference

Confidence intervals for three-sided testing

Question

What confidence interval is consistent with three-sided testing?

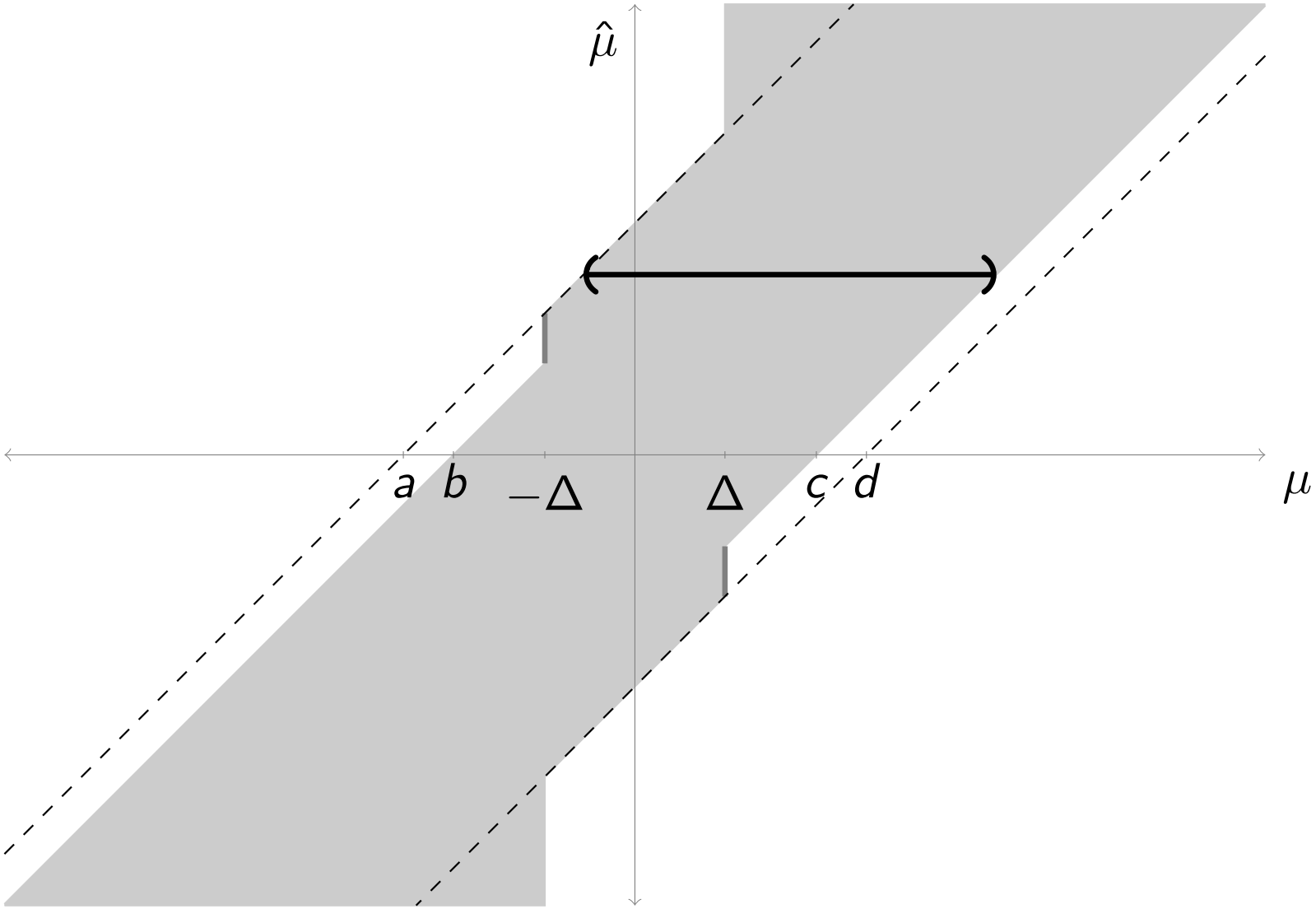
Inverted test

Test $H_x : \mu = x$ for every x

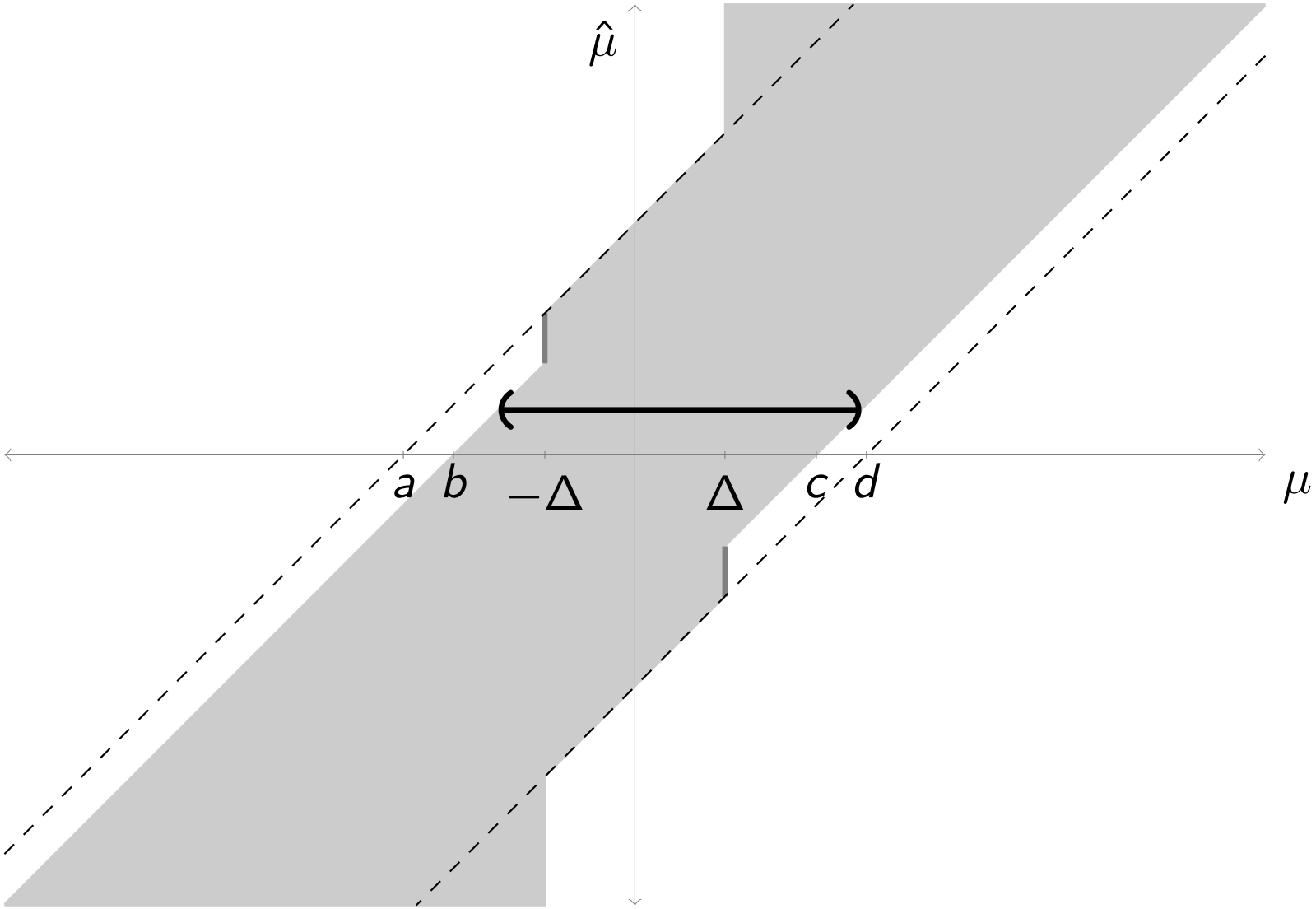
Use

- Two-sided tests for $-\Delta \leq x \leq \Delta$
- One-sided test (left) for $x < -\Delta$
- One-sided test (right) for $x > \Delta$

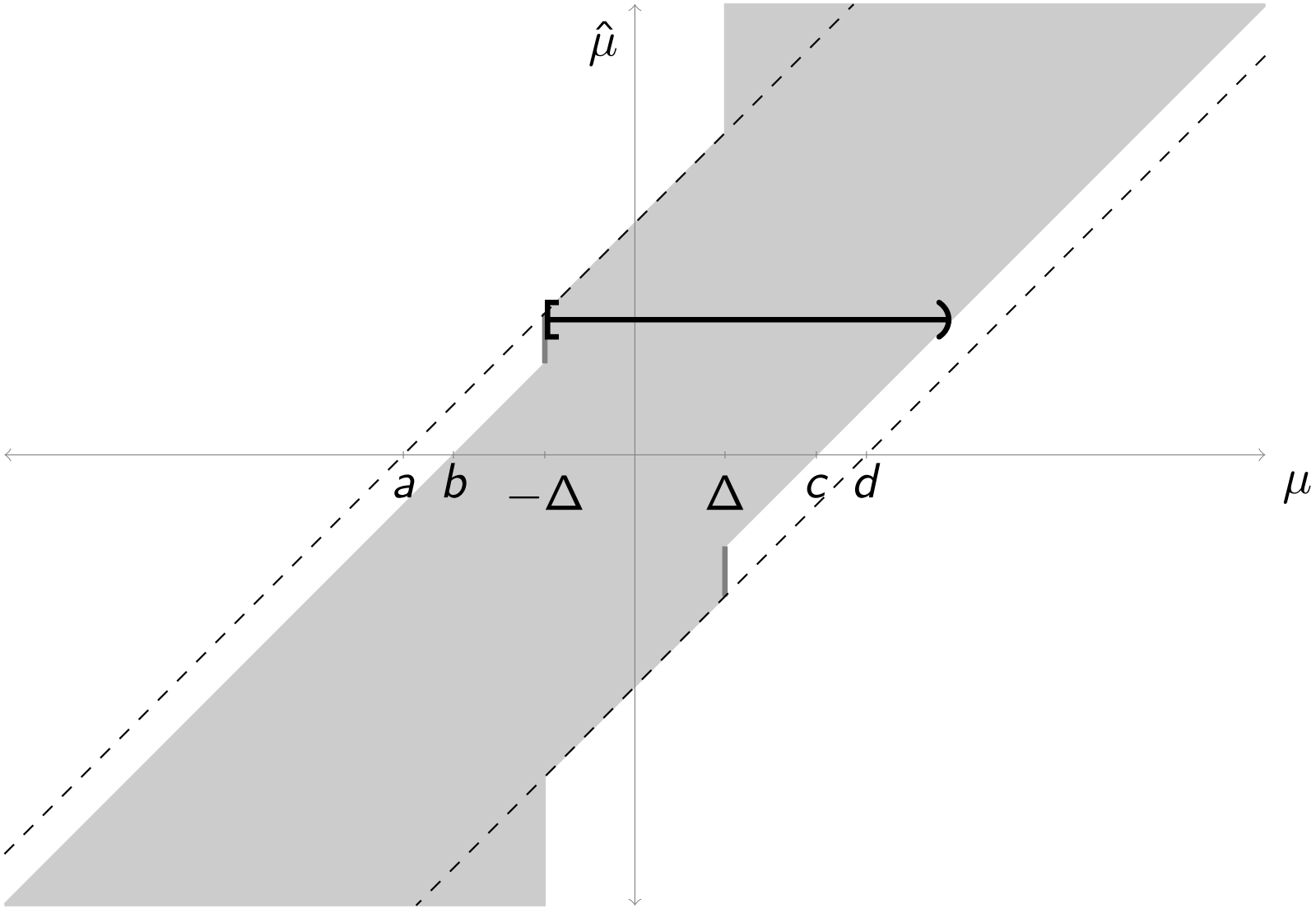
Confidence intervals: diagram



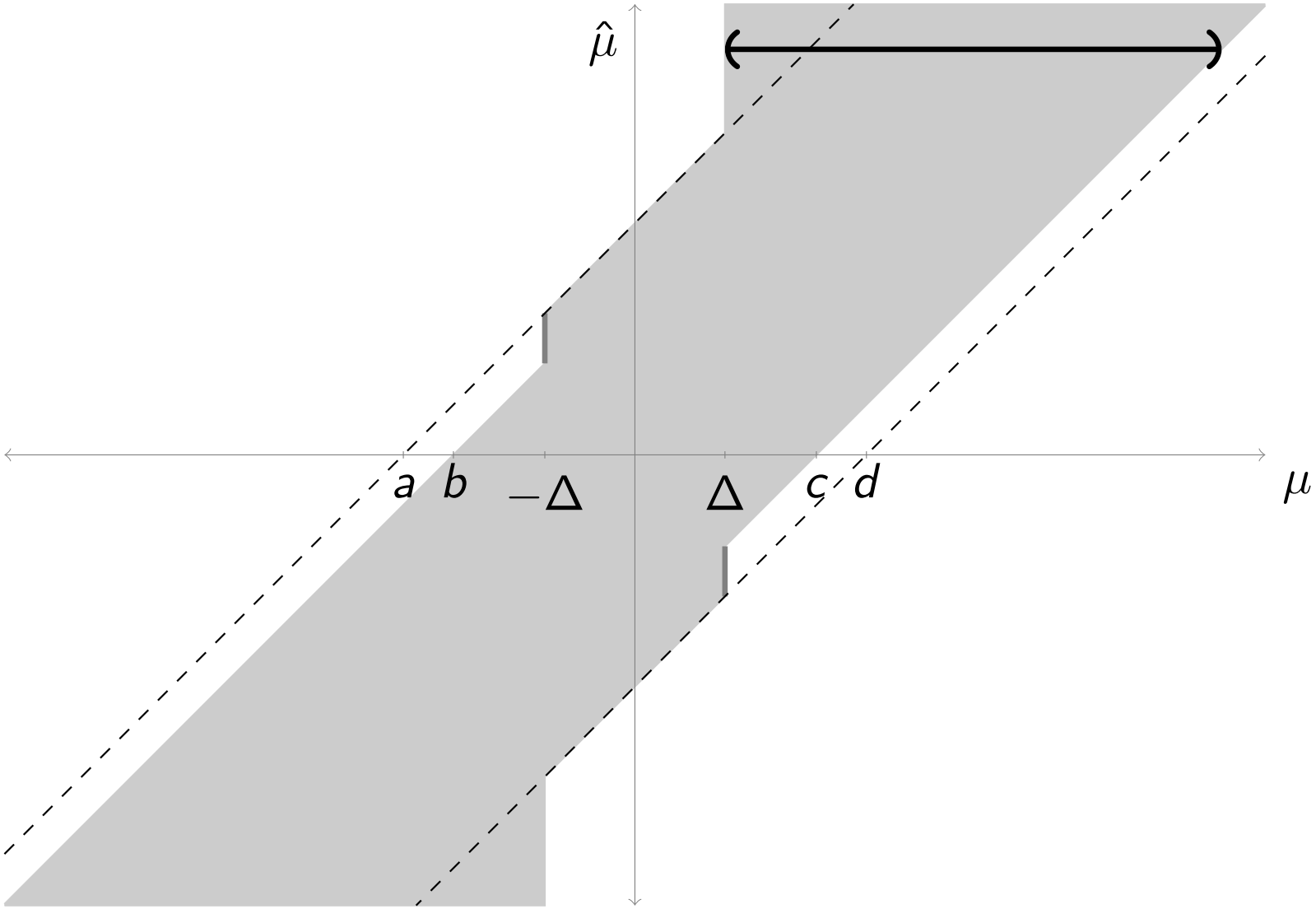
Confidence intervals: diagram



Confidence intervals: diagram



Confidence intervals: diagram



Confidence intervals: gain and loss

Comparison with the usual confidence interval

- Narrower if $-\Delta$ or Δ in classical CI
- Typically broader otherwise

Open and closed

CI sometimes is a half-closed interval $[a, b)$

Lower and upper bound

- Lower bound never above Δ
- Upper bound never below $-\Delta$

Three-sided inference based on regular CI

Consistent with classical CI

- Reject H_0 if CI $\cap H_0 = \varnothing$
- Reject H_+ if CI $\cap H_+ = \varnothing$
- Reject H_- if CI $\cap H_- = \varnothing$

Relative to 3-sided testing

Less powerful to reject H_+, H_-

$\longrightarrow$ less powerful to infer non-inferiority, non-superiority

Outline

Non-inferiority and superiority testing

Clinical trials often asymmetric

- Drug versus placebo
- New versus established treatment
- Drug without side effects versus drug with

Non-inferiority trials

New drug is not worse than established drug

Non-inferiority margin

New drug may be at most Δ worse than established drug

One-sided testing in clinical trials

Asymmetric set-up

- "Placebo outperforms drug" not interesting
- Consequence: one-sided test?
- One-sided testing not allowed by regulatory agencies

Regulatory guidelines

- One-sided tests should be performed at level $\alpha/2$
- Effectively: ban on one-sided tests

What's wrong with the one-sided test?

Post hoc abuse

Following up on a significant result in opposite direction

Suggestive prejudice

One-sided test does not treat placebo and treatment equally

What's wrong with the one-sided test?

Post hoc abuse

Following up on a significant result in opposite direction

Suggestive prejudice

One-sided test does not treat placebo and treatment equally

Symmetry

Interpretation of guidelines: prescribes symmetric procedures

Three-sided testing

Symmetric

- Not biased towards positive or negative
- Still: allows one-sided tests

Flexible

Type of trial (superiority, non-inferiority, equivalence) does not have to be declared beforehand

Choosing Δ

Non-inferiority margin must be declared beforehand

The TORCH trial

Trial outline

- COPD patients
- Salmeterol and Fluticasone combination versus placebo
- Outcome: hazard ratio (death)

Confidence interval

- Traditional: (0.681,1.002)
- Three-sided testing ($\Delta = 0$): (0.702,1]

Conclusion

New CI rules out harmful effect

The COLOR trial

Trial outline

- Colon cancer patients
- Laparoscopic colectomy versus open surgery
- Outcome: 3-year disease-free survival
- Non-inferiority trial $\Delta = 7\%$

Confidence interval

- Traditional: (-7.2%,3.2%)
- Three-sided testing: [-7%,3.2%)

Conclusion

New CI rules out Δ-inferiority of new treatment

The EVA-S3 trial

Trial outline

- Patients with symptomatic carotid stenosis
- Stenting versus Endarterectomy
- Outcome: stroke or death 30 days after treatment
- Non-inferiority trial $\Delta = 2\%$

Confidence interval

- Traditional: (-10.0%,-1.4%)
- Three-sided testing: (-9.3%,-1.4%)

Conclusion

Qualitatively similar conclusion, but narrower CI

The APOLLO trial

Trial outline

- Patients with type II diabetis
- Insulin Glargine versus Prandial Insulin Lispro
- Outcome: haemoglobin decrease
- Non-inferiority trial $\Delta = 0.4$

Confidence interval

- Traditional: (-0.322,0.008)
- Three-sided testing: (-0.322,0.008)

Conclusion

No change

Outline

Discussion

Three-sided testing

- Increased power of one-sided testing
- Symmetry of two-sided testing

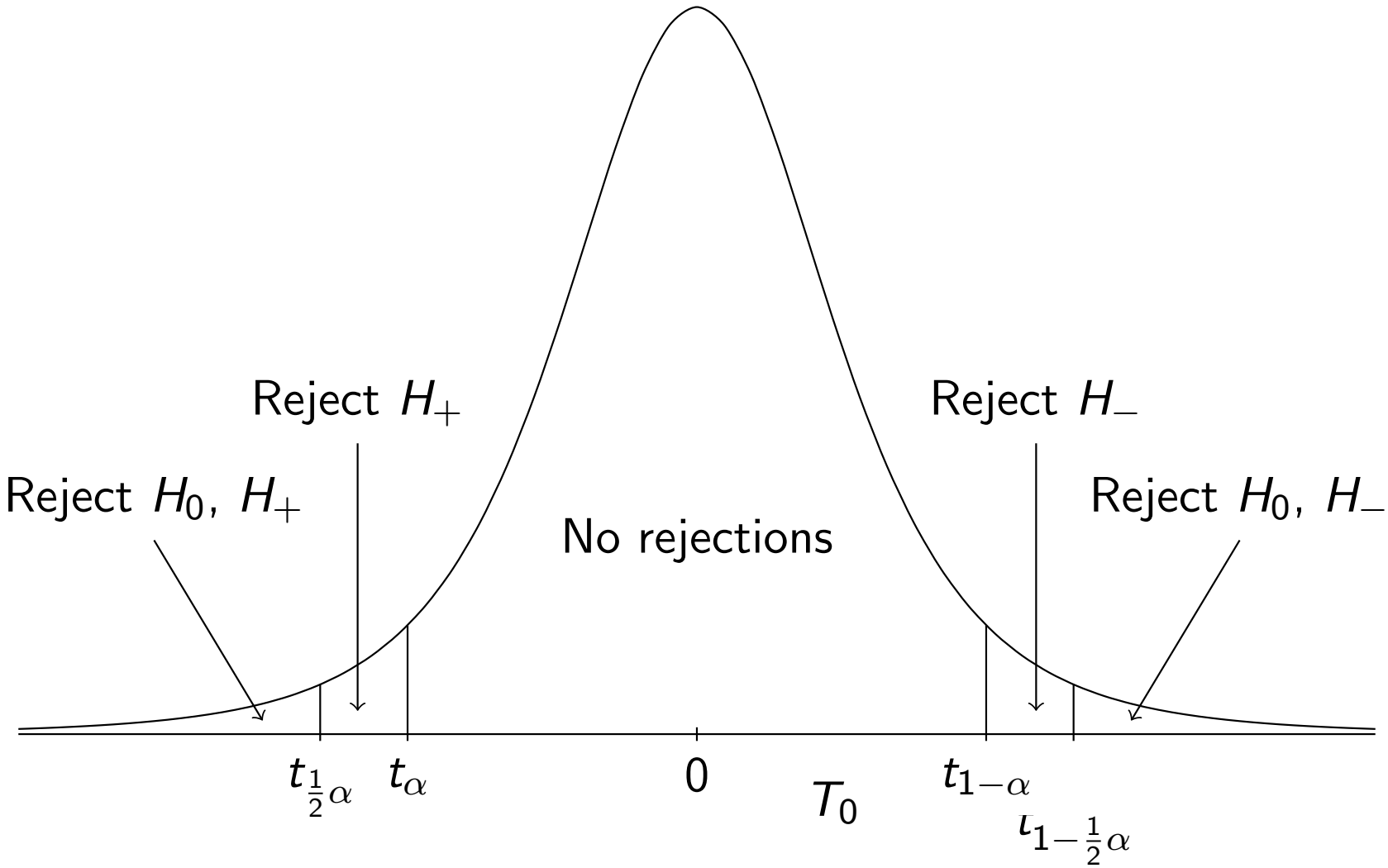
Confidence intervals

- Approach not reconcilable with classical CI
- Alternative CI available (often narrower)

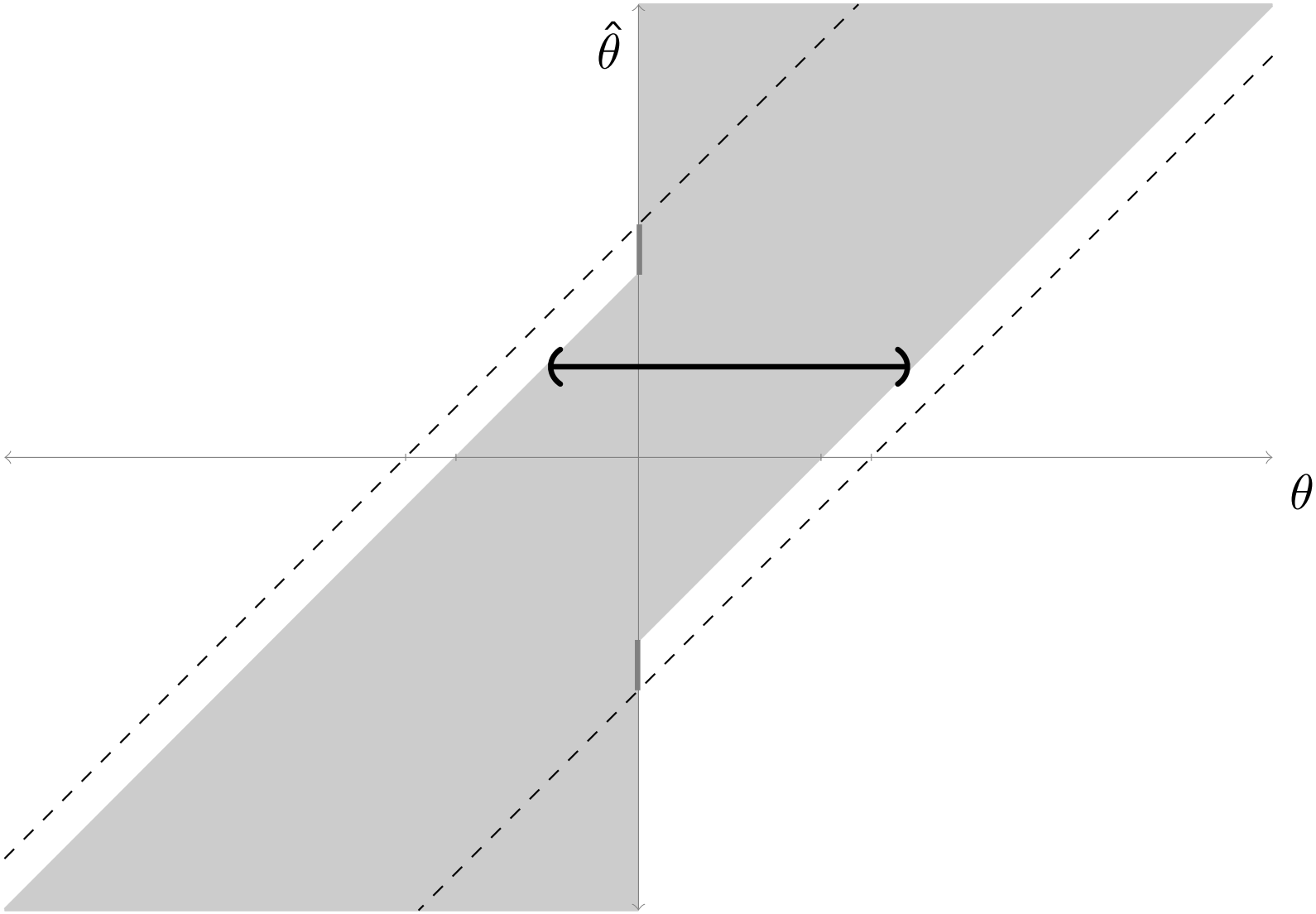
Focussed testing

Uniformly more power than non-focussed procedure

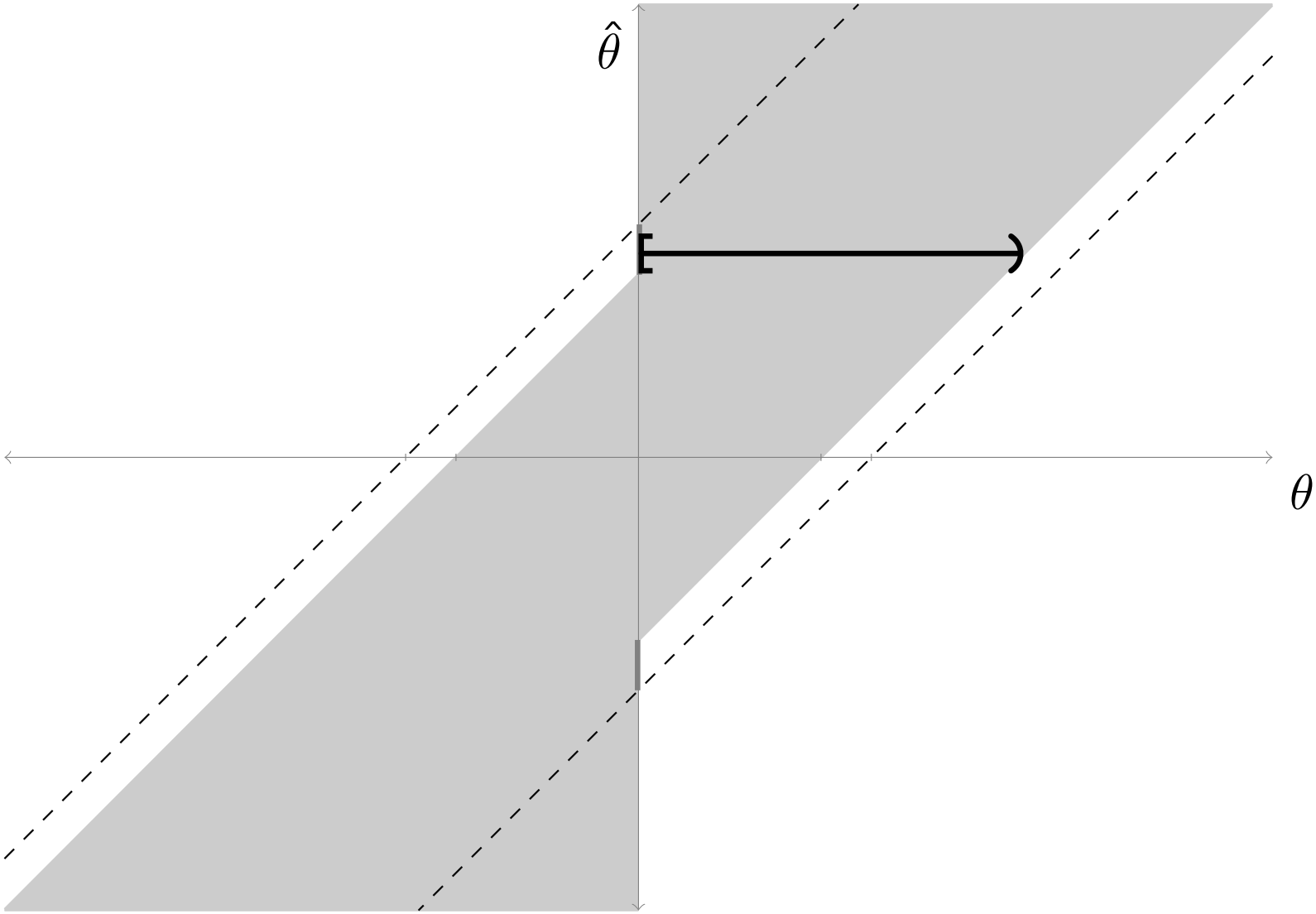
Surprising free inference



Confidence intervals



Confidence intervals



Confidence intervals