Decision making versus inference.

p-values are not the issue.

Kaspar Rufibach, Merck KGaA, Darmstadt, Germany

EORTC, 30th September 2025

2002-2006: Swiss Group for Applied Cancer Research (SAKK).

2002-2012: Uni Bern, Stanford, Zurich.

Since 2013: Roche, Merck KGaA.

Making a clinical decision is a complicated exercise.

It can never be automatized or outsourced.

Even if journals or other stakeholders would like that.

***p*-values are a scientific tool. Banning them is ridiculous.**

Educate people and insist on proper use.

I will sketch decision-making for pharma trials.

Why would decision-making for a collaborative group trial be different?

Hypothesis test

Neyman-Pearson

significant vs. non-significant

Hypothesis test

Scientific question: Primary endpoint score different between Group 1 and Group 2?

Null hypothesis H_0 – statement to be rejected:

$$H_0 \ : \ \mu_{\text{Group 1}} \ = \ \mu_{\text{Group 2}} \quad \Leftrightarrow \quad \delta \ = \ \mu_{\text{Group 1}} \ - \ \mu_{\text{Group 2}} \ = \ 0.$$

Alternative hypothesis H_1 – what researcher is interested in:

$$H_1 \ : \ \delta \ \neq \ 0.$$

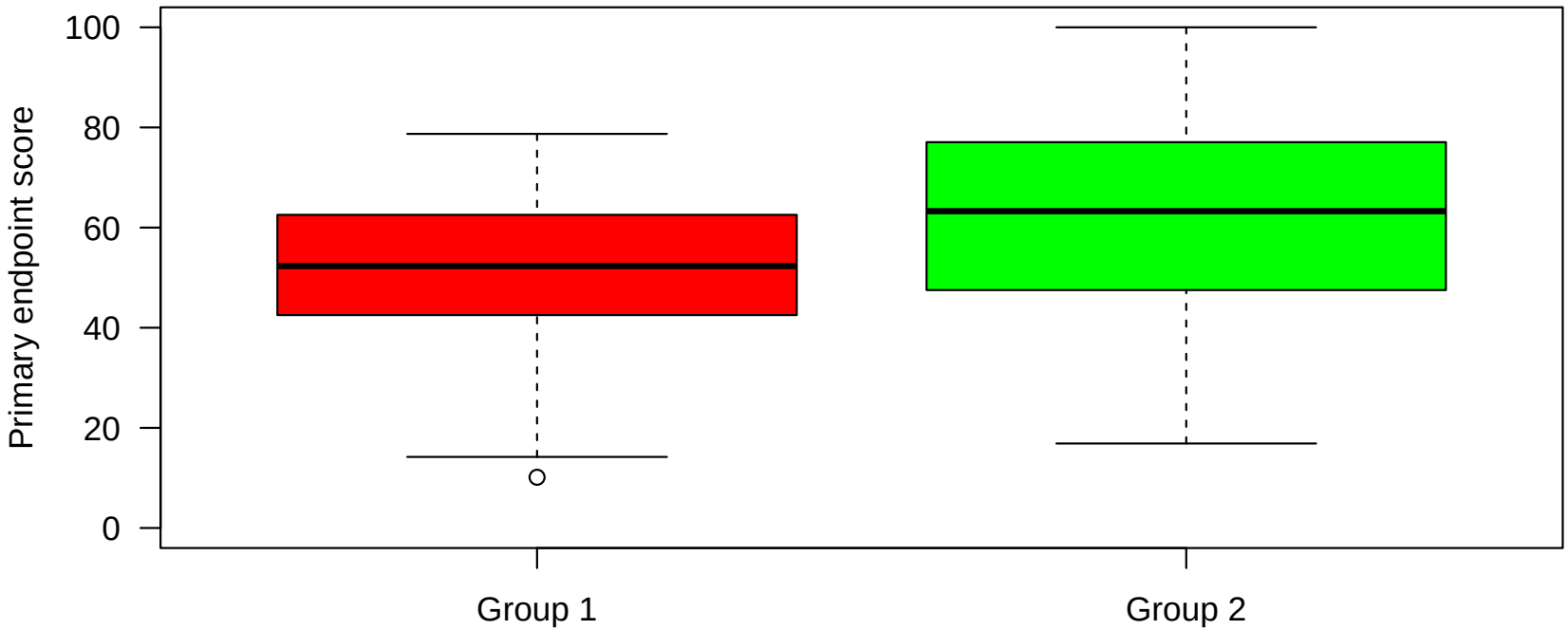
Set significance level α.

Define effect size to be detected with given power.

Compute sample size.

Hypothesis test

Collect data – draw random sample.



Compute **test statistic** $\Rightarrow$ distance between estimated and hypothetical value:

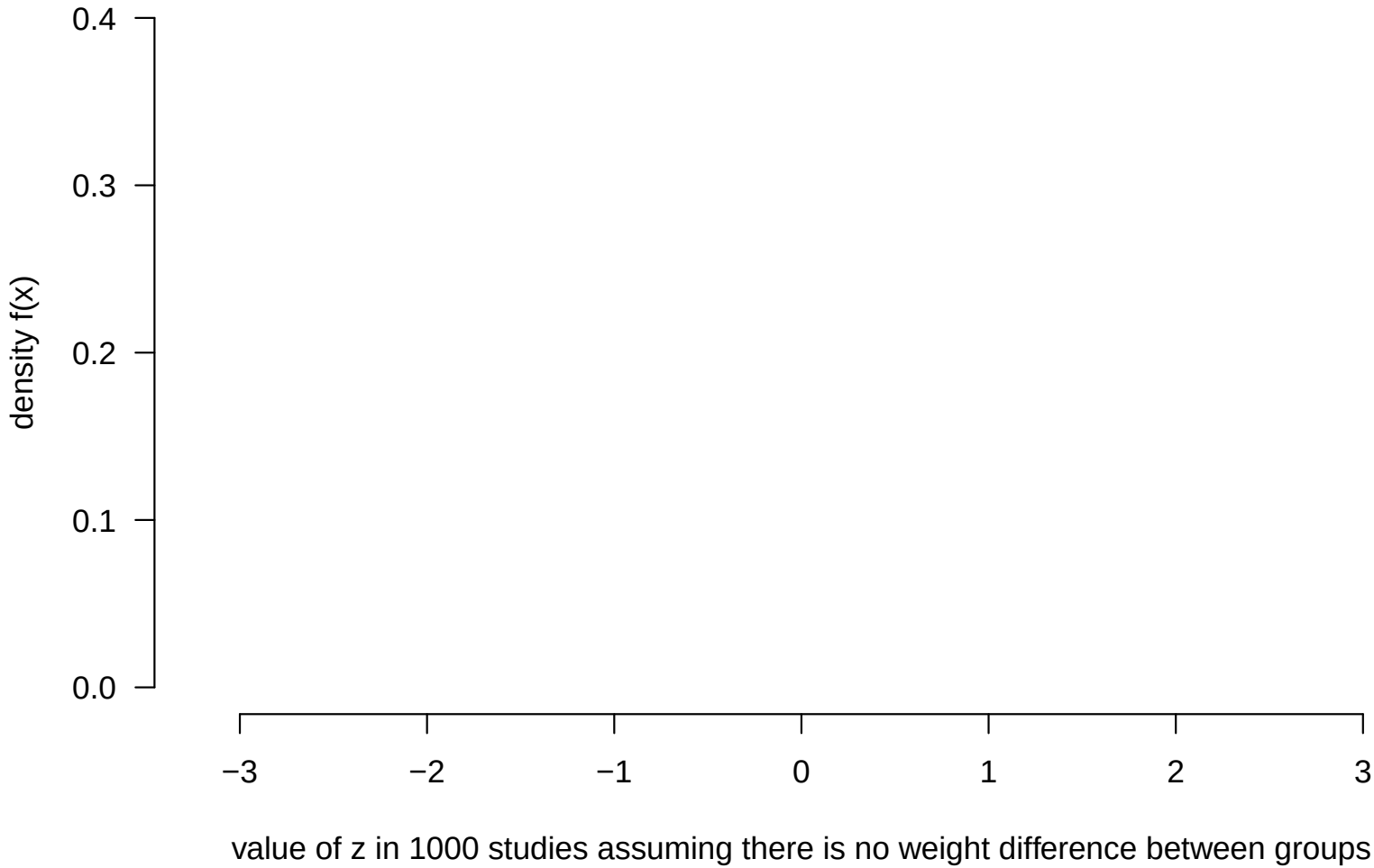
$$z = \frac{\text{estimate} - \text{null value}}{\text{standard error}} = \frac{(51.34 - 63.84) - 0}{4.49} = -2.78.$$

Compare $|z|$ to what would be expected if H_0 were true.

Reject or do not reject H_0.

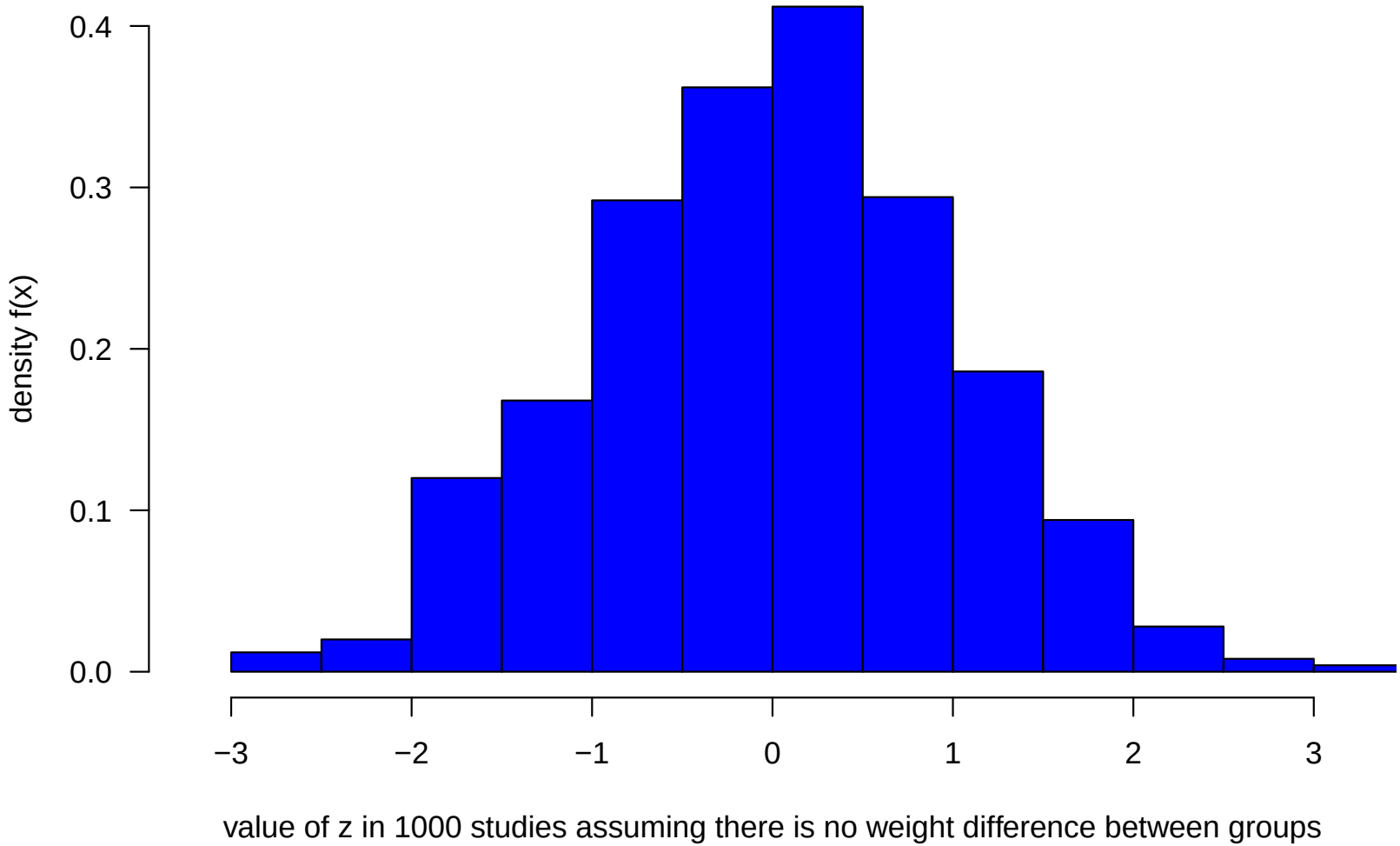
Hypothesis test: example z-Test

How large is $|z|$ to be expected if H_0 holds? Assume we could perform 1000 studies for which H_0 were true.



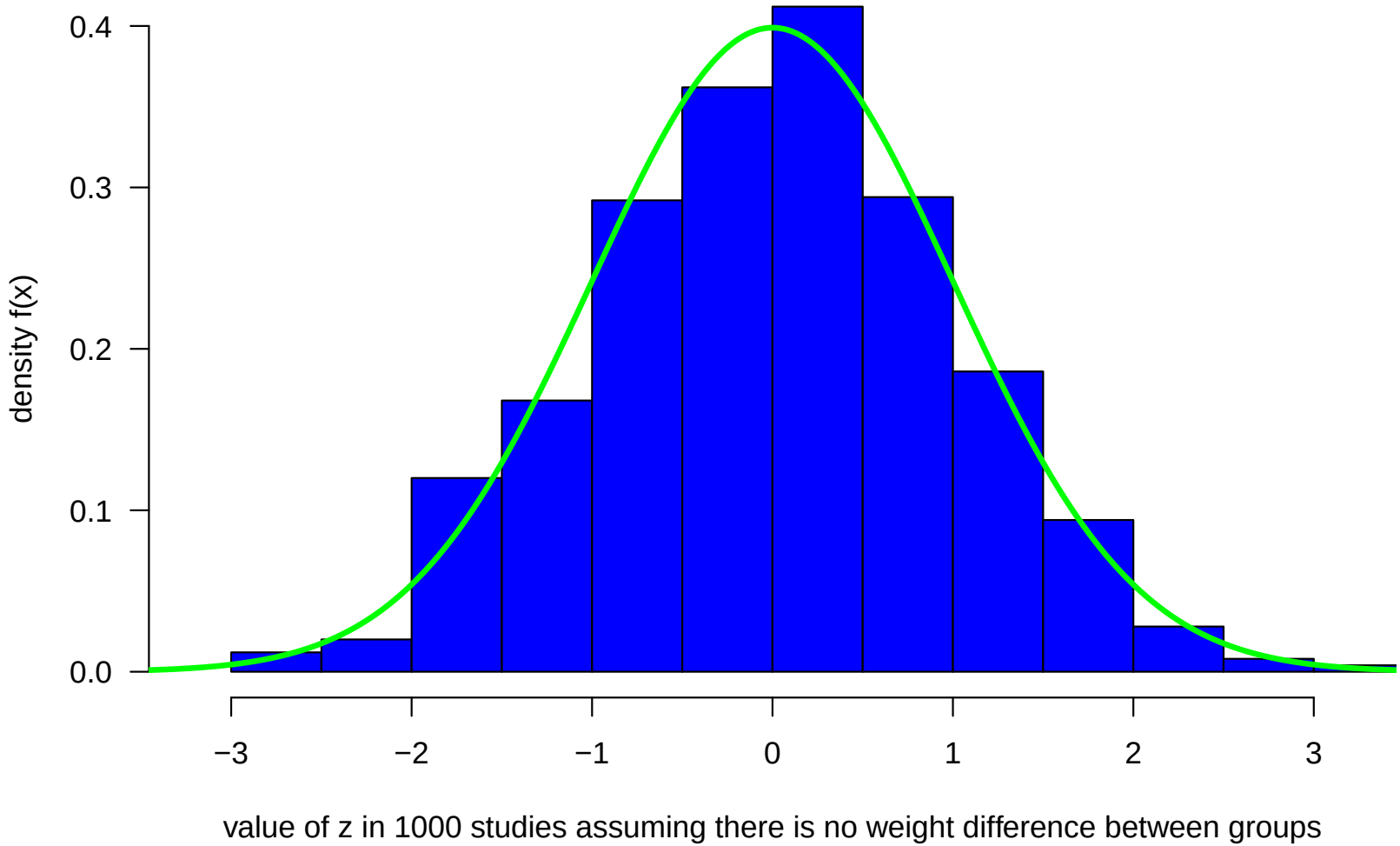
Hypothesis test: example z-Test

How large is $|z|$ to be expected if H_0 holds? Assume we could perform 1000 studies for which H_0 were true.



Hypothesis test: example z-Test

How large is $|z|$ to be expected if H_0 holds? Assume we could perform 1000 studies for which H_0 were true.



We do not need 1000 studies! But **mathematical theory**.

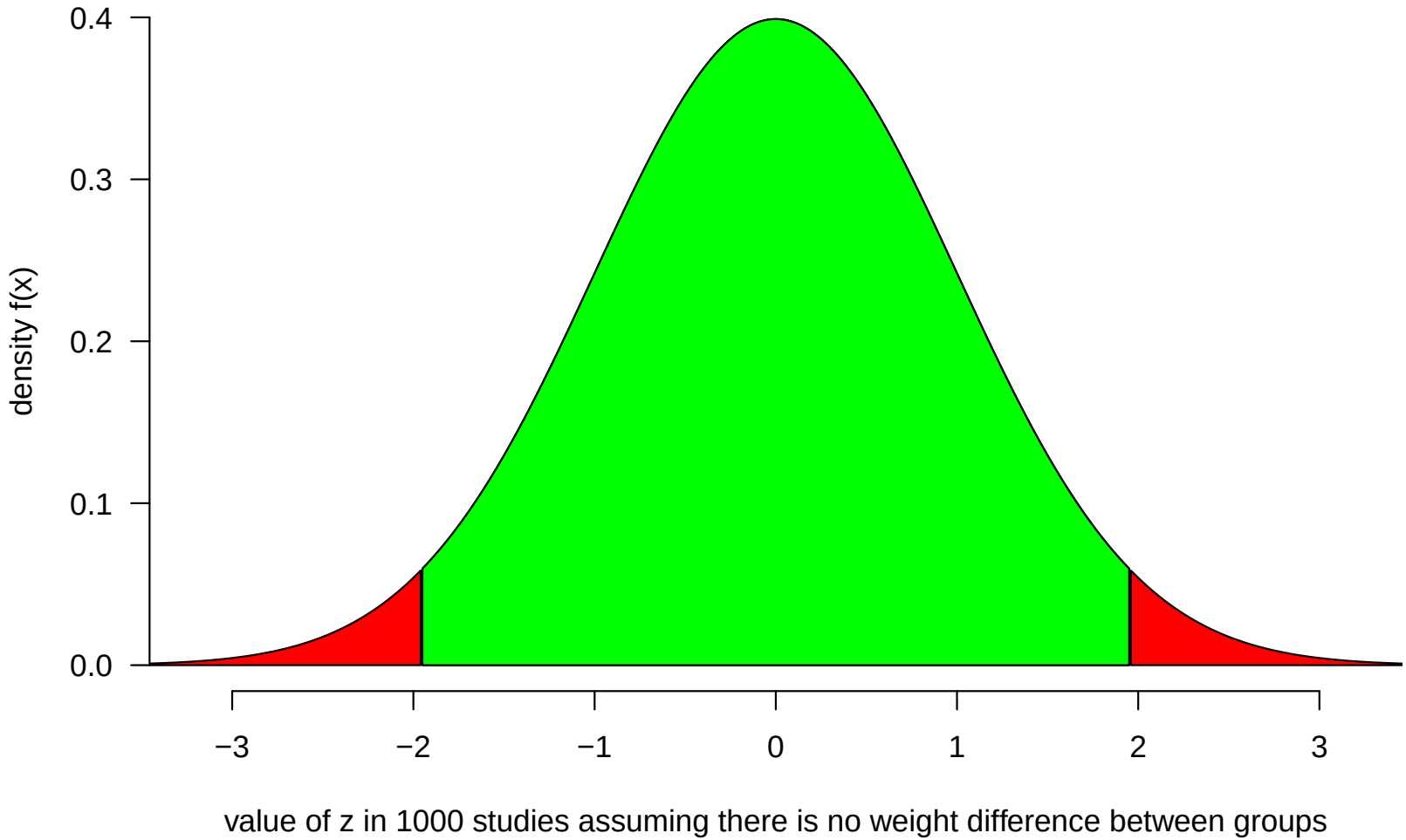
Hypothesis test: example z-Test

How large is $|z|$ to be expected if H_0 holds? Assume we could perform 1000 studies for which H_0 were true.

We do not need 1000 studies! But **mathematical theory**.

Hypothesis test: example z-Test

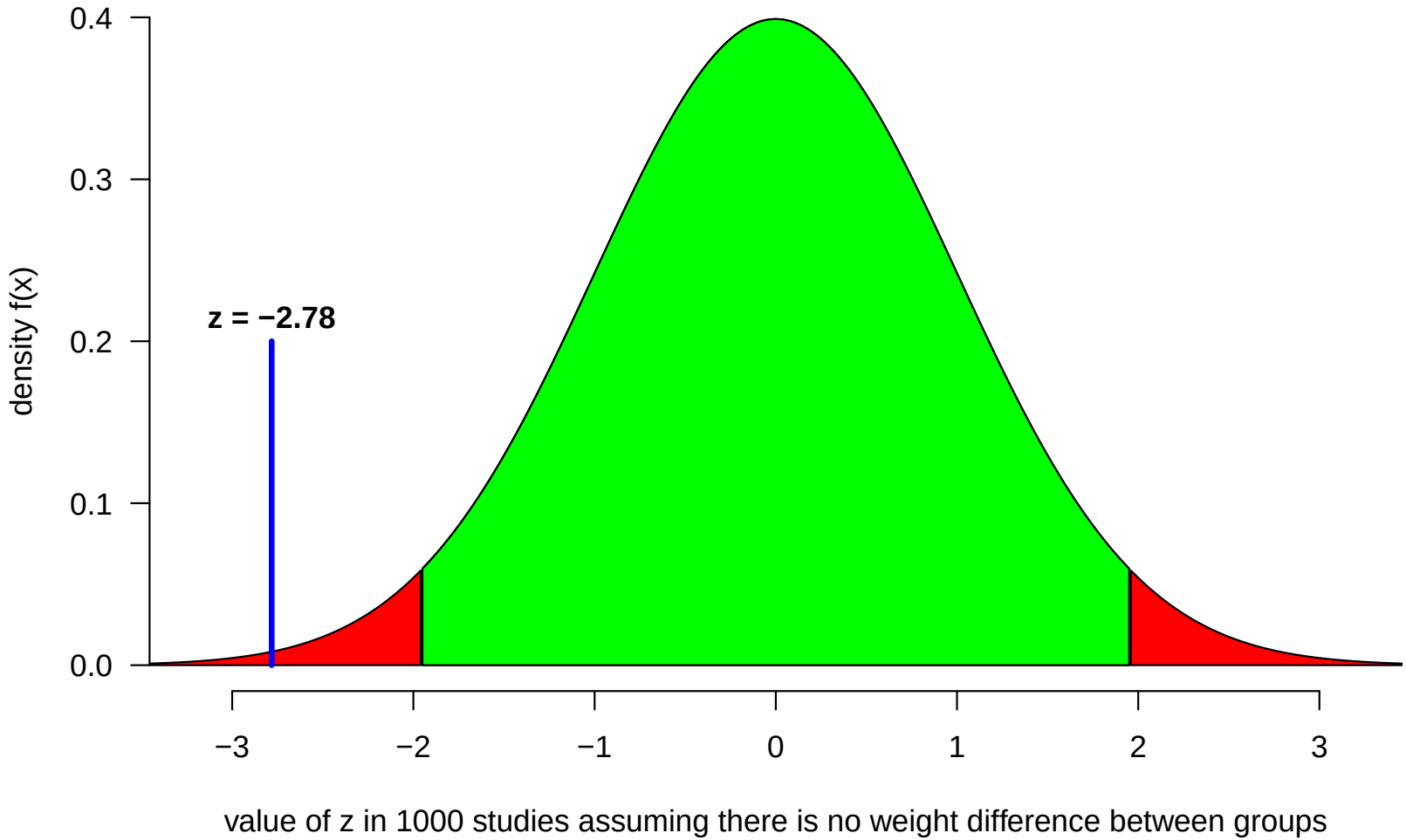
How large is $|z|$ to be expected if H_0 holds? Assume we could perform 1000 studies for which H_0 were true.



We do not need 1000 studies! But **mathematical theory**.

Hypothesis test: example z-Test

How large is $|z|$ to be expected if H_0 holds? Assume we could perform 1000 studies for which H_0 were true.



We do not need 1000 studies! But **mathematical theory**.

Did we need a p-value to make a decision on H_0?

No.

Significance level, power:
operating characteristics.

Pre-specified.

Only maintained if
one sticks to pre-specification.

Significance test

Fisher

***p*-value**

Significance test: p-value

p-value:

- **Quantify evidence** against H_0 with number independent of **test, sample size**.
- Combine **effect estimate** and **uncertainty quantification** into one number.
- Fisher: perform many related experiments, combine p-values to get to final conclusion at some point. **Meta-analysis**.
- **No decision** on null hypothesis! No operating characteristics.

Labels for evidence against null hypothesis:

TABLE 5 A Useful Language for Interpreting p Values

$p < 0.001$	Overwhelming evidence
$0.001 \leq p < 0.01$	Strong evidence
$0.01 \leq p < 0.05$	Some evidence
$0.05 \leq p < 0.10$	Insufficient evidence
$p \geq 0.10$	No evidence

Pocock *et al.* (2015).

Statistical significance: **binary decision.**

Highly significant, *trend towards significance*, **...:**
MEANINGLESS!

Inference vs. decision?

Do we want / need binary decision?

Interest in probability of wrong decision?

Hypothesis test in drug development:

Rejection of H_0 in properly designed trial:
entry ticket for negotations with regulator.

Negotations: effect size, other endpoints, etc.

No automatism!

Experience shows:
leads to reasonably sized trials.

Making a clinical decision is a complicated exercise.

It can never be automatized or outsourced.

Even if journals or other stakeholders would like that.

A p-value is no substitute for a brain.

Stone and Pocock (2010)

In many cases published medical literature requires no firm decision: it contributes incrementally to an existing body of knowledge.

Sterne and Smith (2001)

The reporting of scientific results is not about making decisions, but about collecting, summarizing, and reevaluating evidence.

Blume and Peipert (2003)

So, why are so many people confused?

***p*-value can be used to make decision in a hypothesis test.**

Conceptually, the two frameworks are independent and have different goals.

Making a clinical decision is a complicated exercise.

It can never be automatized or outsourced.

Even if journals or other stakeholders would like that.

***p*-values are a scientific tool. Banning them is ridiculous.**

Educate people and insist on proper use.

References

Blume, J. and Peipert, J. F. (2003). What your statistician never told you about P-values. *J Am Assoc Gynecol Laparosc*, **10**, 439–444.

Pocock, S. J., McMurray, J. J., and Collier, T. J. (2015). Making Sense of Statistics in Clinical Trial Reports: Part 1 of a 4-Part Series on Statistics for Clinical Trials. *J. Am. Coll. Cardiol.*, **66**(22), 2536–2549.

Sterne, J. A. and Smith, G. D. (2001). Sifting the evidence – what's wrong with significance tests? *British Medical Journal*, **322**(7280), 226–231.

Stone, G. W. and Pocock, S. J. (2010). Randomized trials, statistics, and clinical inference. *J. Am. Coll. Cardiol.*, **55**(5), 428–431.

Thank you for your attention.

kaspar.rufibach@merckgroup.com

Slides can be downloaded on

www.kasparrufibach.ch

Backup

Inferential concepts

Feature	Neyman-Pearson	Fisher p-value	Bayes
Specifies H_0	✓	✓	✓
Specifies H_1	✓	✗	✓
Binary decision	✓	✗	$\sim$
Operating characteristics	✓	✗	✓

R version and packages used to generate these slides:

R version: R version 4.4.3 (2025-02-28 ucrt)

Base packages: stats / graphics / grDevices / utils / datasets / methods / base

Other packages: reporttools / xtable

This document was generated on 2025-09-25 at 21:32:42.