UNIVERSITÄT BERN

The p-value – what it is (not)!

Sven Trelle
CTU Bern

FOREWORD

Empiricism

- Ideas formed from observations/empirical evidence
- Induction (↔ deduction)
  - Medical
    - Observed signs & symptoms → (hypothesized) diagnosis
  - Statistical
    - Observed data → (hypothesized) parameter

Personal view on the aims of research

- Natural science classical pradigm → identify deterministic natural laws
  — Example: Newton's law of universal gravitation $F = mg$ (g=9.81 m/s^2)
- "Modern" physics …

- Patient-oriented clinical research
  — Classical paradigm: which treatment works (better)
  — Nowadays (?): How good is this treatment

Current practice

- " … CONCLUSIONS AND RELEVANCE:
  In the final analysis of this randomized clinical trial of patients with glioblastoma who had received standard radiochemotherapy, the addition of TTFields to maintenance temozolomide chemotherapy vs maintenance temozolomide alone, resulted in statistically significant improvement in progression-free survival and overall survival. …"
- Any issues?
  - Where is the scientific conclusion?
  - Observed data and statistical results replace scientific thinking?
  - Context/other evidence, relevance, mechanism, …?

Stupp R et al. 2018

DEFINITION OF THE P-VALUE

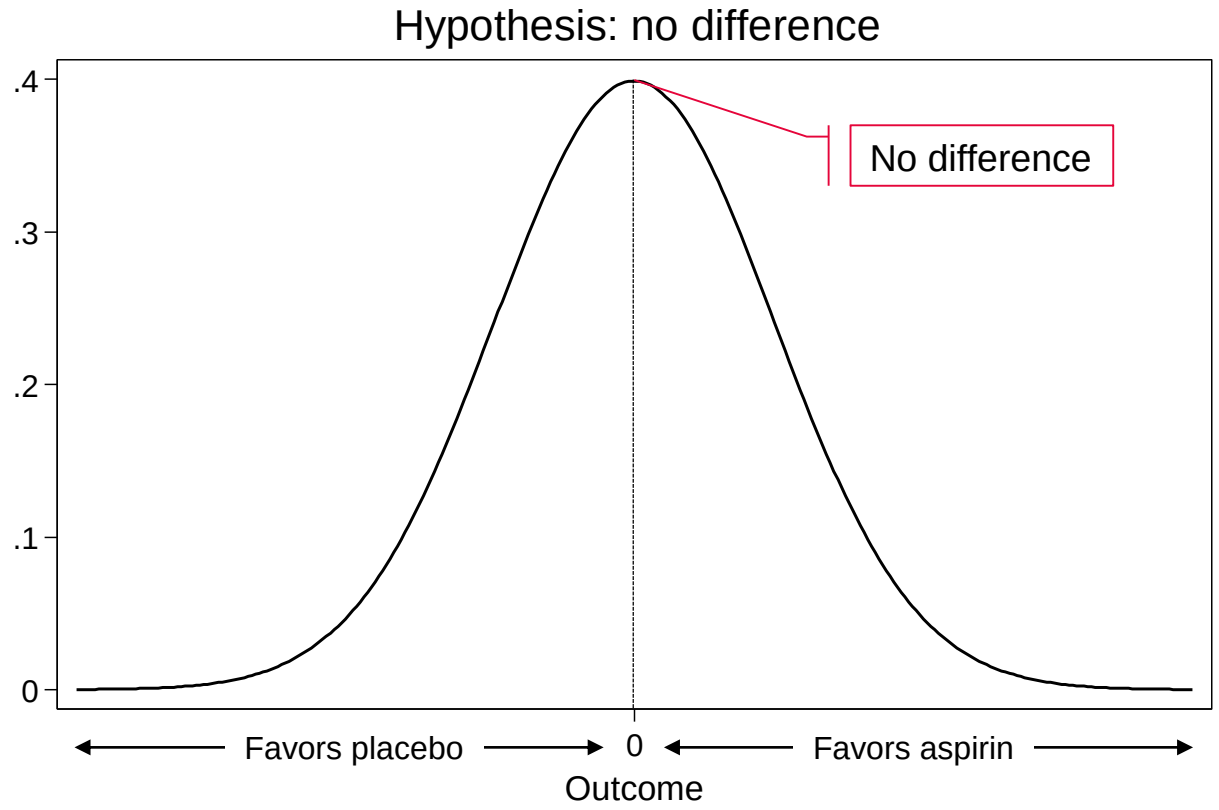
Probability of every possible outcome

- Difference between two interventions e.g. Aspirin and Placebo
- Because we cannot observe the effect in everybody, everytime
  → Sample (trial) → estimate the effect
  — Sample → population
  — Empiricism!

→ Uncertainty

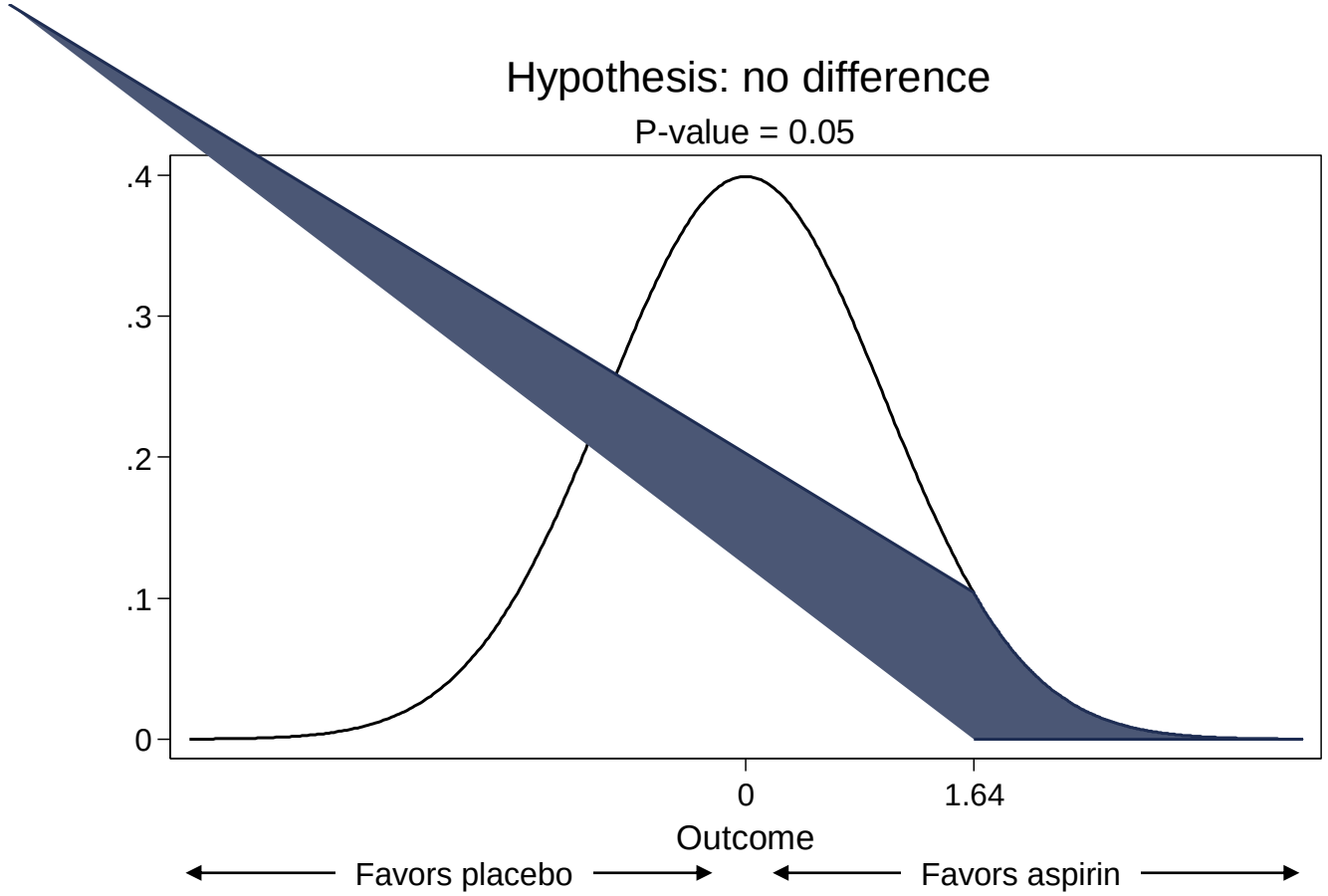
Probability of every possible outcome

- Difference between two interventions e.g. Aspirin and Placebo



The p-value

- Aspirin better than placebo by 1.64



Definition

- After we have observed data (results)

  The P value is defined as the probability,
  under the assumption of no effect or no
  difference (*the null hypothesis*),
  of obtaining a result equal to or more
  extreme than what was actually observed

- Originally proposed by RA Fisher to support scientific conclusion drawing, not as inferential method

Goodman S 1999

Conditional probabilities

- Probabilities are always conditional
  - with respect to available information
  - with respect to context and/or generally accepted assumptions (physics, mathematical truths, etc..)
- Notation P(A|B): probability of A given (conditional on) B
- $P(A|B) \neq P(B|A)$
  - Probability of fever given influenza is very high
  - Probability of influenza given fever is certainly lower if not low
- Bayes theorem: $P(B|A) = \frac{P(A|B) \times P(B)}{P(A)}$

ELSEVIER

Seminars in
HEMATOLOGY

A Dirty Dozen: Twelve *P*-Value Misconceptions

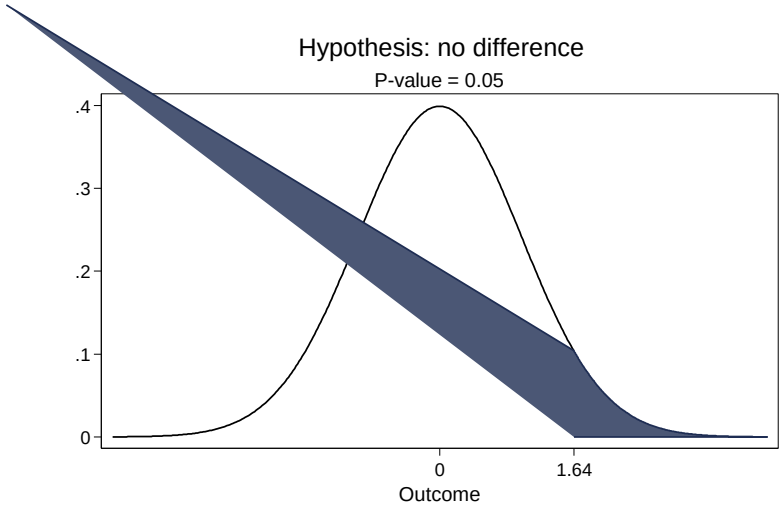
Steven Goodman

The *P* value is a measure of statistical evidence that appears in virtually all medical research papers. Its interpretation is made extraordinarily difficult because it is not part of any formal system of statistical inference. As a result, the *P* value's inferential meaning is widely and often wildly misconstrued, a fact that has been pointed out in innumerable papers and books appearing since at least the 1940s. This commentary reviews a dozen of these common misinterpretations and explains why each is wrong. It also reviews the possible consequences of these improper understandings or representations of its meaning. Finally, it contrasts the *P* value with its Bayesian counterpart, the Bayes' factor, which has virtually all of the desirable properties of an evidential measure that the *P* value lacks, most notably interpretability. The most serious consequence of this array of *P*-value misconceptions is the false belief that the probability of a conclusion being in error can be calculated from the data in a single experiment without reference to external evidence or the plausibility of the underlying mechanism.
Semin Hematol 45:135-140

Misconception 1 (main)

- *P(results|null hypothesis) = P(null hypothesis is true/false|results)*
- *If P=0.05, the null hypothesis has only a 5% chance of being true.*

→ Remember: p-value calculated under assumption that null hypothesis true



Posterior probabilities (reproducibility)

- Remember Bayes theorem i.e. P(A) and P(B) are needed

Table 3 Proportion of false positive significant results with three different criteria for significance

Power of study (proportion (%) of time we reject null hypothesis if it is false)	Percentage of "significant" results that are false positives		
	P=0.05	P=0.01	P=0.001
80% of ideas correct (null hypothesis false)			
20	5.9	1.2	0.10
50	2.4	0.5	0.05
80	1.5	0.3	0.03
50% of ideas correct (null hypothesis false)			
20	20.0	4.8	0.50
50	9.1	2.0	0.20
80	5.9	1.2	0.10
10% of ideas correct (null hypothesis false)			
20	69.2	31.0	4.30
50	47.4*	15.3	1.80
80	36.0	10.1	1.10
1% of ideas correct (null hypothesis false)			
20	96.1	83.2	33.10
50	90.8	66.4	16.50
80	86.1	55.3	11.00

Sterne J & Davey Smith G 2001

Table 3 Proportion of false positive significant results with three different criteria for significance

Power of study (proportion (%) of time we reject null hypothesis if it is false)	Percentage of "significant" results that are false positives		
	P=0.05	P=0.01	P=0.001
80% of ideas correct (null hypothesis false)			
20	5.9	1.2	0.10
50	2.4	0.5	0.05
80	1.5	0.3	0.03
50% of ideas correct (null hypothesis false)			
20	20.0	4.8	0.50
50	9.1	2.0	0.20
80	5.9	1.2	0.10
10% of ideas correct (null hypothesis false)			
20	69.2	31.0	4.30
50	47.4*	15.3	1.80
80	36.0	10.1	1.10
1% of ideas correct (null hypothesis false)			
20	96.1	83.2	33.10
50	90.8	66.4	16.50
80	86.1	55.3	11.00

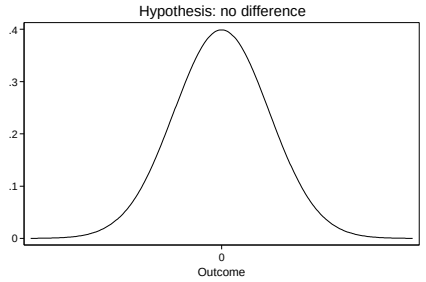
Misconception "1.5"

- > *P-value = hypothesis testing*
- → P-value: probability
- → Hypothesis testing: decision theory to choose between statistical hypothesis (Neyman-Pearson)
  - — Right or wrong decisions
  - — Two types of error (I & II) → misconception 9

Non-significance ≠ no difference (misconception 2)

- "… Median PFS was 6.2 versus 2.8 months (hazard ratio, 1.36; 95% CI, 0.91 to 2.05; P = .11) … There was no apparent difference between the treatment arms …"
- No difference → hazard ratio == 1.0

→ The best estimate supported by the data is always the observed estimate
  - Example: 1.36 not 1.0!



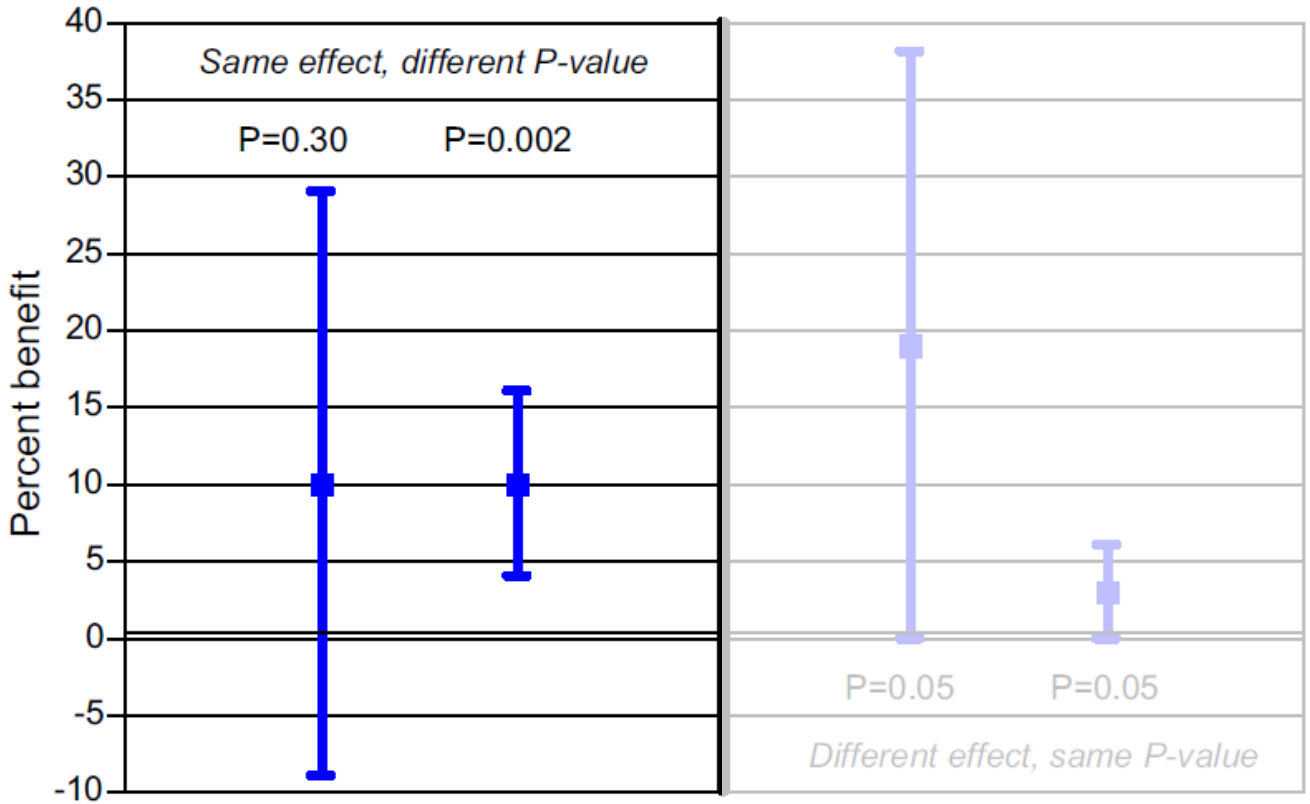
Statistical significance = clinical relevance (misconception 3)

- Clinical relevance determined by
  - Outcome
  - Size of difference

Misconception 4

> *Conflicting p-values = conflicting results*

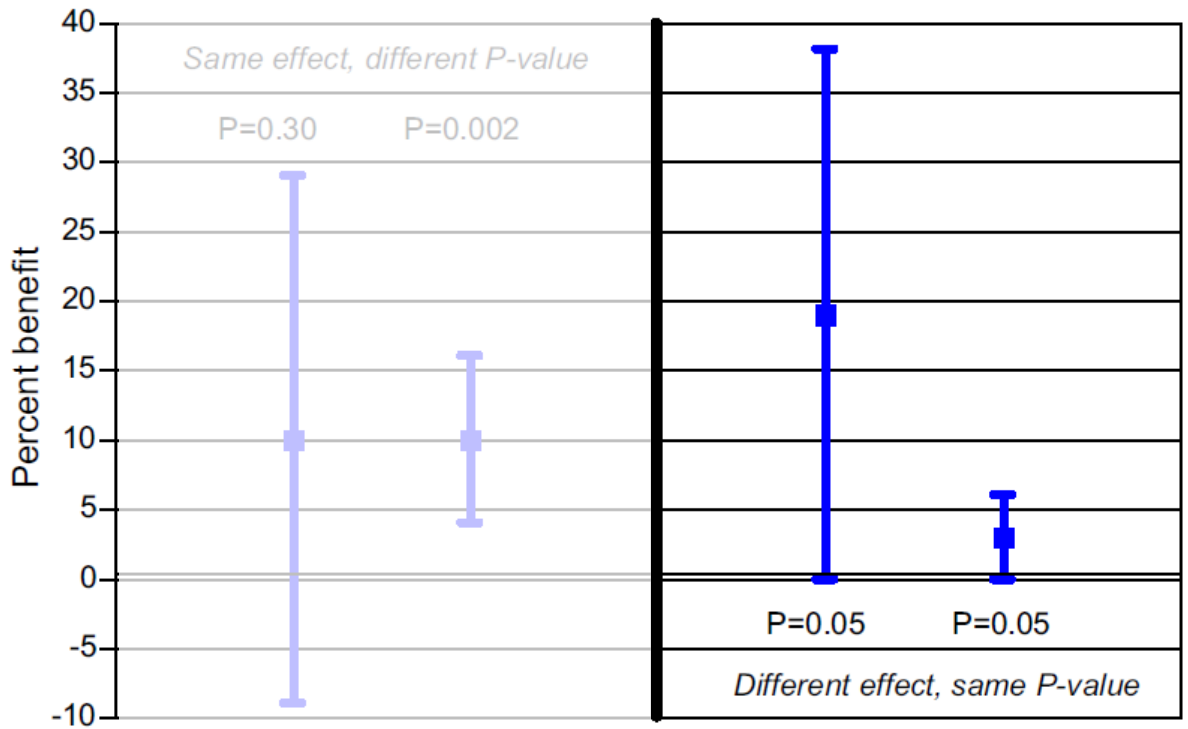
→



P=0.05 == *P=0.05* (misconception 5)

> *Studies with the same P value provide the same evidence against the null hypothesis*

→



Misconception 9

- *P-value == probability for type I error*
- *P=0.05 means that if you reject the null hypothesis, the probability of a type I error is only 5% (after you obtained data/results)*
  - Type I (α) error
    - P(reject null hypothesis|null hypothesis true)
      - E.g. probability to decide intervention is effective whereas in reality it is not

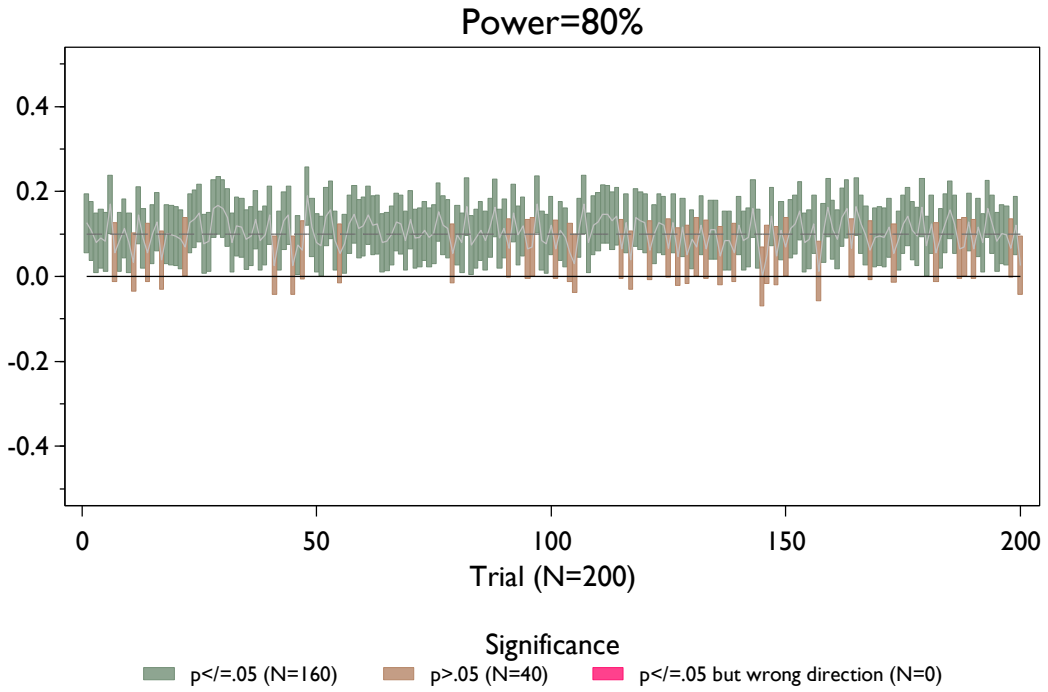
→ See misconception 1 (P(null hypothesis true|data/results))

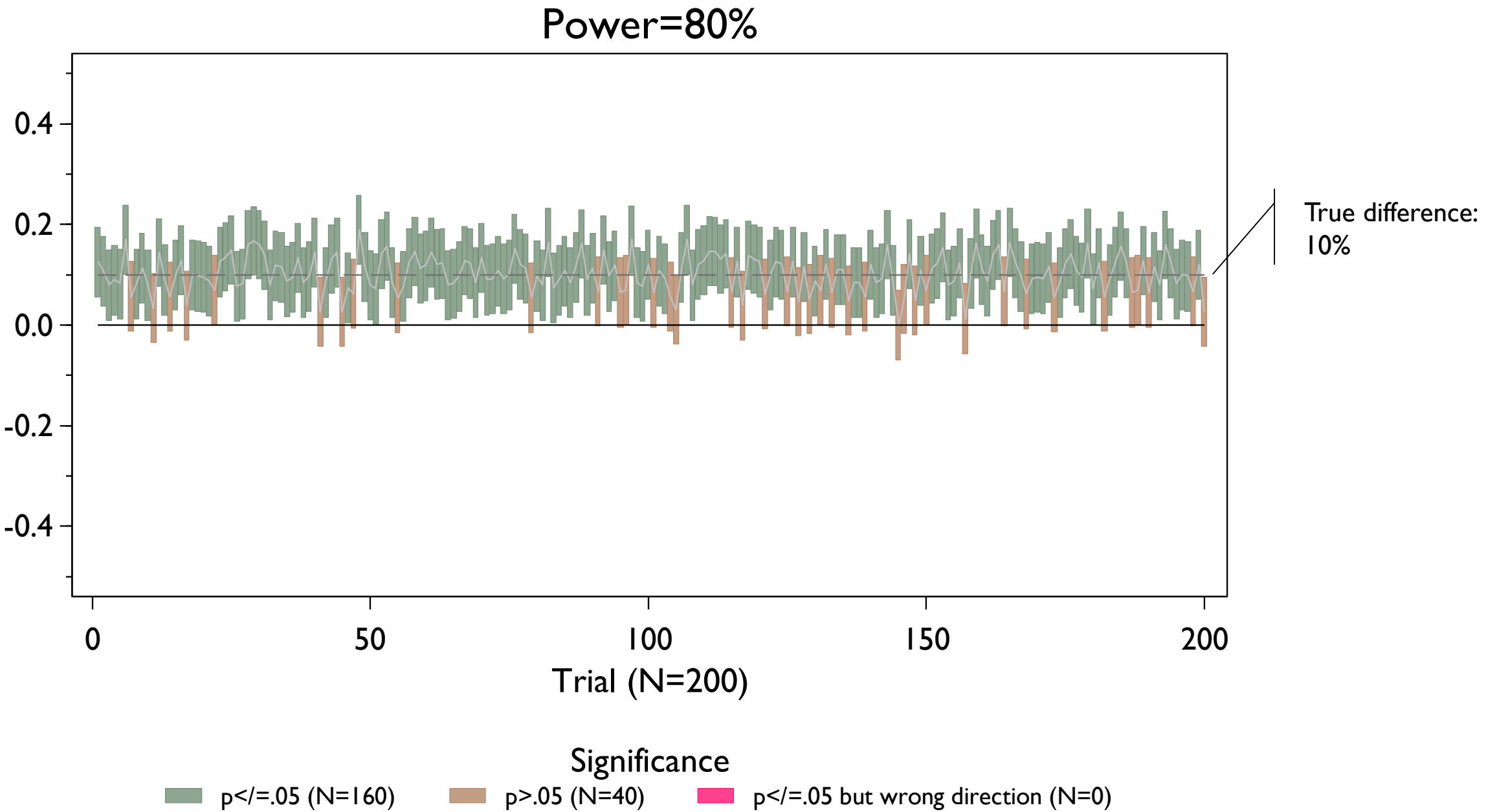
Side note: power

- P-value related to hypothesis testing
- Type I (α) error
  - P(reject null hypothesis|null hypothesis true)
    - E.g. probability to decide intervention is effective whereas in reality it is not
  - α-level → before data acquired
- Type II (β) error
  - P(do not reject null hypothesis|alternative hypothesis true)
    - E.g. probability to decide intervention is not effective whereas in reality it is
  - β-level → before data acquired
- Power = 1 - β
  - P(reject null hypothesis|alternative hypothesis true)

Power by simulation

- Power: 80%; placebo response rate: 40%; effect: 10% improvement




Power=80%
0.4
0.2
0.0
-0.2
-0.4
0
50
100
150
200
Trial (N=200)
True difference:
10%
Significance
p</=.05 (N=160)
p>.05 (N=40)
p</=.05 but wrong direction (N=0)

Misconception 10

- Setting signficance level at p=0.05 will ensure that the chance for a type I error is 5% (before you obtained data)
- → Type I error depends on prior probability of the null hypothesis being true
  - Null hypothesis true → type I error == 5%
  - Null hypothesis false → type I error == 0%
  - Unsure about null hypothesis → type I error == 0-5%

A look on interpretation of p-values from a slightly different angle

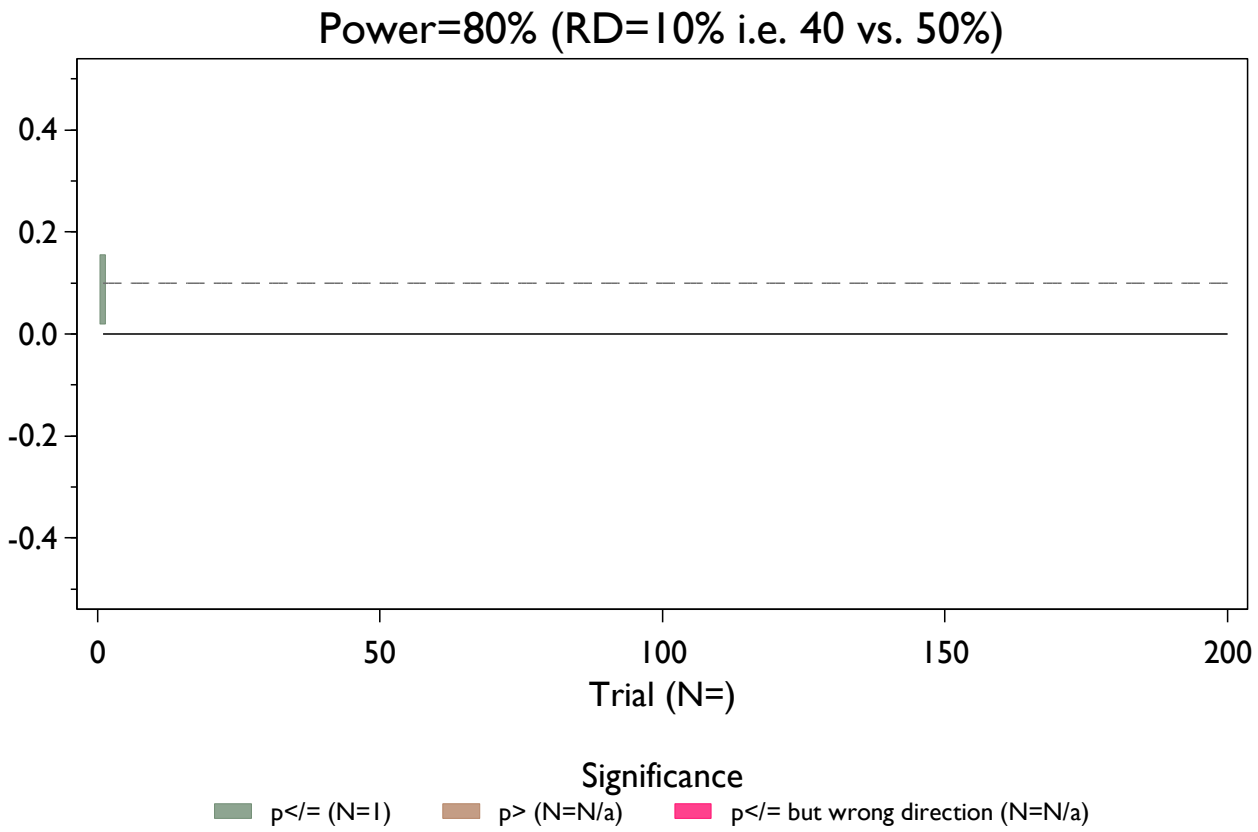
- Randomised-controlled trial comparing placebo with aspirin
- Response rate primary outcome
- A trial with 776 patients (2 x 388)
- Power 80% to detect improvement 40 → 50%

The truth

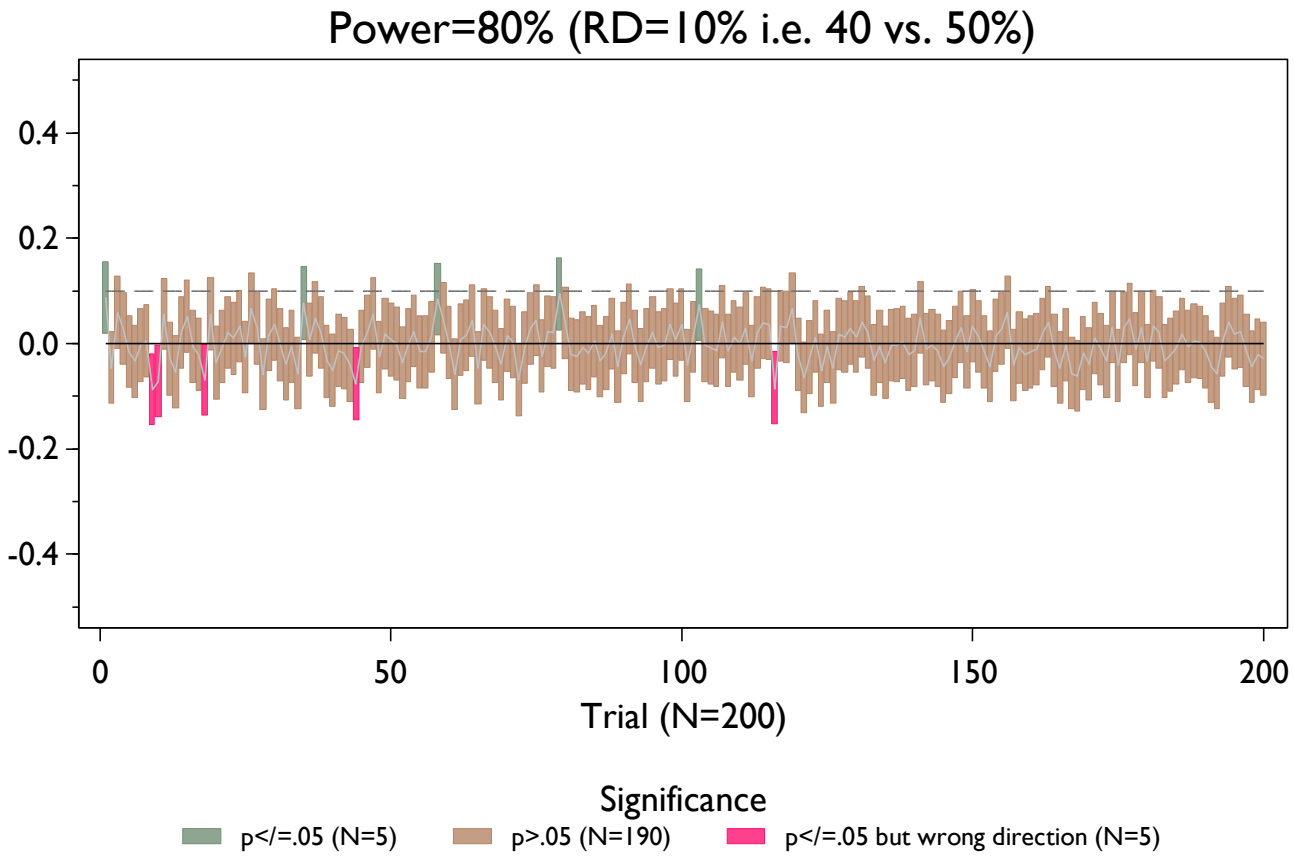
- Interest: difference in response rates placebo-aspirin
- Three possible truths
  - Aspirin > placebo (aspirin better)
    - → Alternative hypothesis
  - Placebo > aspirin (placebo better)
  - Placebo = aspirin (no difference)
    - – With given precision and error of outcome assessment
    - → Null hypothesis

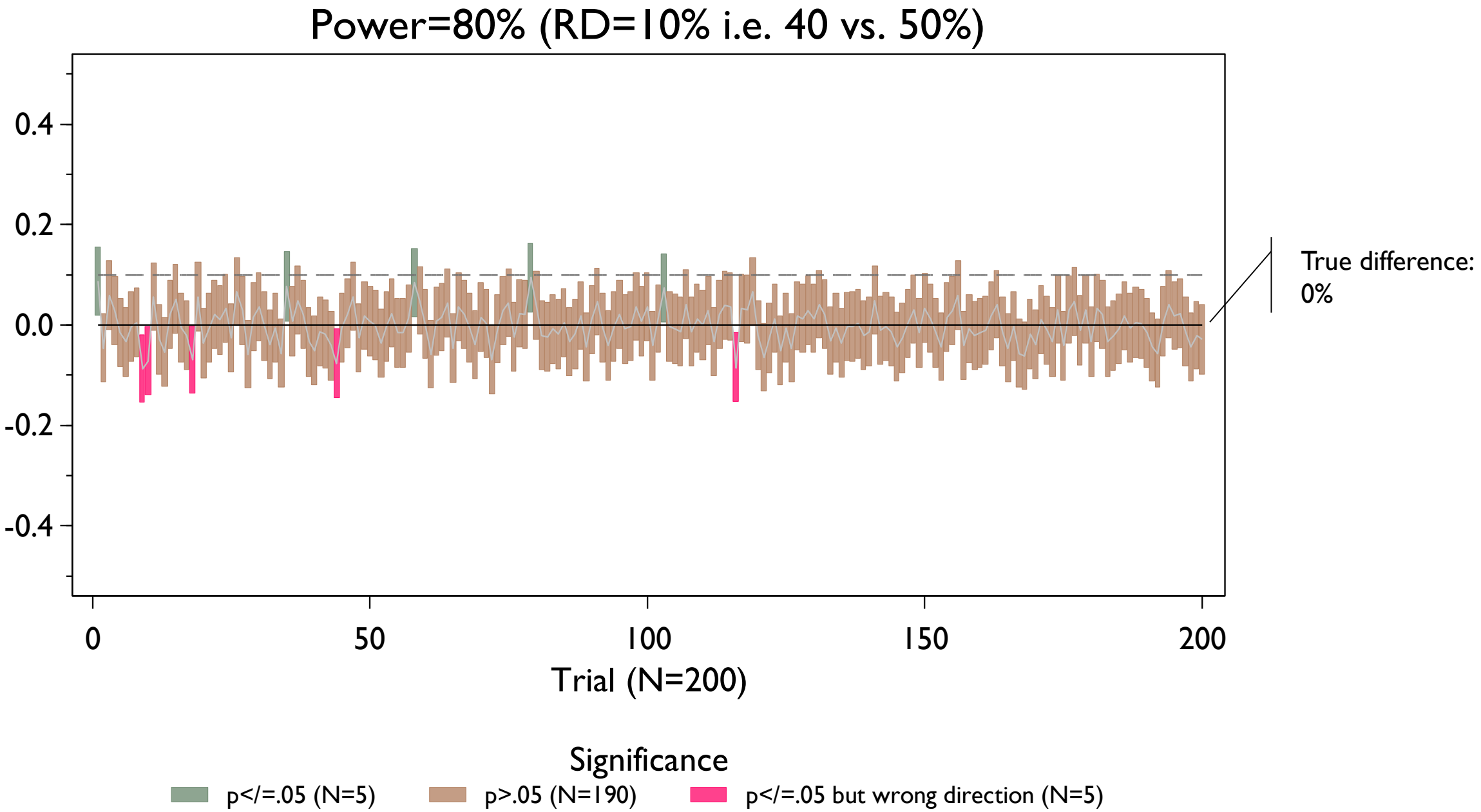
Results

- Response rates
  - Placebo: 34%
  - Aspirin: 43%
- Difference: 9% (95%-CI 2 to 15%)
- P-value: 0.012
  - → Probability that null hypothesis true is 0.012? (WRONG; misconception 1)

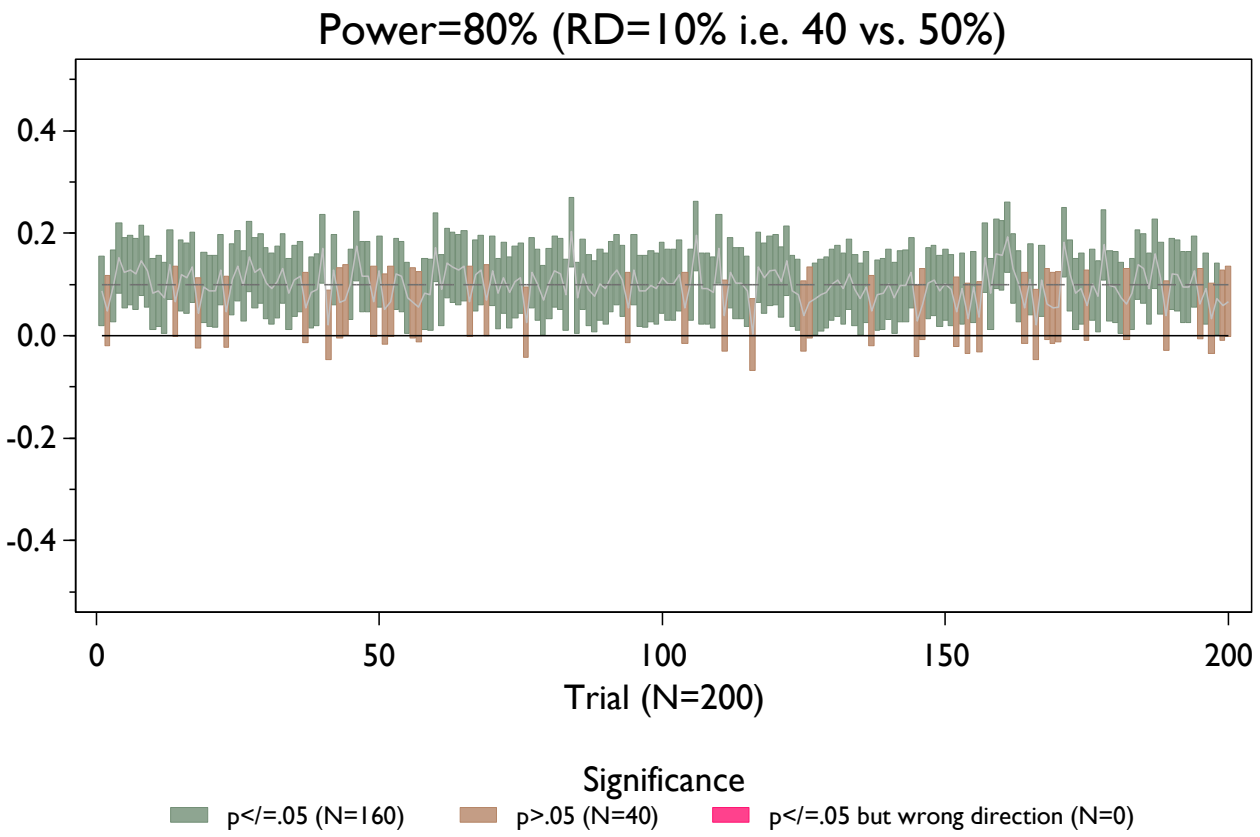


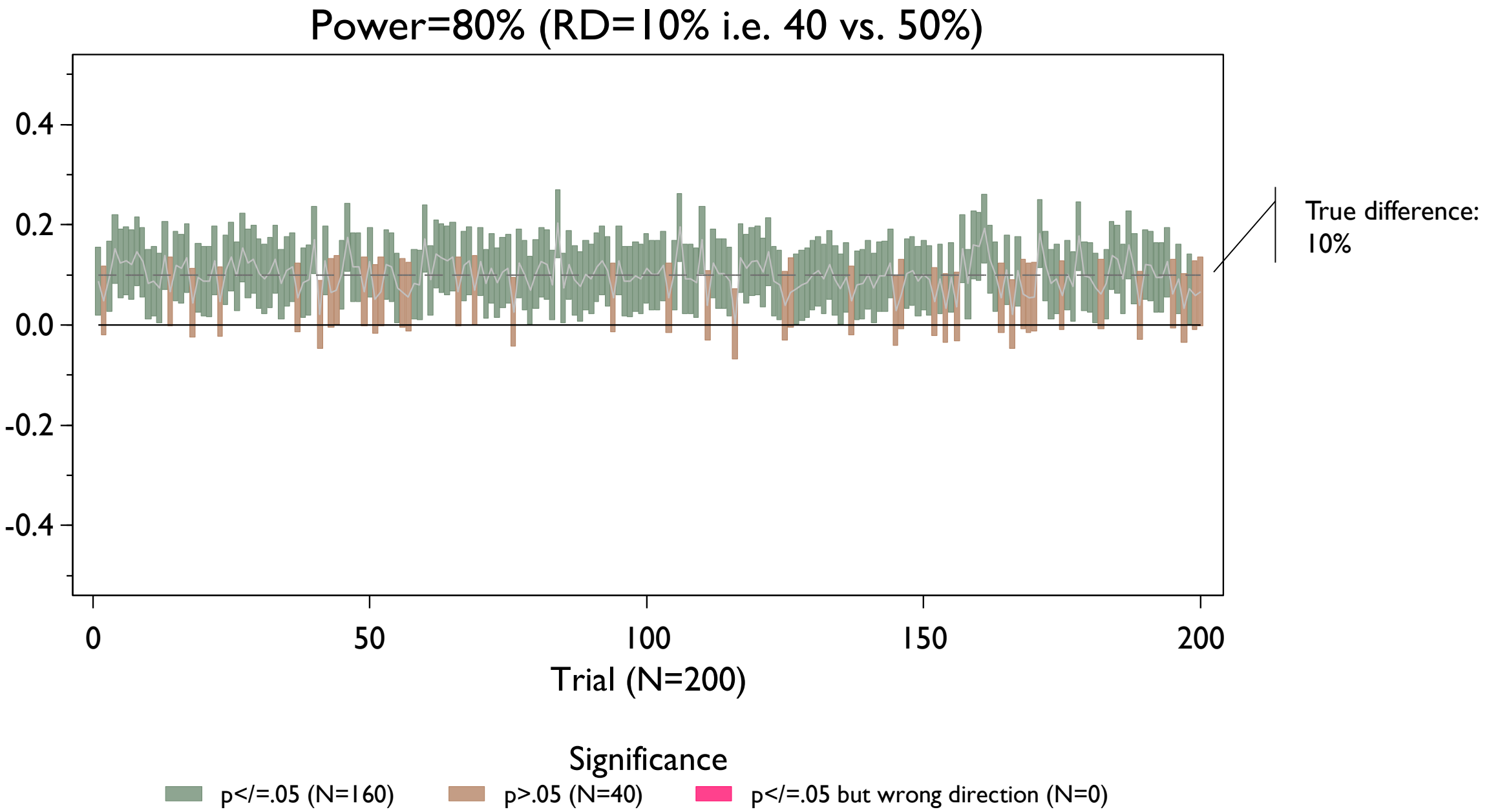
Trials under null hypothesis (no difference)

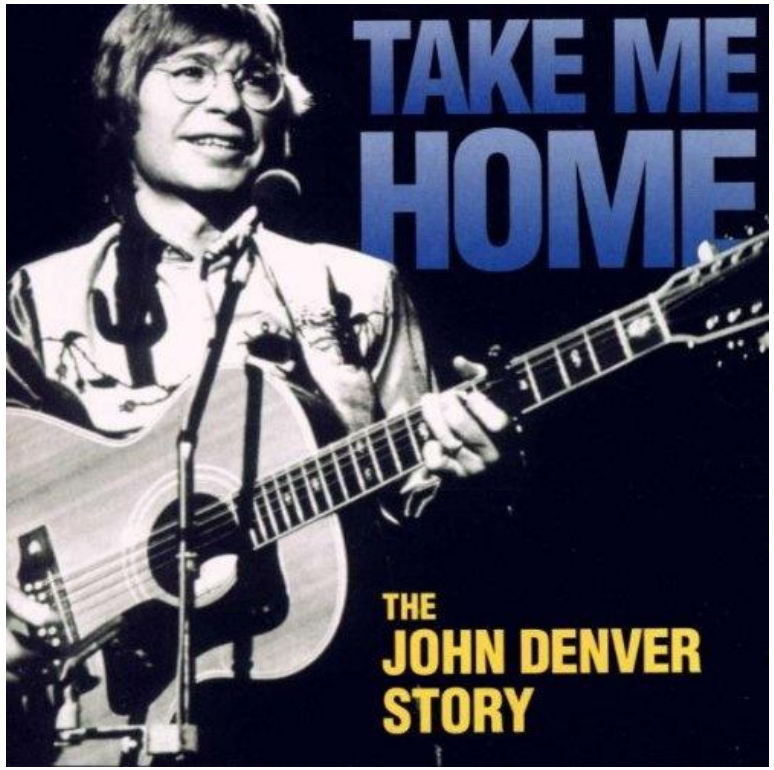



Power=80% (RD=10% i.e. 40 vs. 50%)
0.4
0.2
0.0
-0.2
-0.4
0
50
100
150
200
Trial (N=200)
True difference:
0%
Significance
p</=.05 (N=5)
p>.05 (N=190)
p</=.05 but wrong direction (N=5)

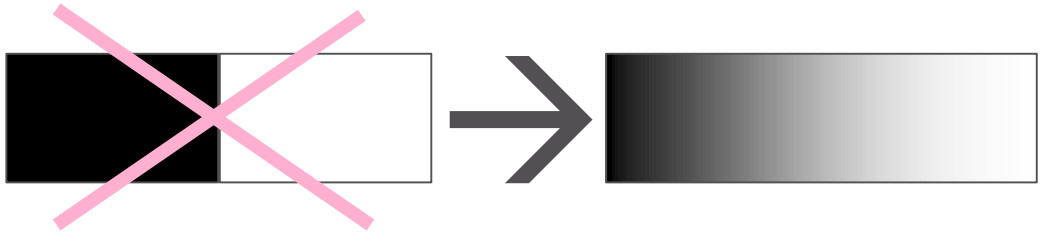
Trials under the alternative hypothesis (aspirin better)




Power=80% (RD=10% i.e. 40 vs. 50%)
0.4
0.2
0.0
-0.2
-0.4
0
50
100
150
200
Trial (N=200)
True difference:
10%
Significance
p</=.05 (N=160)
p>.05 (N=40)
p</=.05 but wrong direction (N=0)



1. P-value ≠ probability of null hypothesis being true!

2. 

Thank You!

References

> Goodman S. A dirty dozen: twelve p-value misconceptions. Semin Hematol. 2008; 45: 135-40.
doi: 10.1053/j.seminhematol.2008.04.003.

> Greenland S et al. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. Eur J Epidemiol. 2016; 31: 337–350.
doi: 10.1007/s10654-016-0149-3