Sequential Alternatives to the Two-Trials Rule

Leonhard Held

University of Zurich UZH

Workshop on Adaptive Designs and Multiple Testing Procedures

Basel, April 20, 2023

The Two-Trials Rule

– FDA requires

"at least two adequate and well-controlled studies, each convincing on its own, to establish effectiveness."

The Two-Trials Rule

– FDA requires

> *"at least two adequate and well-controlled studies, each convincing on its own, to establish effectiveness."*

– Usually implemented requiring one-sided $p \leq \alpha = 0.025$ in two independent studies.

The Two-Trials Rule

– FDA requires

> *"at least two adequate and well-controlled studies, each convincing on its own, to establish effectiveness."*

– Usually implemented requiring one-sided $p \leq \alpha = 0.025$ in two independent studies.

$\rightarrow$ Overall T1E rate is $\alpha_{\text{overall}} = \alpha^2 = 0.025^2 = 0.000625$.

The Two-Trials Rule

– FDA requires

> *"at least two adequate and well-controlled studies, each convincing on its own, to establish effectiveness."*

– Usually implemented requiring one-sided $p \leq \alpha = 0.025$ in two independent studies.

$\rightarrow$ Overall T1E rate is $\alpha_{\text{overall}} = \alpha^2 = 0.025^2 = 0.000625$.

– Nice review in Kennedy-Shaffer (2017)

More Than Two Trials

ORIGINAL RESEARCH

A Generalization of the Two Trials Paradigm

Gerd K. Rosenkranz[1]

More Than Two Trials

ORIGINAL RESEARCH

A Generalization of the Two Trials Paradigm

Gerd K. Rosenkranz[1]

– Requiring at least 2 out of $n > 2$ trials to be significant at 0.025 is too lax.

More Than Two Trials

ORIGINAL RESEARCH

A Generalization of the Two Trials Paradigm

Gerd K. Rosenkranz[1]

– Requiring at least 2 out of $n > 2$ trials to be significant at 0.025 is too lax.

– Requiring all n trials to be significant at 0.025 is too stringent.

More Than Two Trials

ORIGINAL RESEARCH

A Generalization of the Two Trials Paradigm

Gerd K. Rosenkranz[1]

– Requiring at least 2 out of $n > 2$ trials to be significant at 0.025 is too lax.

– Requiring all n trials to be significant at 0.025 is too stringent.

– Rosenkranz (2023) addresses the problem by requiring that

> *"the type-I error rate of any procedure involving more than two trials shall equal the type-I error rate from the two trials rule."*

More Than Two Trials

ORIGINAL RESEARCH

A Generalization of the Two Trials Paradigm

Gerd K. Rosenkranz[1]

– Requiring at least 2 out of $n > 2$ trials to be significant at 0.025 is too lax.

– Requiring all n trials to be significant at 0.025 is too stringent.

– Rosenkranz (2023) addresses the problem by requiring that

> *"the type-I error rate of any procedure involving more than two trials shall equal the type-I error rate from the two trials rule."*

→ "2-of-3 rule" with adjusted significance level $\alpha_* = 0.0145$

The Harmonic Mean χ^2 Test

The harmonic mean χ^2-test to substantiate scientific findings

Leonhard Held

University of Zurich, Switzerland

The Harmonic Mean χ^2 Test

The harmonic mean χ^2-test to substantiate scientific findings

Leonhard Held

University of Zurich, Switzerland

– Compute the harmonic mean Z_H^2 of squared Z-values $Z_1^2, \ldots, Z_n^2$ to obtain

$$X_n^2 = n \cdot Z_H^2 = \frac{n^2}{\sum_{i=1}^{n} 1/Z_i^2} \overset{H_0}{\sim} \chi^2(1 \text{ df})$$

The Harmonic Mean χ^2 Test

Journal of the Royal Statistical Society

Applied Statistics

Series C

The harmonic mean χ^2-test to substantiate scientific findings

Leonhard Held
University of Zurich, Switzerland

– Compute the harmonic mean Z_H^2 of squared Z-values $Z_1^2, \ldots, Z_n^2$ to obtain

$$X_n^2 = n \cdot Z_H^2 = \frac{n^2}{\sum_{i=1}^n 1/Z_i^2} \overset{H_0}{\sim} \chi^2(1 \text{ df})$$

– Exact overall T1E control at level α_{overall} is obtained with critical value c_n.

Necessary Success Bounds on p-Values

The success criterion $X_n^2 \geq c_n \iff T_n = \sum_{i=1}^{n} 1/Z_i^2 \leq d_n = \dfrac{n^2}{c_n}$

requires $1/Z_i^2 \leq d_n$, which implies necessary success bounds on all p-values:

	bound on p-value		
α_{overall}	$n = 2$	$n = 3$	$n = 4$
0.025^2	0.065	0.17	0.26

Formalizes the meaning of

> *"at least two adequate and well-controlled studies, each convincing on its own, to establish effectiveness"*

No such bounds exist for Fisher, Stouffer, 2-of-3 rule

Project Power for 2 Trials

- Of central interest is the overall project power.
- Can be easily calculated through Monte Carlo simulation:

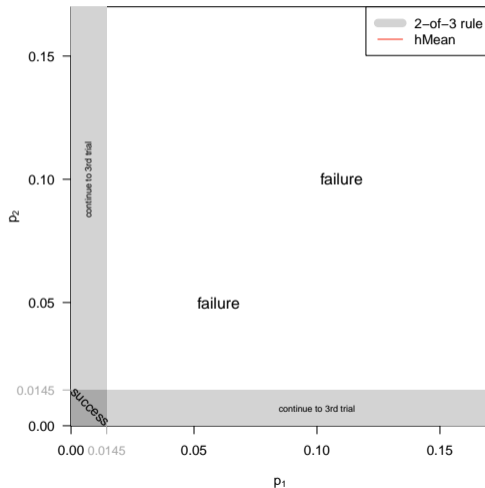| Trial power	Project power (%) two-trials rule	harmonic
80	64	71
90	81	87

Sequential Conduct of 3 Trials

- 2-of-3 rule
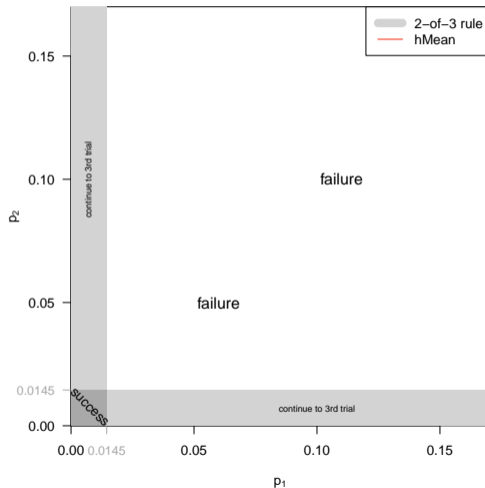 - Cannot stop after trial 1

Sequential Conduct of 3 Trials

- 2-of-3 rule
    - Cannot stop after trial 1
    - → Trial 1 and 2 could be run
      in parallel

Sequential Conduct of 3 Trials

– 2-of-3 rule
  – Cannot stop after trial 1
  $\rightarrow$ Trial 1 and 2 could be run in parallel
  – Can stop after trial 2 for both failure and success

Decisions after trial 1 and 2:
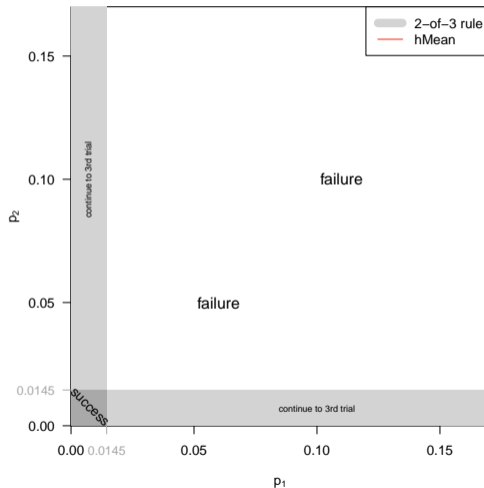
Sequential Conduct of 3 Trials

– 2-of-3 rule
  – Cannot stop after trial 1
  $\rightarrow$ Trial 1 and 2 could be run in parallel
  – Can stop after trial 2 for both failure and success
  – Failure if $p_1, p_2 = 0.02$ (!)

Decisions after trial 1 and 2:

Sequential Conduct of 3 Trials

– 2-of-3 rule
  – Cannot stop after trial 1
  $\rightarrow$ Trial 1 and 2 could be run in parallel
  – Can stop after trial 2 for both failure and success
  – Failure if $p_1, p_2 = 0.02$ (!)
– Harmonic mean test
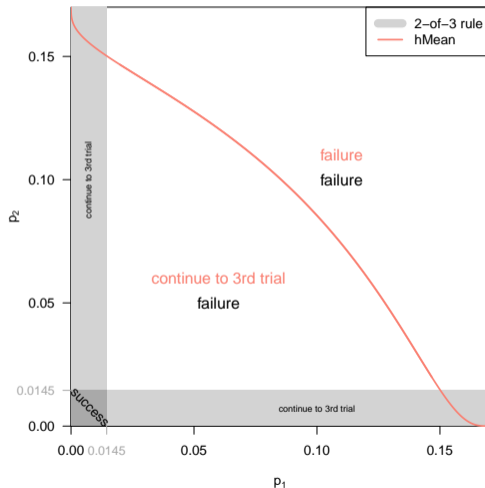 – Can stop after trial 1 for failure (if $p_1 > 0.17$)

Decisions after trial 1 and 2:

Sequential Conduct of 3 Trials

– 2-of-3 rule
  – Cannot stop after trial 1
  $\rightarrow$ Trial 1 and 2 could be run in parallel
  – Can stop after trial 2 for both failure and success
  – Failure if $p_1, p_2 = 0.02$ (!)
– Harmonic mean test
  – Can stop after trial 1 for failure (if $p_1 > 0.17$)
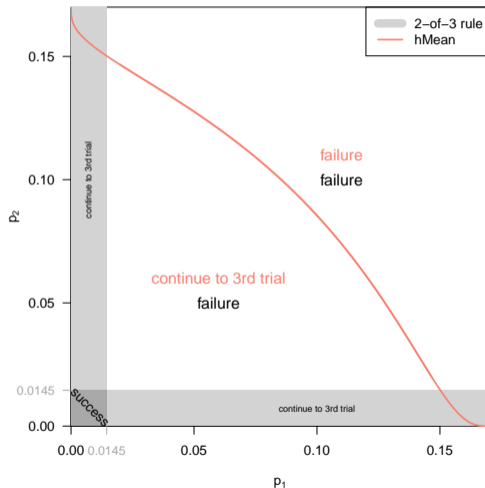 – Can stop after trial 2 for failure

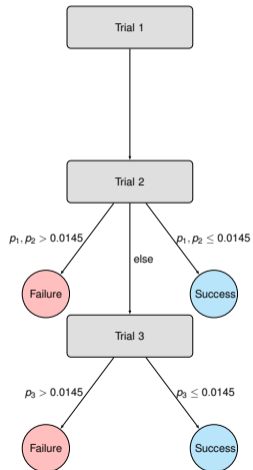Decisions after trial 1 and 2:

Sequential Conduct of 3 Trials

- 2-of-3 rule
    - Cannot stop after trial 1
  $\rightarrow$ Trial 1 and 2 could be run
    in parallel
    - Can stop after trial 2 for
    both failure and success
    - Failure if $p_1, p_2 = 0.02$ (!)
- Harmonic mean test
    - Can stop after trial 1 for
    failure (if $p_1 > 0.17$)
    - Can stop after trial 2 for
    failure
    - Success is only possible
    with 3 trials

Decisions after trial 1 and 2:
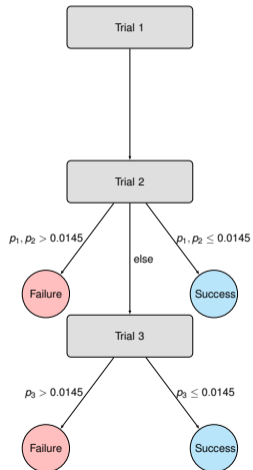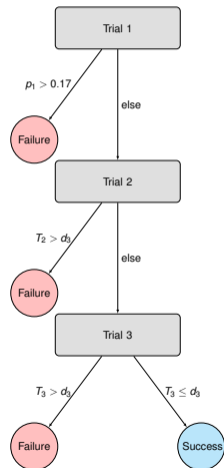
Sequential Conduct of 3 Trials

2-of-3 rule

```
                    ┌─────────┐
                    │ Trial 1 │
                    └────┬────┘
                         │
                    ┌────┴────┐
                    │ Trial 2 │
                    └────┬────┘
```
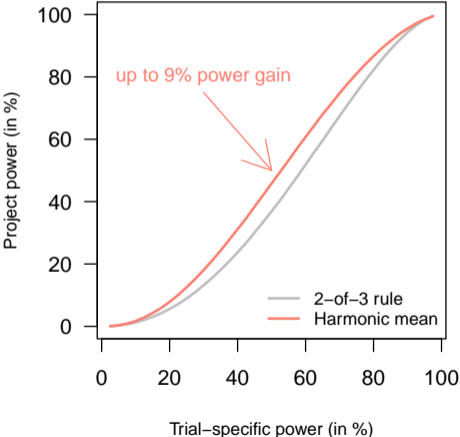
$p_1, p_2 > 0.0145$ else $p_1, p_2 \leq 0.0145$

Failure   Trial 3   **Success**

$p_3 > 0.0145$ $p_3 \leq 0.0145$

Failure   **Success**

Sequential Conduct of 3 Trials

2-of-3 rule

Harmonic mean

Project Power and Expected Number of Trials

Project power

Project power (in %) vs Trial–specific power (in %)

up to 9% power gain

2–of–3 rule
Harmonic mean

Expected number of trials

Expected number of trials vs Trial–specific power (in %)

more well–powered trials needed

2–of–3 rule
Harmonic mean

The Sequential Harmonic Mean Test

– Test for success after 2nd and 3rd trial at level α_{adjusted}.

– Exact T1E control at level α_{overall} requires multiplicity adjustments based on
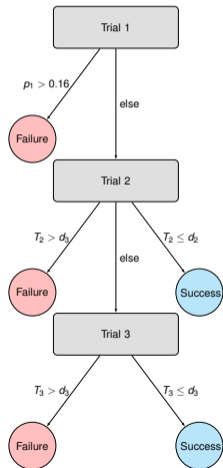
$$T_3 = \underbrace{1/Z_1^2 + 1/Z_2^2}_{= T_2 \, \sim \, \text{IG}(1/2,2)} + \underbrace{1/Z_3^2}_{\sim \, \text{IG}(1/2,1/2)}$$

– Can be extended to test also after 1st trial.

$\rightarrow$ Adjusted level and thresholds d_i for $\alpha_{\text{overall}} = 0.025^2$:

		threshold		
	α_{adjusted}	d_1	d_2	d_3
hMean	0.025^2			1.14
2|3 seq hMean	0.018^2		0.39	1.00
1|2|3 seq hMean	0.015^2	0.081	0.36	0.92

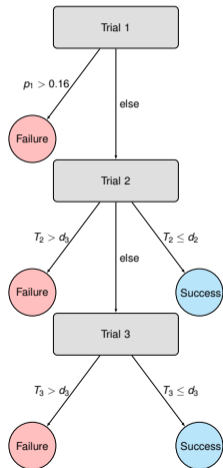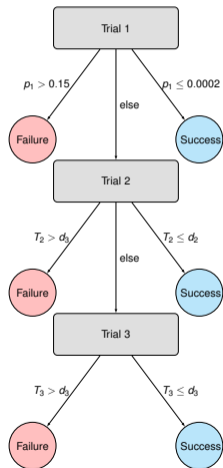# The Sequential Harmonic Mean Test

2|3 version

The Sequential Harmonic Mean Test
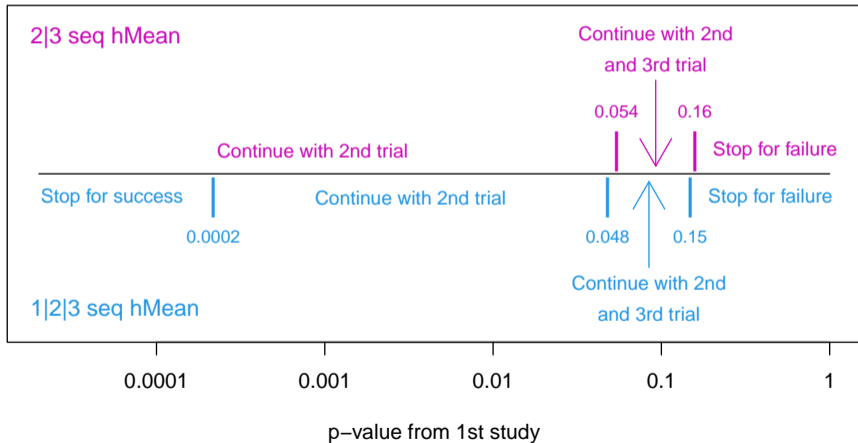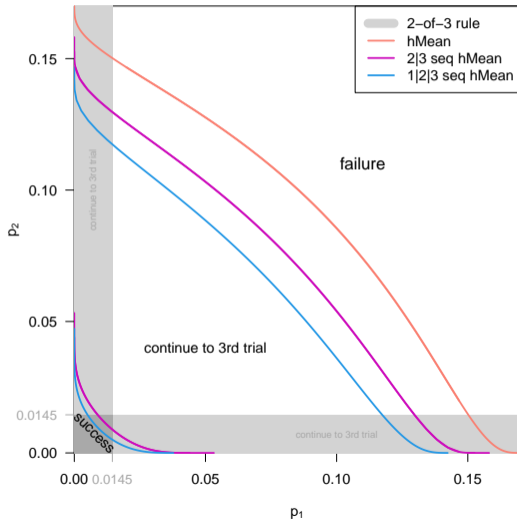
Possible Decisions after Trial 1

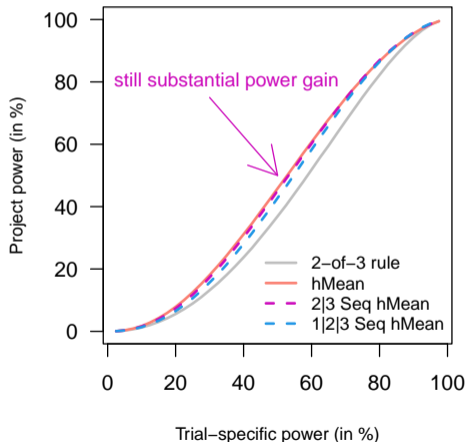Trials 2 and 3 can sometimes be conducted in parallel:
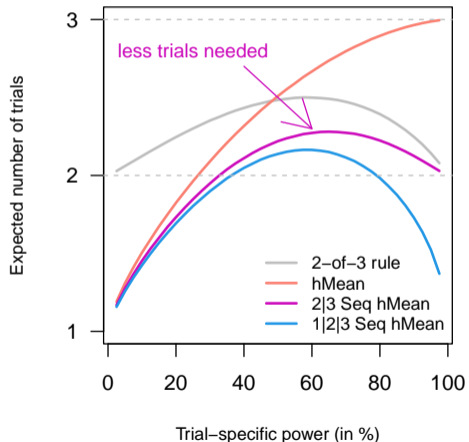
Possible Decisions after Trial 2

Project Power and Expected Number of Trials

Project power

Expected number of trials

Summary

The harmonic mean χ^2 test
– requires each trial to be convincing on its own

Summary

The harmonic mean χ^2 test
- requires each trial to be convincing on its own
- allows for stopping after the 1st trial

Summary

The harmonic mean χ^2 test
  – requires each trial to be convincing on its own
  – allows for stopping after the 1st trial
  – leads to more appropriate inferences than the 2-of-3 rule

Summary

The harmonic mean χ^2 test

– requires each trial to be convincing on its own

– allows for stopping after the 1st trial

– leads to more appropriate inferences than the 2-of-3 rule

– can be applied sequentially to stop for success after the 2nd or 1st trial

Summary

The harmonic mean χ^2 test
- requires each trial to be convincing on its own
- allows for stopping after the 1st trial
- leads to more appropriate inferences than the 2-of-3 rule
- can be applied sequentially to stop for success after the 2nd or 1st trial
- has increased project power and requires less trials

References

Held, L. (2020). The harmonic mean χ^2-test to substantiate scientific findings. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 69(3):697–708.

Kennedy-Shaffer, L. (2017). When the alpha is the omega: *P*-values, "substantial evidence," and the 0.05 standard at FDA. *Food and Drug Law Journal*, 72(4):595–635.

Rosenkranz, G. K. (2023). A generalization of the two trials paradigm. *Therapeutic Innovation & Regulatory Science*, 57:316–320.