The harmonic mean χ^2 test to substantiate scientific findings

Leonhard Held

University of Zurich[UZH]

GMDS-CEN Conference 2020

Satellite Webinar "Long-run behaviour of Bayesian procedures"

September 16, 2020

Introduction

The Harmonic Mean χ^2 Test

Discussion

Introduction

- Replicability of research findings is crucial to the credibility of science.
- Large-scale replication projects have been conducted in the last years.
- Such efforts help to assess to what extent results from original studies can be confirmed in independent replication studies.

UZH

Center for Reproducible Science

Replication is Standard in Drug Regulation

- FDA/EMA 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 ("two-trials rule").
- However, this may not reflect the available evidence:
  - $p_1 = p_2 = 0.024$ leads to claim of success.
  - $p_1 = 0.026$ and $p_2 = 0.001$ leads to no claim of success.

The harmonic mean χ^2 test leads to more appropriate inferences.

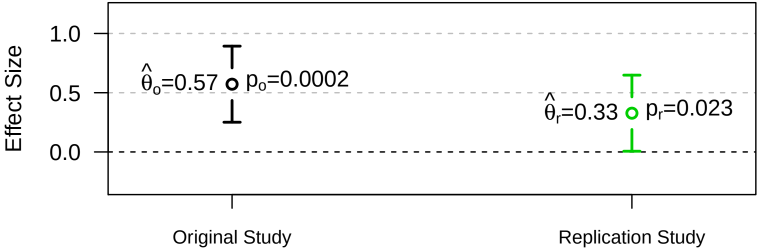
Combining and Pooling P-Values

- It is not clear how to extend the rule to results from $n > 2$ studies:
  - Requiring at least 2 out of n studies to be significant is too lax.
  - Requiring all n studies to be significant is too stringent.
- Fisher's combined or Stouffer's pooled method is sometimes used, but not without problems:
  - $p_1 = 0.0001$ and $p_2 = 0.5$ gives Fisher's $p = 0.0005 < 0.025^2$.
  - $p_1 = 0.01$ and $p_2 = 0.01$ gives Fisher's $p = 0.001 > 0.025^2$.

The harmonic mean χ^2 test leads to more appropriate inferences.

Analysis of Replication Studies

Effect estimates with 95% confidence interval



$\hat{\theta}_o$	Effect estimate	$\hat{\theta}_r$
σ_o	Standard error	σ_r
z_o	Test statistic	z_r
p_o	p-value (one-sided)	p_r

Analysis and Design of Replication Studies

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistics in Society
Series A

J. R. Statist. Soc. A (2020)

A new standard for the analysis and design of replication studies

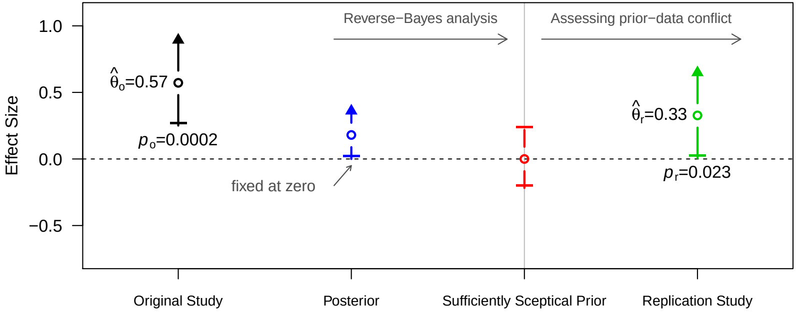
Leonhard Held

University of Zurich, Switzerland

[*Read before* The Royal Statistical Society *at a meeting on 'Signs and sizes: understanding and replicating statistical findings' at the Society's 2019 annual conference in Belfast on Wednesday, September 4th, 2019, the President,* Professor D. Ashby, *in the Chair*]

https://doi.org/10.1111/rssa.12493

Reverse-Bayes Analysis



A New Standard for the Analysis and Design of Replication Studies

A combination of

- Analysis of Credibility (Matthews, 2001, 2018)
- Assessment of Prior-Data Conflict (Box, 1980)

leads to

- A new definition of replication success
- The degree of replication success quantified by the sceptical p-value p_S
- If the two studies are equally sized and $\text{sign}(\hat{\theta}_o) = \text{sign}(\hat{\theta}_r)$ then

$$p_S = 1 - \Phi(z_S) \text{ where } z_S^2 = \frac{1}{1/z_o^2 + 1/z_r^2}$$

Introduction

The Harmonic Mean χ^2 Test

Discussion

Publication

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Applied Statistics
Series C

Appl. Statist. (2020)

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

Leonhard Held

University of Zurich, Switzerland

https://doi.org/10.1111/rssc.12410

The Harmonic Mean χ^2 Test

$n = 2$ studies

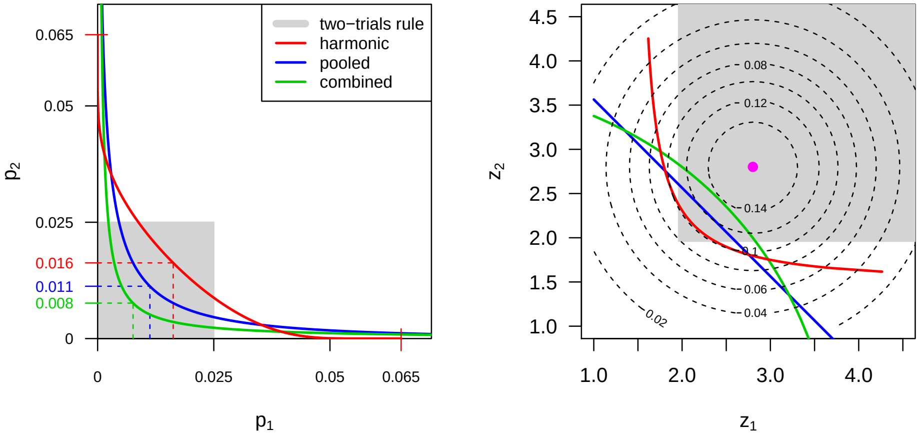
- Transform two (one-sided) p-values p_1, p_2 to z-values $z_i = \Phi^{-1}(1 - p_i)$.
- Compute

$$X^2 = \frac{4}{1/z_1^2 + 1/z_2^2}$$

- The null distribution of X^2 is $\chi^2(1)$.
- A one-sided p-value can be calculated.
- Exact Type-I error rate control can be achieved.

Comparison With the Two-Trials Rule

Type-I error rate control at 0.025^2



Project Power

– Of central interest is the overall power for the project (project power).
– Project power can easily be calculated through Monte Carlo simulation.

Project Power

	Project power (%)			
Trial power	two-trials rule	harmonic	combined	pooled
80	64	71	74	77
90	81	87	90	91

The General Harmonic Mean χ^2 Test

- The approach can be generalized

to n studies:

$$X^2 = \frac{n^2}{\sum_{i=1}^{n} 1/z_i^2}$$

and can include weights w_i:

$$X_w^2 = \frac{w^2}{\sum_{i=1}^{n} w_i/z_i^2} \text{ where } w = \sum_{i=1}^{n} \sqrt{w_i}$$

- The null distribution of X^2 resp. X_w^2 is still $\chi^2(1)$.
- A one-sided p-value can be calculated.

Bounds on p-Values

Bounds for p-values from n studies at level 0.025^2

bound	$n = 2$	$n = 3$	$n = 4$	$n = 5$	$n = 6$
necessary	0.065	0.17	0.26	0.32	0.37
sufficient	0.016	0.053	0.099	0.15	0.20

Formalizing the meaning of

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

Bounds on p-Values

Bounds for p-values from n studies at level 0.025^2

bound	$n = 2$	$n = 3$	$n = 4$	$n = 5$	$n = 6$
necessary	0.065	0.17	0.26	0.32	0.37
sufficient	0.016	0.053	0.099	0.15	0.20

Formalizing the meaning of

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

Application

Results from 5 clinical trials on the effect of Carvedilol on mortality for the treatment of patients with moderate to severe heart failure (from Fisher, 1999):

study number	p-value	hazard ratio	standard error
220	0.00025	0.27	0.41
240	0.0245	0.22	0.85
223	0.128	0.72	0.29
221	0.1305	0.57	0.51
239	0.2575	0.53	1.02

combined $p = 0.00013 < 0.025^2$

pooled $p = 0.00009 < 0.025^2$

harmonic $p = 0.00048 < 0.025^2$

weighted harmonic $p = 0.00034 < 0.025^2$

Application

Modified data: Double the *p*-value of study 223

study number	*p*-value	hazard ratio	standard error
220	0.00025	0.27	0.41
240	0.0245	0.22	0.85
223	0.256	0.83	0.29
221	0.1305	0.57	0.51
239	0.2575	0.53	1.02

combined $p = 0.00021 < 0.025^2$

pooled $p = 0.00022 < 0.025^2$

harmonic $p = 0.0012 > 0.025^2$

weighted harmonic $p = 0.0027 > 0.025^2$

Confidence intervals

The harmonic χ^2 test can be inverted to obtain a confidence interval:

- Consider test statistic $Z_i = (\hat{\theta}_i - \mu)/\sigma_i$ for general H_0: $\theta = \mu$.
- Consider two-sided p-values to represent the common scenario that an initial study is two-sided and all following studies are one-sided.
- Derive confidence interval from p-value function.

study number	p-value	hazard ratio	standard error
220	0.00025	0.27	0.41
240	0.0245	0.22	0.85
223	0.128	0.72	0.29
221	0.1305	0.57	0.51
239	0.2575	0.53	1.02

$\rightarrow$ 95% confidence interval for hazard ratio: 0.21 to 0.73.

Discussion

"p-values are just too familiar and useful to ditch"
David Spiegelhalter (2017)

The harmonic mean χ^2 test

- leads to more appropriate inferences than the two-trials rule
- has more project power than the two-trials rule
- provides a principled extension to analyse results from more than two trials
- allows for weights
- implies restrictions on study-specific *p*-values, requesting each trial to be convincing on its own
- Software available in R-package `ReplicationSuccess` on `R-Forge`

Backup: Conditional Power

Power to detect the observed effect from the first study with an identical second study