Sample Size Calculation for Replication Studies

Charlotte Micheloud, Manuela Ott, Leonhard Held

University of Zurich
Department of Biostatistics
Center for Reproducible Science

May 23, 2019

Introduction

- Replicability: ability of confirming the result of a study when new data are collected
- Replication crisis
- Increasing interest in large-scale replication projects

Replicability of psychological science

RESEARCH ARTICLE

PSYCHOLOGY

Estimating the reproducibility of psychological science

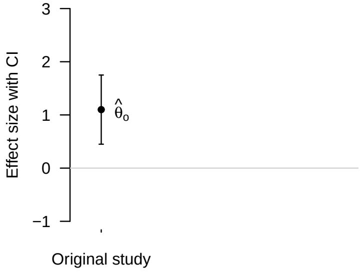
Open Science Collaboration*†

Reproducibility is a defining feature of science, but the extent to which it characterizes current research is unknown. We conducted replications of 100 experimental and correlational studies published in three psychology journals using high-powered designs and original materials when available. Replication effects were half the magnitude of original effects, representing a substantial decline. Ninety-seven percent of original studies had statistically significant results. Thirty-six percent of replications had statistically significant results; 47% of original effect sizes were in the 95% confidence interval of the replication effect size; 39% of effects were subjectively rated to have replicated the original result; and if no bias in original results is assumed, combining original and replication results left 68% with statistically significant effects. Correlational tests suggest that replication success was better predicted by the strength of original evidence than by characteristics of the original and replication teams.

- Open Science Collaboration (2015)
- Replication of 100 studies
  $\rightarrow$ Statistical significance in
  - 97% of original studies
  - 36% of replication studies

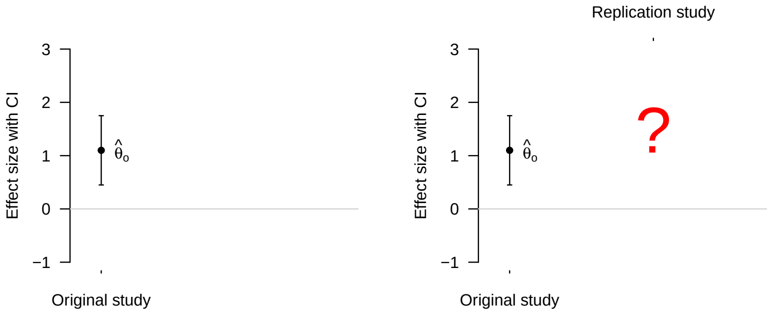
Setup

- $\hat{\theta}_o$ effect estimate of the original study
- Outcome assumed to be normally distributed



Setup

- $\hat{\theta}_o$ effect estimate of the original study
- We want to conduct a replication study and find $\hat{\theta}_r$



Same sample size as in the original study

- Taking the same sample size as in the original study
- Relative sample size $c = n_r / n_o = 1$

STATISTICS IN MEDICINE, VOL. 11, 875–879 (1992)

A COMMENT ON REPLICATION, *P*-VALUES AND EVIDENCE

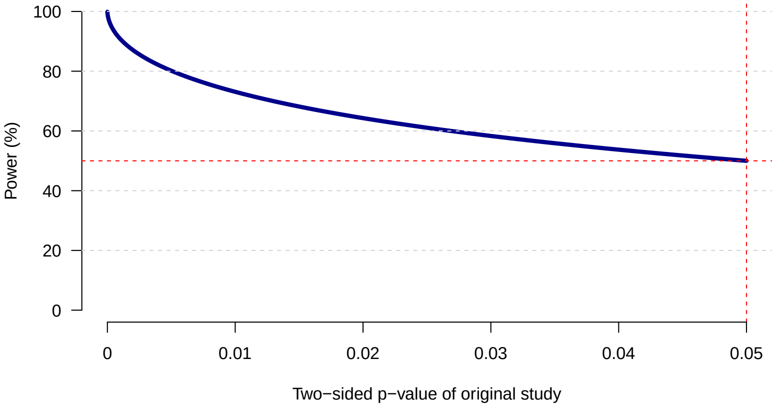
STEVEN N. GOODMAN

Johns Hopkins University School of Medicine, Department of Oncology, Division of Biostatistics, 550 N. Broadway, Suite 1103, Baltimore MD 21205, U.S.A.

$\rightarrow$ Low power even if the original effect estimate is the true effect

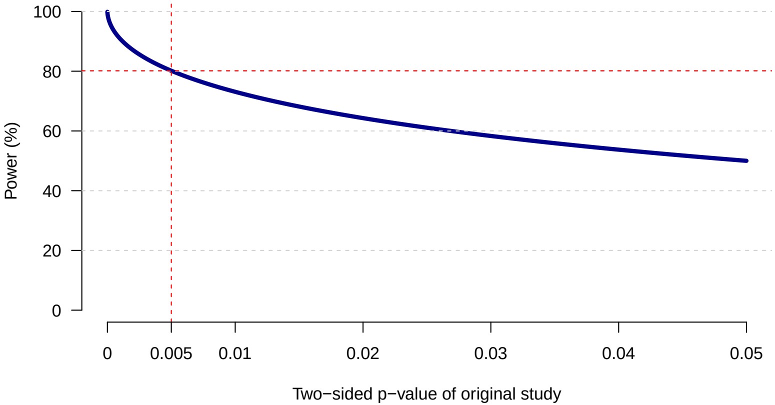
Same sample size as in the original study

Replication power for $c = 1$ assuming $\theta = \hat{\theta}_o$



Same sample size as in the original study

Replication power for $c = 1$ assuming $\theta = \hat{\theta}_o$



Standard method

Conditional power

$$\Pr\left(\text{reject } H_0 \mid \theta = \hat{\theta}_o\right)$$

$\rightarrow$ Does not incorporate the uncertainty of $\hat{\theta}_o$

Incorporation of the uncertainty

How to incorporate the uncertainty of $\hat{\theta}_o$?

$\rightarrow$ By using a prior distribution for θ:

$$\theta \sim \mathrm{N}\left(\hat{\theta}_o, \sigma^2/n_o\right)$$

$\rightarrow$ Design vs. analysis prior

$\rightarrow$ Spiegelhalter et al. (2004)

Power calculation methods

		Analysis	
		Flat prior	Normal prior
Design	Point prior	Standard	
	Normal prior		

Power calculation methods

		Analysis	
		Flat prior	Normal prior
Design	Point prior	Standard	
	Normal prior	Hybrid	

Power calculation methods

		Analysis	
		Flat prior	Normal prior
Design	Point prior	Standard	
	Normal prior	Hybrid	Bayesian

Power calculation methods

Design		Analysis: Flat prior	Analysis: Normal prior
	Point prior	Standard	Cond. Bayesian
	Normal prior	Hybrid	Bayesian

Power calculation methods

		Analysis	
		Flat prior	Normal prior
Design	Point prior	Standard	Cond. Bayesian
	Normal prior	Hybrid	Bayesian

→ Power only depends on $c = n_r/n_o$, p_o and α

Power calculation methods

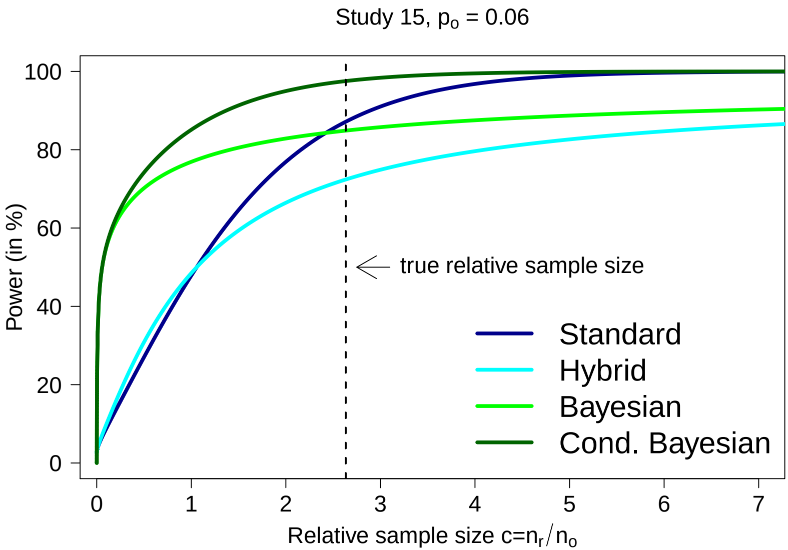
Application to the OSC replication project

$\rightarrow$ Power as a function of relative sample size $c = n_r / n_o$

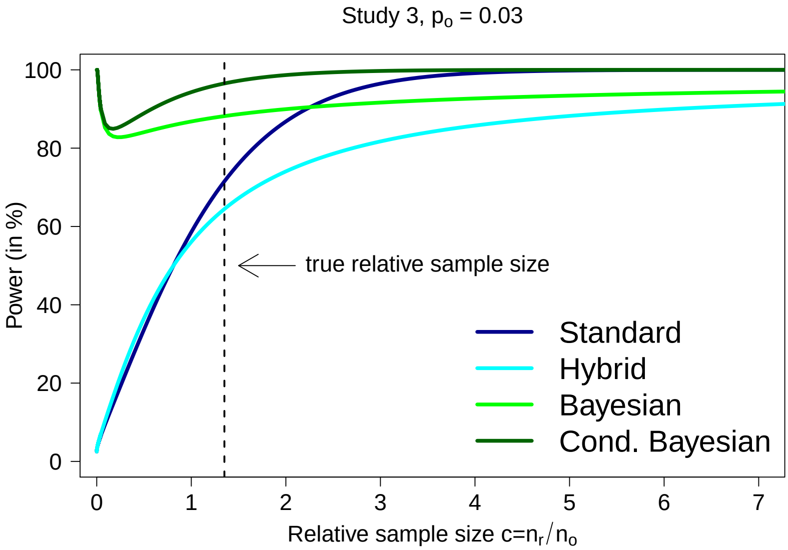
$\rightarrow$ Three studies with different p-values p_o

$\rightarrow$ $\alpha = 5\%$

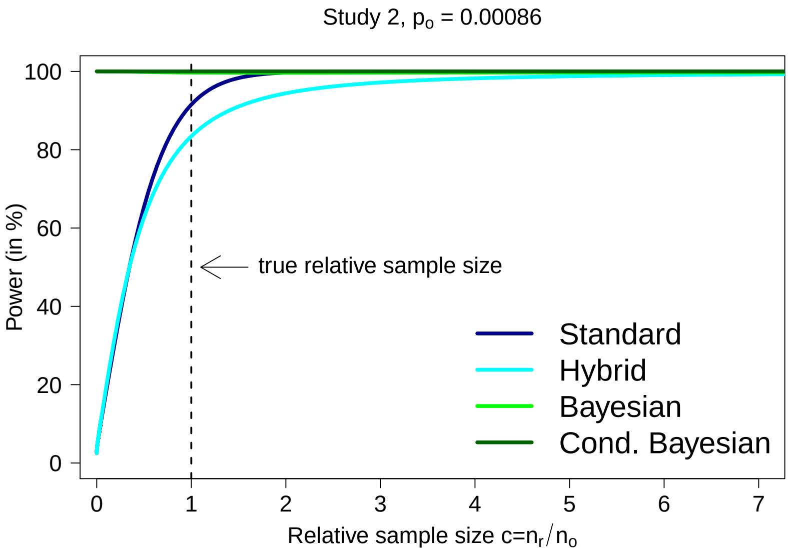
Application to OSC replication project



Application to OSC replication project

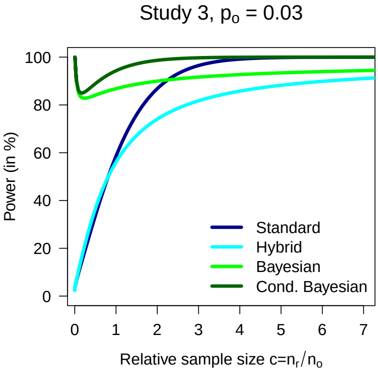


Application to OSC replication project



Theory

Predictive power



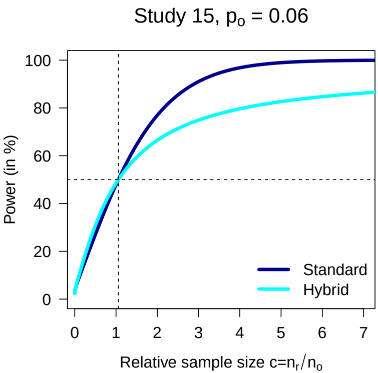
Hybrid and Bayesian power

$\rightarrow \lim_{c\rightarrow\infty}(\text{pred. pow}) = 1 - p_o/2$

$\rightarrow$ Grouin et al. (2007)

Theory

Conditional and predictive power



Standard and hybrid power

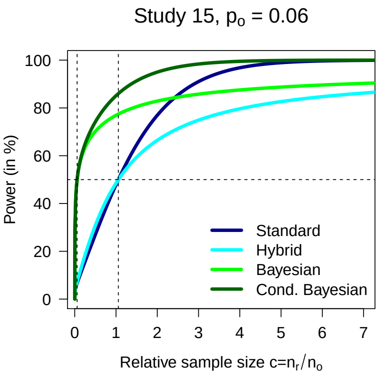
→ Cross at power = 50%

→ Spiegelhalter et al. (2004)

→ At $c = z^2_{1-\alpha/2}/t^2_o$

Theory

Conditional and predictive power



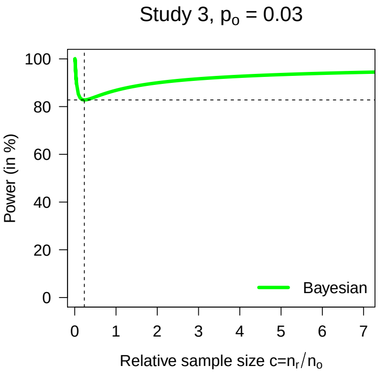
Bayesian and conditional Bayesian power

→ Cross at power = 50%

→ At $c = z^2_{1-\alpha/2} / t^2_o - 1$

Theory

Bayesian power



Bayesian power

$\rightarrow$ Non-monotone for significant original studies

$\rightarrow$ Minimum at power $= \Phi\left(\sqrt{t_o^2 - z_{1-\alpha/2}^2}\right)$

$\rightarrow$ At $c = t_o^2/z_{1-\alpha/2}^2 - 1$

$\rightarrow$ Dallow and Fina (2011)

Outlook

Comparison with sceptical *p*-value (Held, 2019)

- New definition of replication success
- Based on a reverse-Bayes approach
- Incorporates $\hat{\theta}_o$ and $\hat{\theta}_r$

  $\rightarrow$ Possible to compute conditional and predictive power for replication success

References I

Nigel Dallow and Paolo Fina.
The perils with the misuse of predictive power.
Pharmaceutical Statistics, 10(4):311–317, 2011.
doi: 10.1002/pst.467.

Steven N Goodman.
A comment on replication, *p*-values and evidence.
Statistics in Medicine, 11(7):875–879, 1992.
doi: 10.1002/sim.4780110705.

Jean-Marie Grouin, Maylis Coste, Pierre Bunouf, and Bruno Lecoutre.
Bayesian sample size determination in non-sequential clinical trials: Statistical aspects and some regulatory considerations.
Statistics in Medicine, 26(27):4914–4924, 2007.

Leonhard Held.
A new standard for the analysis and design of replication studies.
2019.
URL https://arxiv.org/abs/1811.10287.

Open Science Collaboration.
Estimating the reproducibility of psychological science.
Science, 112(517):1–10, 2015.
doi: 10.1126/science.aac4716.

David J Spiegelhalter, Keith R Abrams, and Jonathan P Myles.
Bayesian Approaches to Clinical Trials and Health-Care Evaluation, volume 13.
John Wiley & Sons, 2004.