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
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.
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 $\hat{\theta}_r$

![Graph showing original study and replication study effect sizes with confidence intervals.]

Original study

Replication study

$\hat{\theta}_o$
Same sample size as in the original study

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

→ 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**

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

→ 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 $\theta$:

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

$\rightarrow$ Design vs. analysis prior

$\rightarrow$ Spiegelhalter et al. (2004)
## Power calculation methods

<table>
<thead>
<tr>
<th>Design prior</th>
<th>Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flat prior</td>
<td>Normal prior</td>
</tr>
<tr>
<td>Point</td>
<td>Standard</td>
</tr>
<tr>
<td>Normal</td>
<td></td>
</tr>
</tbody>
</table>

Power only depends on \( c = \frac{n_r}{n_o} \), \( p \), and \( \alpha \).
## Power calculation methods

<table>
<thead>
<tr>
<th>Design</th>
<th>Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Point prior</td>
<td>Flat prior</td>
</tr>
<tr>
<td>Normal prior</td>
<td>Normal prior</td>
</tr>
<tr>
<td></td>
<td>Standard</td>
</tr>
<tr>
<td></td>
<td>Hybrid</td>
</tr>
</tbody>
</table>
### Power calculation methods

<table>
<thead>
<tr>
<th>Design prior</th>
<th>Analysis: Flat prior</th>
<th>Analysis: Normal prior</th>
</tr>
</thead>
<tbody>
<tr>
<td>Point prior</td>
<td>Standard</td>
<td></td>
</tr>
<tr>
<td>Normal prior</td>
<td>Hybrid</td>
<td>Bayesian</td>
</tr>
</tbody>
</table>
### Power calculation methods

<table>
<thead>
<tr>
<th>Design</th>
<th>Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Point prior</td>
<td>Flat prior</td>
</tr>
<tr>
<td>Normal prior</td>
<td>Standard</td>
</tr>
<tr>
<td></td>
<td>Hybrid</td>
</tr>
</tbody>
</table>

Power only depends on \( c = \frac{n_r}{n_o}, p_0 \) and \( \alpha \).
## Power calculation methods

<table>
<thead>
<tr>
<th>Design</th>
<th>Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Point prior</td>
<td>Flat prior</td>
</tr>
<tr>
<td></td>
<td>Standard</td>
</tr>
<tr>
<td>Normal prior</td>
<td>Hybrid</td>
</tr>
</tbody>
</table>

→ Power only depends on $c = \frac{n_r}{n_o}$, $p_o$ and $\alpha$
Power calculation methods

Application to the OSC replication project

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

→ Three studies with different $p$-values $p_o$

→ $\alpha = 5\%$
Application to OSC replication project

Study 15, \( p_o = 0.06 \)

<table>
<thead>
<tr>
<th>Power (in %)</th>
<th>Standard</th>
<th>Hybrid</th>
<th>Bayesian</th>
<th>Cond. Bayesian</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>40</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>60</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>80</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Relative sample size \( c = \frac{n_r}{n_o} \)

True relative sample size

0 1 2 3 4 5 6 7

0
20
40
60
80
100

Relative sample size c=n_r/n_o
Application to OSC replication project

Study 3, $p_o = 0.03$

Power (in %)

Relative sample size $c = n_r / n_o$

- Standard
- Hybrid
- Bayesian
- Cond. Bayesian

true relative sample size
Application to OSC replication project

Study 2, $p_o = 0.00086$

Power (in %)

- Standard
- Hybrid
- Bayesian
- Cond. Bayesian

Relative sample size $c = n_r/n_o$

true relative sample size
Theory

Predictive power

Hybrid and Bayesian power

\[ \lim_{c \to \infty} (\text{pred. pow}) = 1 - \frac{p_0}{2} \]

\[ \rightarrow \quad \text{Grouin et al. (2007)} \]
Theory

Conditional and predictive power

Study 15, $p_o = 0.06$

- Cross at power = 50%
- Spiegelhalter et al. (2004)
- At $c = \frac{z_{1-\alpha/2}^2}{t_o^2}$
Theory

Conditional and predictive power

Study 15, $p_o = 0.06$

Bayesian and conditional Bayesian power

→ Cross at power = 50%

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

→ Non-monotone for significant original studies

→ Minimum at power
  \[ \Phi \left( \sqrt{t^2_o - z^2_{1-\alpha/2}} \right) \]

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

→ 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 $\hat{\theta}_r$
  
  → Possible to compute conditional and predictive power for replication success
Nigel Dallow and Paolo Fina.
The perils with the misuse of predictive power.

Steven N Goodman.
A comment on replication, *p*-values and evidence.

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.

Leonhard Held.
A new standard for the analysis and design of replication studies.
2019.

Open Science Collaboration.
Estimating the reproducibility of psychological science.
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.