Sample Size Calculation for Replication Studies

Master Thesis in Biostatistics (STA495)

by

Charlotte Micheloud
13 - 824 - 107

supervised by

Prof. Dr. Leonhard Held
Dr. Manuela Ott

Zurich, May 2019

Contents

Preface

Abstract

As a consequence of the so-called 'replication crisis' (Ioannidis, 2005), an increasing number of replication studies have been conducted to determine the reliability of the original findings. Ideally, the procedures of the replication study should be as closely matched to the original study as possible. However, selecting the same sample size in the replication study as in the original study may lead to a severely underpowered design and as a result, true effects may not be detected. Furthermore, using standard sample size calculations is not well suited because the uncertainty of the original effect estimate is ignored.

One way of tackling this issue is to use Bayesian approaches by incorporating a normal prior centered around the original effect estimate and with variance inversely proportional to the original sample size (Spiegelhalter *et al.*, 2004). This corresponds to the concept of predictive power and generally leads to larger sample sizes than the standard method. Furthermore, the resulting power tends to one minus the one-sided p-value of the original study as the replication sample size increases. When the normal prior is also incorporated in the analysis of the replication study, the resulting Bayesian power increases as a function of the replication sample size for non-significant original studies. However, adding more subjects to the replication study may lead to a decrease of the Bayesian power if the p-value of the original study is only 'suggestive', *i.e.* only slightly below the significance level.

In a second part, we investigate an approach to declare replication success based on the sceptical p-value, a new metric introduced by Held (2019b). Conditional and predictive power calculations to reach replication success lead to larger sample sizes and emphasize the importance of intrinsically credible original studies (Held, 2019a; Matthews, 2018). We illustrate these properties using data from the Open Science Collaboration project on the replicability of psychological science (Open Science Collaboration, 2015).

Acknowledgments

First, I would like to thank my thesis supervisor Prof. Dr. Leonhard Held for his guidance and the interesting and inspiring discussions we had. I am also grateful to Dr. Manuela Ott for her precious help and availability. Undertaking this master thesis was an inspiring challenge which made me grow at many levels and this would not have been possible without them. My sincere gratitude goes to Dr. Eva Furrer for making this Master program an amazing place to progress as a scientist. I would also like to thank my fellow colleagues from the Master Program in Biostatistics and Sandra in particular, for the mental support, the valuable advice and for living in the library with me. Last but not least, a huge thanks to my family and Simon for their unconditional support and for always believing in me whatever path I choose to take.

Charlotte Micheloud
May 2019

Chapter 1

Introduction

The replicability of research findings is a defining feature of science. However, many of the scientific claims are false and thus irreplicable, as shown by Ioannidis (2005). One reason for this so-called 'replication crisis' is the use of low-powered study designs. Underpowered studies combined with the use of thresholds for statistical significance lead to inflated effect estimates (Ioannidis, 2008). Moreover, publication bias is more likely to affect small underpowered studies (Button *et al.*, 2013).

The rising awareness of the low replicability of scientific findings has led to a substantial increase of replication projects in various fields. These include psychology (Open Science Collaboration, 2015; Johnson *et al.*, 2017), social sciences (Camerer *et al.*, 2018) and economics (Camerer *et al.*, 2016) among others. Such efforts help to determine whether claims of new discoveries can be confirmed in independent replication studies whose procedures are as closely matched to the original studies as possible (Held, 2019b).

In the design phase of a replication study, the sample size determination is a crucial step. A replication study with low power to detect an effect may result in a waste of time and money. Using the same sample size as in the original study may lead to a severely underpowered replication study, even if the original study correctly estimated the true effect size (Goodman, 1992). The common approach is to use standard power calculations to estimate the sample size that is necessary to achieve a certain level of power in the replication study. In this approach, the probability of rejecting the null hypothesis H_0 given that the alternative hypothesis H_1 is true is computed. When used in the context of replication studies, the probability of rejecting H_0 is conditioned on the effect estimate of the original study which is assumed to be the true effect. However, this approach is not well suited as it ignores the uncertainty which accompanies the original effect estimate. Addressing this issue, my Master thesis aims to optimize the design of replication studies focusing on reasonable approaches for power calculation and thus sample size recommendations.

The uncertainty of the original effect estimate is taken into account by incorporating a normal prior centered around the original effect estimate and with variance inversely proportional to the original sample size (Spiegelhalter *et al.*, 2004). This corresponds to the concept of predictive power. Chapter 2 presents and deepens this approach in the context of a replication study aiming at reaching significance at a pre-specified level. Because of the lack of a single standard to assess if the replication was successful, Held (2019b) proposes the sceptical p-value, a new metric to

define replication success. This approach gives rise to new methods of calculating conditional and predictive power, as further investigated in Chapter 3. The obtained findings are illustrated using data from the Open Science Collaboration (2015) on the replicability of psychological science. In Chapter 4, the relevance of each power calculation method in the replication framework is assessed.

For simplicity, we will mostly refer to the power formulas in this report. By fixing this power to the desired level, the sample size required in the replication study can be computed.

Chapter 2

Power for Significance

In this chapter, we focus on methods calculating the power for significance of the replication study. Power calculations for significance aim to detect the effect estimate from the original study with a standard two-sided significance test. The replication study is then declared significant if the p-value of the replication study is smaller than or equal to the significance level, namely $p_r \leq \alpha$. Section 2.1 specifies the notation used in this chapter. In Section 2.2, the standard power formula is derived and followed by alternative methods that acknowledge the uncertainty of the original effect estimate $\hat{\theta}_o$. These different methods of power calculation are studied in detail in Section 2.3. Using the Open Science Collaboration (2015) data, the findings are then illustrated in Section 2.4. Formulas of this chapter are based on Spiegelhalter *et al.* (2004, Sections 6.5 & 6.6) and adapted to the replication framework.

2.1 Notation

Table 2.1 presents the notation which will be followed in this chapter.

Notation	Meaning
n_o	sample size of original study
n_r	sample size of replication study
c	relative sample size n_r/n_o
θ	true effect size
$\hat{\theta}_o$	effect estimate of the original study
$Y_{1:n_r}$	future data of the replication study
$\overline{Y}_{n_r}$	future parameter estimate of the replication study
σ	common standard deviation of one observation
$2\epsilon = \alpha$	significance level
z_ϵ	ϵ-quantile of the standard normal distribution
t_o	test statistic of the original study
t_r	test statistic of the replication study
p_o	two-sided p-value of the original study
p_r	two-sided p-value of the replication study

Table 2.1: Table presenting the notation.

2.2 Theory

This section gathers the formulas of the standard, hybrid, Bayesian and conditional Bayesian power and explains the related theory.

2.2.1 Standard method

Suppose researchers conducted a study and declared the results significant at a pre-specified level $\alpha = 2\epsilon$. In order to confirm this finding, a replication study is planned. Let us assume that the future data of the replication study are normally distributed as follows,

$$Y_1, \ldots, Y_{n_r} \overset{\text{iid}}{\sim} \mathrm{N}\left(\theta, \sigma^2\right) ,$$

where θ is the true effect size and σ is the known standard deviation of one observation. Let the sample mean of $Y_{1:n_r}$ be the parameter estimate of the replication study. The parameter estimate $\overline{Y}_{n_r}$ has distribution

$$\overline{Y}_{n_r} \sim \mathrm{N}\left(\theta, \frac{\sigma^2}{n_r}\right) , \tag{2.1}$$

with n_r being the planned sample size in the replication study. Let us suppose the null hypothesis of the replication study is H_0: $\theta = 0$ and we want to detect an alternative hypothesis H_1: $\theta = \hat{\theta}_o > 0$, where $\hat{\theta}_o$ is the effect estimate of the original study. The corresponding standardized test statistic t_r is $\overline{Y}_{n_r}\sqrt{n_r}/\sigma$ and we declare the result statistically significant at the two-sided $\alpha = 2\epsilon$ level if $|t_r| > z_{1-\alpha/2} = -z_\epsilon$. In the following, we focus on $t_r > -z_\epsilon$ as $t_r < z_\epsilon$ is relatively small for $\hat{\theta}_o > 0$. H_0 will thus be rejected when the parameter estimate $\overline{Y}_{n_r}$ obeys

$$\overline{Y}_{n_r} > -\frac{1}{\sqrt{n_r}} z_\epsilon \sigma \, . \tag{2.2}$$

This event is denoted by S_ϵ^C and is called the 'Classical significance' as a classical (frequentist) analysis will be conducted at the end of the replication study and is opposed to 'Bayesian significance' which will become a relevant concept later in this report. Under H_1, $\overline{Y}_{n_r}$ is normally distributed with mean $\mathrm{E}\left(\overline{Y}_{n_r}\right) = \hat{\theta}_o$ and variance $\mathrm{Var}\left(\overline{Y}_{n_r}\right) = \sigma^2/n_r$. In order to calculate the power of the replication study, we compute the probability of Classical significance given that the effect estimate of the original study is the true effect,

$$\begin{aligned}
\Pr\left(S_\epsilon^C \,|\, \theta = \hat{\theta}_o\right) &= \Pr\left(\overline{Y}_{n_r} > -\frac{1}{\sqrt{n_r}} z_\epsilon \sigma\right) \\
&= 1 - \Pr\left(\overline{Y}_{n_r} \leq -\frac{1}{\sqrt{n_r}} z_\epsilon \sigma\right) \\
&= 1 - \Pr\left(\frac{\overline{Y}_{n_r} - \mathrm{E}(\overline{Y}_{n_r})}{\sqrt{\mathrm{Var}(\overline{Y}_{n_r})}} \leq \frac{-z_\epsilon \sigma/\sqrt{n_r} - \mathrm{E}(\overline{Y}_{n_r})}{\sqrt{\mathrm{Var}(\overline{Y}_{n_r})}}\right) \\
&= 1 - \Phi\left[\frac{-z_\epsilon \sigma/\sqrt{n_r} - \mathrm{E}(\overline{Y}_{n_r})}{\sqrt{\mathrm{Var}(\overline{Y}_{n_r})}}\right]
\end{aligned}$$

$$
\begin{aligned}
&= \Phi\left[\frac{z_\epsilon \sigma/\sqrt{n_r} + \mathrm{E}(\overline{Y}_{n_r})}{\sqrt{\mathrm{Var}(\overline{Y}_{n_r})}}\right] \\
&= \Phi\left[\frac{z_\epsilon \sigma/\sqrt{n_r} + \hat{\theta}_o}{\sqrt{\sigma^2/n_r}}\right] \\
&= \Phi\left[\frac{\hat{\theta}_o \sqrt{n_r}}{\sigma} + z_\epsilon\right] . \qquad (2.3)
\end{aligned}
$$

Equation (2.3) specifies the conditional power of a replication study with n_r subjects assuming a classical analysis of the results. The necessary sample size to achieve a pre-specified level of power in the replication study can be derived from equation (2.3) and is given by

$$n_r = \frac{(z_{1-\beta} - z_\epsilon)^2 \sigma^2}{\hat{\theta}_o^2} ,$$

where $1-\beta$ denotes the power and $z_{1-\beta}$ the $(1-\beta)$-quantile of the standard normal distribution for notational simplicity.

As stated in Chapter 1, conditioning the power on the original effect estimate $\hat{\theta}_o$, which is assumed to be the true effect size, ignores the uncertainty surrounding the estimate and may contribute to sub-optimal designs of replication studies. This issue motivates the need for alternative power calculation methods taking this uncertainty into account.

2.2.2 Alternative methods acknowledging the uncertainty of the original effect estimate $\hat{\theta}_o$

One way of incorporating the uncertainty of the original effect estimate $\hat{\theta}_o$ in the power calculation of the replication study is to consider the use of a normal prior for the true effect size,

$$\theta \sim \mathrm{N}\left(\hat{\theta}_o, \frac{\sigma^2}{n_o}\right) , \qquad (2.4)$$

which is centered around the original effect estimate $\hat{\theta}_o$ and with variance inversely proportional to the original sample size n_o. This prior can be used as a design, but also as an analysis prior (O'Hagan and Stevens, 2001). The design prior, also called sampling prior by some authors (Wang and Gelfand, 2002; Sahu and Smith, 2006), is used before the data are collected in order to quantify prior beliefs about the true effect size (Schönbrodt and Wagenmakers, 2018). It contributes to the study design but is not used in the subsequent statistical analysis. In our case, a point design prior at $\theta = \hat{\theta}_o$ corresponds to the concept of conditional power while the normal design prior (2.4) corresponds to the concept of predictive power (Spiegelhalter *et al.*, 1986). The predictive power averages the conditional power over the possible values of the true effect according to its prior distribution. Conversely, the analysis prior determines the type of analysis which will be used after the data collection. If a flat analysis prior is used, a classical analysis takes place when reporting the results whereas using the normal prior (2.4) as an analysis prior indicates a Bayesian analysis approach. Table 2.2 summarizes the four methods of power calculation resulting from the different combinations of design and analysis priors.

The standard method, as explored in Section 2.2.1, computes a conditional power and assumes a classical analysis at the end of the replication study. Both the hybrid and the Bayesian methods can be used to compute the predictive power but the hybrid method assumes a classical analysis while the Bayesian method will carry out a Bayesian analysis at the end of the replication study. Alternative names for predictive power in the literature are assurance (O'Hagan *et al.*, 2005), probability of study success (Wang *et al.*, 2013) and Bayesian predictive power (Spiegelhalter *et al.*, 1986). Finally, the conditional Bayesian method does not acknowledge the uncertainty of the original effect estimate $\hat{\theta}_o$ and assumes a Bayesian analysis at the end of the replication study.

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

Table 2.2: Table summarizing the methods of power calculation resulting from the different combinations of design and analysis priors.

Derivation of power formulas

In the following, the hybrid, Bayesian and conditional Bayesian power formulas are derived. In order to make the understanding of the derivations easier, we present here the key steps that will systematically be followed in each derivation. For the sake of consistency with the derivation order, we first cover the analysis phase and then the design phase.

As mentioned before, using a flat analysis prior is equivalent to performing a classical analysis. Classical significance S_ϵ^C is declared if $\overline{Y}_{n_r} > -z_\epsilon \sigma/\sqrt{n_r}$, as stated in equation (2.2). The power is then obtained by calculating the probability $\Pr\left(S_\epsilon^C\right)$ of this event happening. On the other hand, using a normal analysis prior introduces a new concept, the 'Bayesian significance' denoted by S_ϵ^B. Such power is then obtained by calculating the probability $\Pr\left(S_\epsilon^B\right)$ of the event S_ϵ^B.

Using a point design prior is equivalent to conditioning the power on the original effect estimate $\hat{\theta}_o$, which is assumed to be the true effect size. The resulting power is thus a conditional power. On the contrary, incorporating the uncertainty of $\hat{\theta}_o$ in the design results in a predictive power. In terms of calculation, the conditional power needs to be integrated with respect to the design prior in (2.4). Integration can be demanding and a more direct way is to use the predictive distribution of $\overline{Y}_{n_r}$,

$$\overline{Y}_{n_r} \sim \mathrm{N}\left(\hat{\theta}_o, \sigma^2\left(\frac{1}{n_o} + \frac{1}{n_r}\right)\right), \tag{2.5}$$

obtained by combining the prior (2.4) and the likelihood (2.1).

Hybrid method

The hybrid approach is used when we want to include the uncertainty of the original effect estimate $\hat{\theta}_o$ in the design, while performing a classical analysis at the end of the replication study. Hence this method is a hybrid of Bayesian and classical methods. As a classical analysis will be performed at the end of the replication study, we look at the probability of Classical significance,

$$\Pr\left(S_\epsilon^C\right) = \Pr\left(\overline{Y}_{n_r} > -\frac{1}{\sqrt{n_r}} z_\epsilon \sigma\right).$$

We incorporate the design prior by using the predictive distribution (2.5) of $\overline{Y}_{n_r}$,

$$\begin{aligned}
\Pr\left(S_\epsilon^C\right) &= 1 - \Phi\left[\frac{-z_\epsilon\sigma/\sqrt{n_r} - \mathrm{E}(\overline{Y}_{n_r})}{\sqrt{\mathrm{Var}(\overline{Y}_{n_r})}}\right] \\
&= \Phi\left[\frac{z_\epsilon\sigma/\sqrt{n_r} + \hat{\theta}_o}{\sigma\sqrt{1/n_o + 1/n_r}}\right] \\
&= \Phi\left[\sqrt{\frac{n_o}{n_o + n_r}}\left(\frac{\hat{\theta}_o\sqrt{n_r}}{\sigma} + z_\epsilon\right)\right]. \qquad (2.6)
\end{aligned}$$

Equation (2.6) specifies the predictive power of a replication study with n_r subjects assuming a classical analysis of the results. By fixing the hybrid power $\Pr\left(S_\epsilon^C\right)$ to the desired value, we can calculate the required sample size n_r in the replication study. Unlike the standard method, the hybrid method has no closed-form expression of the sample size n_r for a fixed hybrid power $\Pr\left(S_\epsilon^C\right)$. Applications of root-finding algorithms are required.

Bayesian method

Incorporating the prior (2.4) not only in the design but also in the analysis of the replication study gives rise to the Bayesian method, a new approach to power calculation. 'Bayesian significance' is denoted as

$$S_\epsilon^B = \Pr\left(\theta < 0 \,|\, \text{replication data}\right) < \epsilon$$

and is the predictive probability of obtaining a significant Bayesian result when testing the null hypothesis $\theta < 0$ against an alternative $\theta > 0$. Assuming a future parameter estimate $\overline{Y}_{n_r}$, the posterior distribution of θ is given by

$$\theta \,|\, \overline{Y}_{n_r} \sim \mathrm{N}\left(\frac{n_o\hat{\theta}_o + n_r\overline{Y}_{n_r}}{n_o + n_r}, \frac{\sigma^2}{n_o + n_r}\right). \qquad (2.7)$$

From (2.7) we can deduce that S_ϵ^B will occur when the parameter estimate $\overline{Y}_{n_r}$ obeys

$$\overline{Y}_{n_r} > \frac{-\sqrt{n_o + n_r}\, z_\epsilon \sigma - n_o\hat{\theta}_o}{n_r}. \qquad (2.8)$$

Derivation details are omitted here and can be found in Appendix A.1. The Bayesian power is

given by the probability of S_ϵ^B,

$$\Pr\left(S_\epsilon^B\right) = \Pr\left(\overline{Y}_{n_r} > \frac{-\sqrt{n_o+n_r}z_\epsilon\sigma - n_o\hat{\theta}_o}{n_r}\right).$$

We again incorporate the design prior by using the predictive distribution (2.5) of $\overline{Y}_{n_r}$,

$$\begin{aligned}
\Pr\left(S_\epsilon^B\right) &= 1 - \Phi\left[\frac{-\sqrt{n_o+n_r}z_\epsilon\sigma - n_o\hat{\theta}_o - n_r\,\mathrm{E}(\overline{Y}_{n_r})}{n_r\sqrt{\mathrm{Var}(\overline{Y}_{n_r})}}\right] \\
&= 1 - \Phi\left[\frac{-\sqrt{n_o+n_r}z_\epsilon\sigma - n_o\hat{\theta}_o - n_r\hat{\theta}_o}{n_r\sigma\sqrt{(n_o+n_r)/n_r n_o}}\right] \\
&= \Phi\left[\frac{\hat{\theta}_o\sqrt{n_o}\sqrt{n_o+n_r}}{\sigma\sqrt{n_r}} + \sqrt{\frac{n_o}{n_r}}z_\epsilon\right]. \qquad (2.9)
\end{aligned}$$

Equation (2.9) specifies the predictive power of a replication study with n_r subjects assuming a Bayesian analysis of the results. By fixing the Bayesian power $\Pr\left(S_\epsilon^B\right)$, we are able to calculate the required sample size n_r in the replication study with root-finding algorithms.

Conditional Bayesian method

The conditional Bayesian method assumes a Bayesian analysis at the end of the replication study but conditions the power on the original effect estimate $\hat{\theta}_o$, which is assumed to be the true effect size. This approach has not been described in Spiegelhalter *et al.* (2004). This is a new method we design and briefly present in order to have the fourth combination of design and analysis priors exposed in Table 2.2. The conditional Bayesian power is the probability of S_ϵ^B conditioned on $\theta = \hat{\theta}_o$,

$$\begin{aligned}
\Pr\left(S_\epsilon^B \,|\, \theta = \hat{\theta}_o\right) &= \Pr\left(\overline{Y}_{n_r} > \frac{-\sqrt{n_o+n_r}z_\epsilon\sigma - n_o\hat{\theta}_o}{n_r}\right) \\
&= 1 - \Phi\left[\frac{-\sqrt{n_o+n_r}z_\epsilon\sigma - n_o\hat{\theta}_o - n_r\,\mathrm{E}(\overline{Y}_{n_r})}{n_r\sqrt{\mathrm{Var}(\overline{Y}_{n_r})}}\right] \\
&= 1 - \Phi\left[\frac{-\sqrt{n_o+n_r}z_\epsilon\sigma - n_o\hat{\theta}_o - n_r\hat{\theta}_o}{n_r\sigma/\sqrt{n_r}}\right] \\
&= \Phi\left[\sqrt{\frac{n_o+n_r}{n_r}}z_\epsilon + \frac{\hat{\theta}_o(n_o+n_r)}{\sigma\sqrt{n_r}}\right]. \qquad (2.10)
\end{aligned}$$

Equation (2.10) is the conditional power of a replication study with n_r subjects assuming a Bayesian analysis of the results. By fixing the conditional Bayesian power $\Pr\left(S_\epsilon^B \,|\, \hat{\theta}_o\right)$, the required sample size in the replication study can be calculated with root-finding algorithms. As the uncertainty incorporation is the main goal of this thesis, the conditional Bayesian method will not be as thoroughly investigated as the other methods.

2.3 Results

In this section, we present our findings concerning the different methods of power calculation. We outline their properties, their differences, their common characteristics and examine the behavior of the power curves under special circumstances.

2.3.1 Alternative expressions based on the relative sample size c and the original p-value p_o

For a fixed significance level $\alpha = 2\epsilon$, the standard, the hybrid and the Bayesian power formulas can be rewritten as a function of the original test statistic $t_o = \hat{\theta}_o\sqrt{n_r}/\sigma$ and the relative sample size $c = n_r/n_o$ only. The test statistic t_o can easily be transformed into the two-sided p-value p_o with

$$p_o = 2\left(1 - \Phi[t_o]\right).$$

The standard power formula becomes

$$\Pr\left(S_\epsilon^C \mid \theta = \hat{\theta}_o\right) = \Phi\left[t_o\sqrt{c} + z_\epsilon\right], \tag{2.11}$$

the hybrid power formula becomes

$$\Pr\left(S_\epsilon^C\right) = \Phi\left[\sqrt{\frac{1}{c+1}}\left(t_o\sqrt{c} + z_\epsilon\right)\right], \tag{2.12}$$

and finally, the Bayesian power formula becomes

$$\Pr\left(S_\epsilon^B\right) = \Phi\left[t_o\sqrt{1 + \frac{1}{c}} + \sqrt{\frac{1}{c}}z_\epsilon\right]. \tag{2.13}$$

Similarly, it can be deduced that for a fixed significance level $\alpha = 2\epsilon$, the relative sample size c only depends on the original test statistic t_o and the power. This interesting property reduces the number of arguments needed in the power calculation. Based on the context, we will either use the initial or the alternative formulas in the following.

Illustration Let us consider two hypothetical studies, study A and study B. They have already been conducted and were declared significant at the 5% level with a p-value of 0.02. Suppose the two studies have different sample sizes, effect sizes and standard deviations, see Table 2.3. We now want to conduct two replication studies in order to confirm the findings of the two studies. For each replication study, we calculate the replication power as a function of the relative sample size c, as shown in Figure 2.1. We see that for each method, the same replication power is achieved given the same relative sample size c in both replication studies. This example illustrates the dependence of the power on the original p-value p_o and the relative sample size c only.

	A	B
n_o	26	240
$\hat{\theta}_o$	0.5	0.15
σ	1.1	1
p-value	0.02	0.02

Table 2.3: Original sample size, effect estimate, standard deviation and p-value of studies A and B.

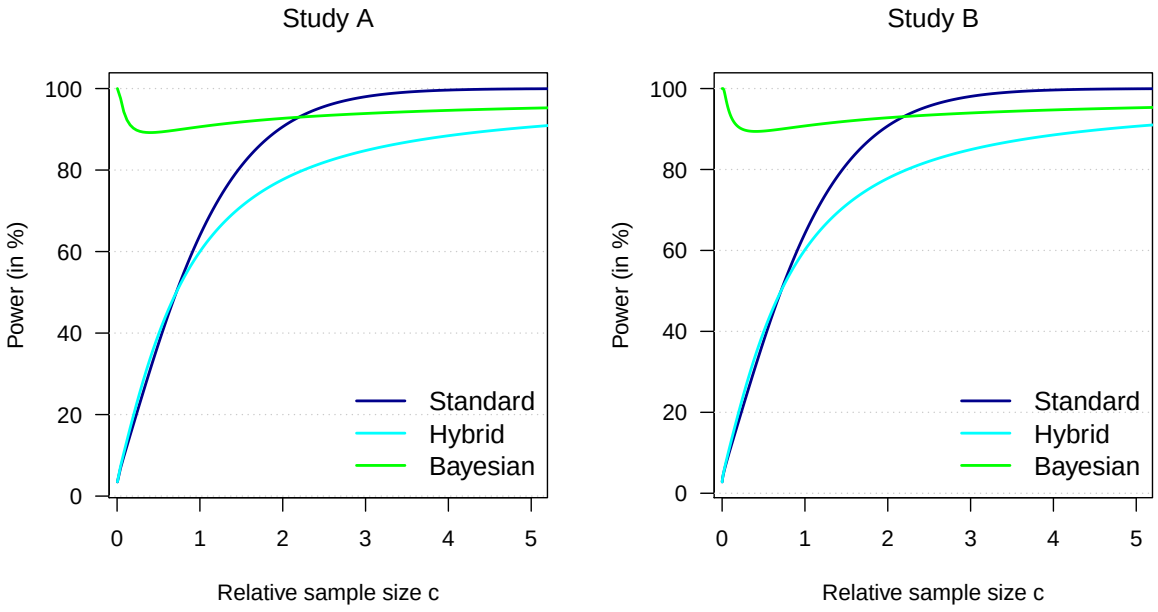


Figure 2.1: Replication power with the standard, the hybrid and the Bayesian methods as a function of the relative sample size c for studies A and B.

2.3.2 Manipulation of the original study

In this part, we investigate how the replication power changes in the hypothetical situation where we could manipulate the settings of the original study.

Replication power as a function of the original p-value p_o

The standard, the hybrid and the Bayesian formulas in (2.11), (2.12) and (2.13) imply that the power will increase as the original test statistic t_o increases. This entails that for a fixed relative sample size c, more convincing original studies (smaller p-value p_o) will lead to more powerful replication studies. Similarly, the required sample size to achieve a certain power will get smaller when the original p-value p_o decreases. This property is illustrated in Figures 2.2 and 2.3.

Illustration In Figure 2.2, the replication study is assumed to have the same size as the original study ($c = 1$). Less convincing original studies lead to a non-negligible lower power than more convincing original studies. Although we do not obtain the same power with the three methods, the observed trend is respected in all of them. Remarkably, an original finding with a p-value $p_o = 0.05$ will reach a replication power of only 50% with the standard and the hybrid methods if the same sample size is used as in the original study. The Bayesian method returns

the largest power for all shown original p-values p_o. In Figure 2.3, the relative sample size c to reach a power of 90% with the different methods is shown. As expected, original studies with larger p-values p_o need larger sample sizes than more convincing original studies. Remarkably, the difference between the standard and the hybrid sample sizes increases as p_o gets larger. Moreover, there is no replication sample size for small original p-value p_o with the Bayesian method. This feature is investigated later.

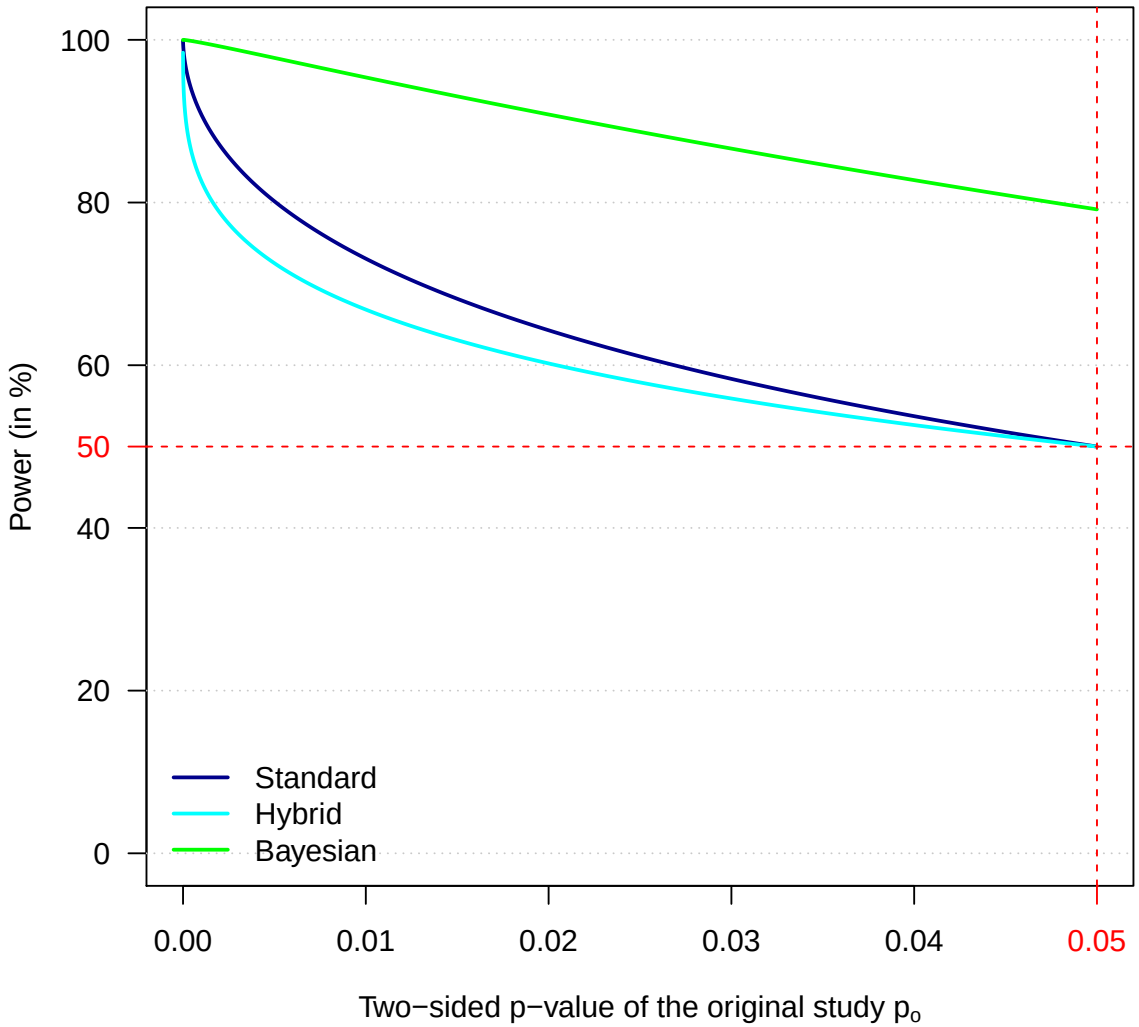


Figure 2.2: Power calculation with the standard, the hybrid and the Bayesian methods for a replication study with sample size equal to the original sample size ($c = 1$) as a function of the original p-value p_o at the traditional 5% level. The vertical red line indicates a p-value p_o of 0.05 and the horizontal red line a power of 50%.

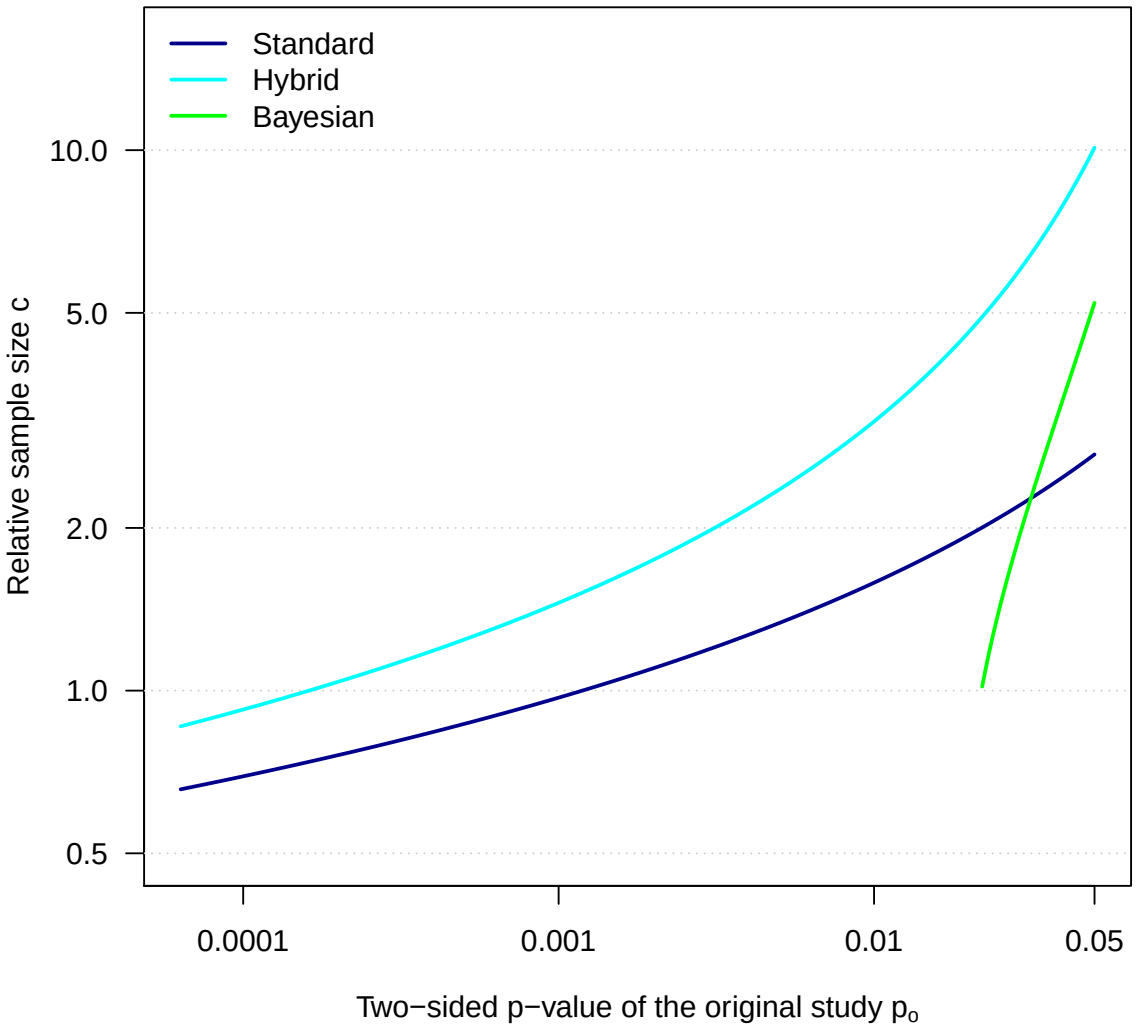


Figure 2.3: Relative sample size c to achieve a power of 90% as a function of the original p-value p_o with the standard, the hybrid and the Bayesian methods at the traditional 5% level.

Replication power as a function of the original sample size n_o

Here we investigate the power with the hybrid and the Bayesian methods in the hypothetical situation of an original study with infinitely large sample size. For this task, we use the initial formulas (2.6) and (2.9), where we let the original sample size n_o go to infinity.

For an infinitely large original sample size n_o, the hybrid power tends to the standard power,

$$\lim_{n_o \to +\infty} \Phi\left[\sqrt{\frac{n_o}{n_o + n_r}}\left(\frac{\hat{\theta}_o\sqrt{n_r}}{\sigma} + z_\epsilon\right)\right] = \Phi\left[\frac{\theta_o\sqrt{n_r}}{\sigma} + z_\epsilon\right].$$

This is reasonable since the prior (2.4) variance tends to zero as n_o tends to infinity and the normal prior becomes then a point prior at $\theta = \hat{\theta}_o$. However, the Bayesian power behaves differently.

The Bayesian replication power of an infinitely large original study is 100%, as outlined by

$$\lim_{n_o \to +\infty} \Phi \left[\frac{\hat{\theta}_o \sqrt{n_o} \sqrt{n_o + n_r}}{\sigma \sqrt{n_r}} + \sqrt{\frac{n_o}{n_r}} z_\epsilon \right] = 1 \,.$$

This is an intriguing property of the Bayesian power, as it means that a replication study will always reach a large Bayesian power if the original study is large enough.

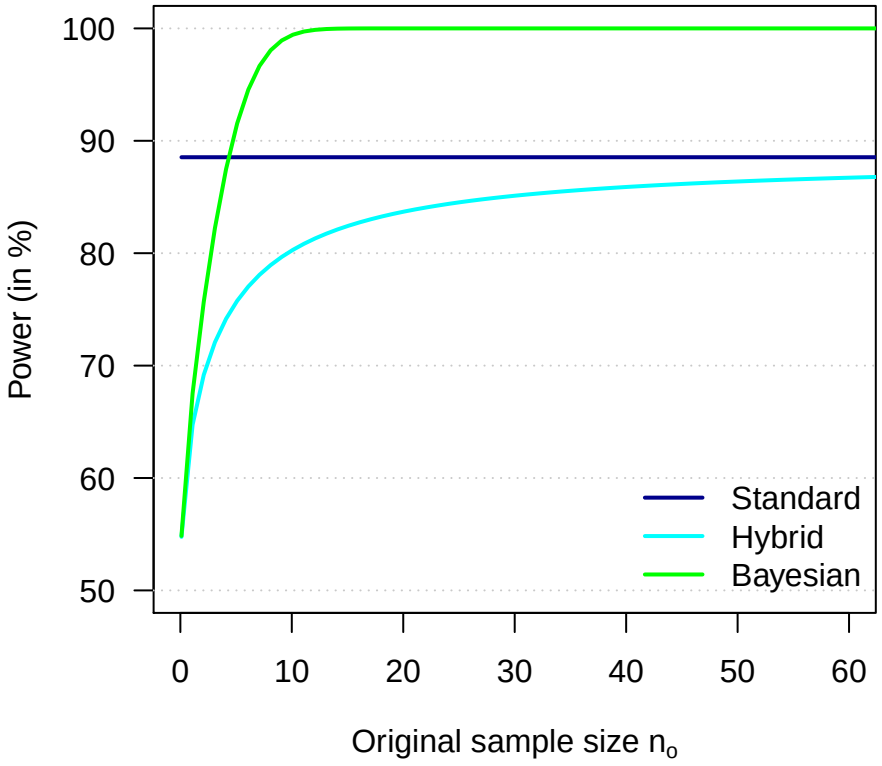


Figure 2.4: Power calculation with the standard, the hybrid and the Bayesian methods as a function of the original sample size n_o assuming $\hat{\theta}_o = 1$, $\sigma = 1$ and $n_r = 10$ at the 5% level.

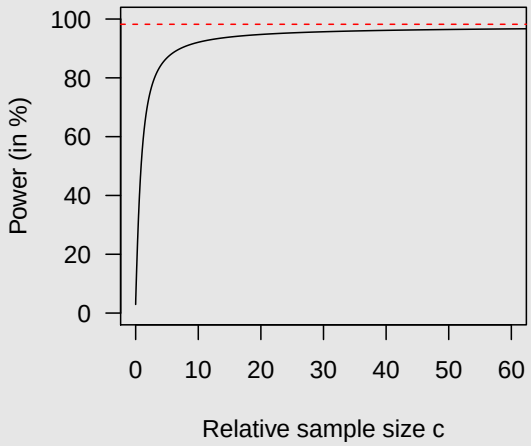
Illustration Figure 2.4 illustrates the behavior of the power with the hybrid and the Bayesian methods as a function of the original sample size n_o and compares it to the power with the standard method. Suppose we conducted an original study which detected an effect estimate $\hat{\theta}_o = 1$, with $\sigma = 1$. We now plan a replication study with a sample size n_r of ten. With the standard method, the power is 88.5% regardless of the original sample size. In contrast, the hybrid power is 54.8% if only one subject was included in the original experiment and increases until it meets the standard power for large values of n_o. The Bayesian power rapidly increases to 100% as n_o increases.

2.3.3 Manipulation of the replication study

Here we investigate the more interesting situation where the original study has already been conducted and we are planning the replication study. In order to choose the most adequate sample size for the replication study, we examine the hybrid and the Bayesian power formulas as a function of the replication sample size n_r or analogously of the relative sample size c. The important results are first briefly summarized and then derived in detail.

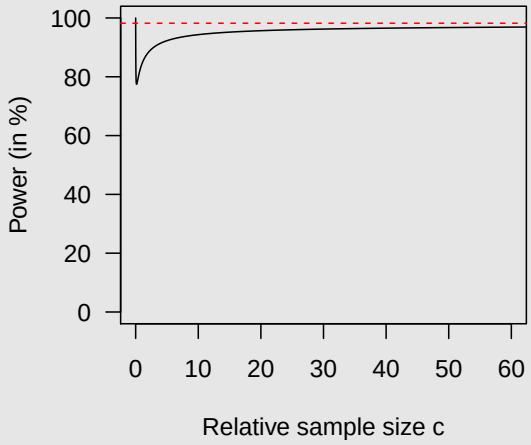
Hybrid power

- monotonically increasing
- limiting value: $1 - p_o/2$



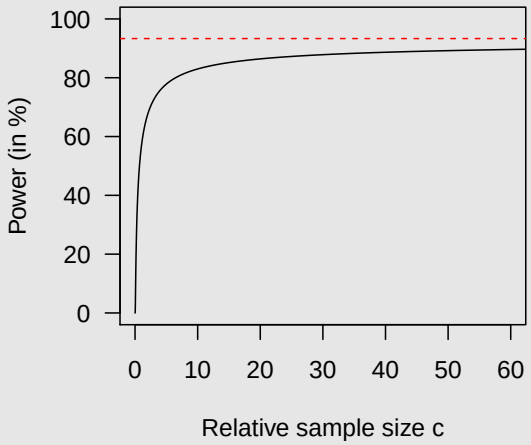
Bayesian power for significant original studies

- non-monotone
- power of 1 for extremely small relative sample size
- minimum at $\Phi\left[\sqrt{t_o^2 - z_\epsilon^2}\right]$
- limiting value: $1 - p_o/2$



Bayesian power for non-significant original studies

- monotonically increasing
- limiting value: $1 - p_o/2$



Hybrid power

As a first step, we compute the power of a replication study with no subject. It may seem meaningless to investigate this situation as it will never happen in practice. However, it will become an interesting property in the Bayesian method. The replication power of a study without subjects is found by replacing n_r with 0 in equation (2.6),

$$\Phi\left[\sqrt{\frac{n_o}{n_o+0}}\left(\frac{\hat{\theta}_o\sqrt{0}}{\sigma}+z_\epsilon\right)\right]=\Phi\left[z_\epsilon\right]=\epsilon\,,$$

and is the significance level divided by two. If we assume a 5% significance level, the replication power of a study without subjects is still 2.5%. Similarly, we want to know the hybrid power of a replication study with an infinitely large sample size,

$$\begin{aligned}\lim_{n_r\to+\infty}\Phi\left[\sqrt{\frac{n_o}{n_o+n_r}}\left(\frac{\hat{\theta}_o\sqrt{n_r}}{\sigma}+z_\epsilon\right)\right]&=\Phi\left[\lim_{n_r\to+\infty}\left(\sqrt{\frac{n_o}{n_o+n_r}}\left(\frac{\hat{\theta}_o\sqrt{n_r}}{\sigma}+z_\epsilon\right)\right)\right]\\&=\Phi\left[\lim_{n_r\to+\infty}\left(\frac{\hat{\theta}_o\sqrt{n_o}}{\sigma}\frac{\sqrt{n_r}}{\sqrt{n_o+n_r}}+\frac{\sqrt{n_o}z_\epsilon}{\sqrt{n_o+n_r}}\right)\right]\\&=\Phi\left[\frac{\hat{\theta}_o\sqrt{n_o}}{\sigma}\right]\,.\end{aligned}\tag{2.14}$$

As $\hat{\theta}_o\sqrt{n_o}/\sigma$ is the test statistic t_o of the original study, we have

$$\Phi\left[\frac{\hat{\theta}_o\sqrt{n_o}}{\sigma}\right]=\Phi\left[t_o\right]$$

$$= 1 - \Phi\left[-t_o\right]$$
$$= 1 - p_o/2\,.$$

The same result is obtained with the alternative formula of the hybrid power (2.12),

$$\lim_{c\to+\infty} \Phi\left[\sqrt{\frac{1}{c+1}}\left(t_o\sqrt{c} + z_\epsilon\right)\right] = \Phi[t_o]\,.$$

This result implies that by increasing the replication sample size, the hybrid power tends to $1 - p_o/2$, one minus the one-sided p-value of the original study. The more conclusive the original study, the larger power the replication study can achieve.

Illustration Figure 2.5 illustrates the dependence of the limiting hybrid power on the original one-sided p-value $p_o/2$. The hybrid power of three hypothetical studies with $p_o = 0.08$, 0.03 and 0.005 is plotted as a function of the relative sample size c. With a sufficiently large relative sample size c, each study reaches a power of $1 - p_o/2$. A larger relative sample size c is required in less convincing studies.

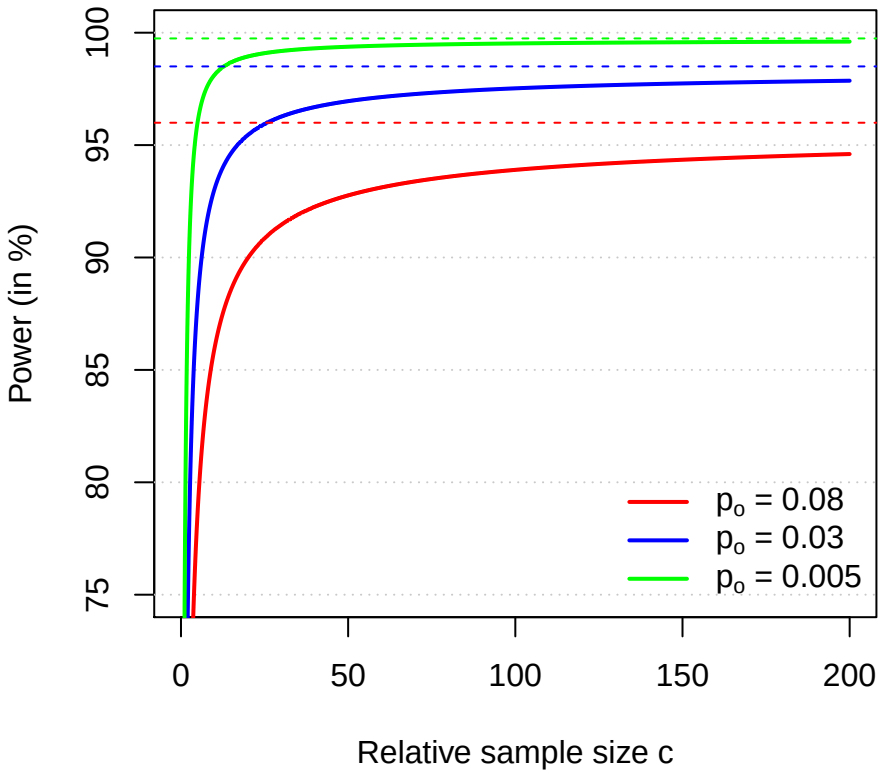


Figure 2.5: Hybrid power as a function of the relative sample size c for three hypothetical studies with $p_o = 0.08$, 0.03 and 0.005 respectively and at the 5% level. The red, blue and green horizontal dashed lines indicate $1 - p_o/2$, (1 - 0.04), (1 - 0.015) and (1 - 0.0025) respectively.

We know now the hybrid power of a study without subjects and of a study with an infinite number of subjects. However, we do not know the behavior of hybrid power between these two extremes. In the following, we prove the monotonicity of the hybrid power as a function of the relative sample size c. We first recall the formula of the hybrid power:

$$\Pr\left(S_\epsilon^C\right) = \Phi\left[\sqrt{\frac{n_o}{n_o+n_r}}\left(\frac{\hat{\theta}_o\sqrt{n_r}}{\sigma} + z_\epsilon\right)\right].$$

Since the function $\Phi[\cdot]$ is monotonically increasing, we only need to show that its argument is monotone in n_r. To do so, we take the first partial derivative with respect to n_r,

$$\frac{d}{dn_r}\left[\sqrt{\frac{n_o}{n_o+n_r}}\left(\frac{\hat{\theta}_o\sqrt{n_r}}{\sigma} + z_\epsilon\right)\right] = \frac{(n_o/(n_r+n_o))^{3/2}(\hat{\theta}_o n_o - \sqrt{n_r}\sigma z_\epsilon)}{2(\sqrt{n_r}n_o\sigma)}. \tag{2.15}$$

Throughout the thesis, we assume positive values for n_o and n_r. As the alternative hypothesis H_1 is $\theta = \hat{\theta}_o > 0$ and as there would not be much interest in replicating a study with a negative effect, we assume that $\hat{\theta}_o$ is also positive. Moreover, z_ϵ is always negative and is equal to zero in the extreme case of $\epsilon = 0.5$ meaning $\alpha = 1$. Hence equation (2.15) is always positive and the replication power with the hybrid method is monotonically increasing as a function of the replication sample size n_r. In other words, increasing the replication sample size cannot decrease the hybrid power. This last statement may seem trivial but its relevance will become clear when investigating the monotonicity of the Bayesian power curve.

Bayesian power

The same procedure is applied to the Bayesian power formula. We first look at the Bayesian power in the hypothetical case of a replication study without subjects. For simplicity, let us consider the alternative Bayesian formula (2.13) here and calculate its limit as the relative sample size c tends to zero,

$$\begin{aligned}
\lim_{c\to 0}\Phi\left[t_o\sqrt{1+\frac{1}{c}} + \sqrt{\frac{1}{c}}z_\epsilon\right] &= \Phi\left[\lim_{c\to 0}\left(t_o\sqrt{1+\frac{1}{c}} + \sqrt{\frac{1}{c}}z_\epsilon\right)\right] \\
&= \Phi\left[(t_o - |z_\epsilon|)\infty\right] \\
&= \begin{cases} 0 & \text{if} \quad t_o < |z_\epsilon| \\ 1 & \text{if} \quad t_o > |z_\epsilon| \\ 0.5 & \text{if} \quad t_o = |z_\epsilon| \end{cases}.
\end{aligned}$$

The replication power when the relative sample size c tends to zero is 0% for non-significant original studies, 100% for significant original studies and 50% for original studies with p-values equal to the significance level. This intriguing property highlights the relevance of studying the power of a replication study without subjects.

In a second step, we look at the Bayesian power of a study with an infinitely large replication sample size n_r,

$$\begin{aligned}\lim_{n_r\to+\infty}\Phi\left[\frac{\hat{\theta}_o\sqrt{n_o+n_r}\sqrt{n_o}}{\sigma\sqrt{n_r}}+\sqrt{\frac{n_o}{n_r}}z_\epsilon\right] &= \Phi\left[\lim_{n_r\to+\infty}\left(\frac{\hat{\theta}_o\sqrt{n_o+n_r}\sqrt{n_o}}{\sigma\sqrt{n_r}}+\sqrt{\frac{n_o}{n_r}}z_\epsilon\right)\right]\\ &= \Phi\left[\frac{\hat{\theta}_o\sqrt{n_o}}{\sigma}\right]\\ &= 1-p_o/2\,. \end{aligned} \tag{2.16}$$

The limiting power with the Bayesian method is the same as with the hybrid method. This property implies that irrespective of the magnitude of the replication sample size, certain levels of power cannot be reached with the hybrid and the Bayesian methods.

Illustration Figure 2.6 represents the Bayesian replication power of the same three original studies as in Figure 2.5.

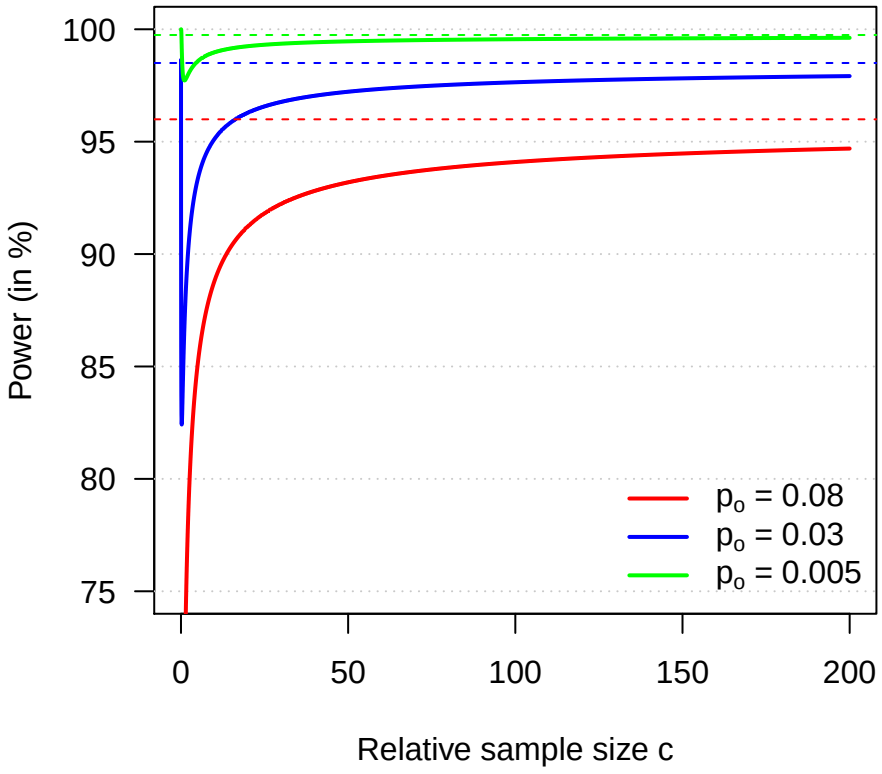


Figure 2.6: Bayesian power as a function of the relative sample size c for three hypothetical studies with p_o = 0.08, 0.03 and 0.005 respectively and at the 5% level. The red, blue and green horizontal dashed lines indicate (1 - 0.04), (1 - 0.015) and (1 - 0.0025) respectively.

We also notice in Figure 2.6 that the Bayesian replication power as a function of the relative

sample size c can be non-monotone. We investigate this property in the following. Derivation details can be found in Appendix A.2. By taking the first derivative of the Bayesian power formula, we learn that the Bayesian power is minimal when

$$n_r = n_o \left[\frac{\hat{\theta}_o^2 n_o}{\sigma^2 z_\epsilon^2} - 1 \right] \Leftrightarrow c = \frac{t_o^2}{z_\epsilon^2} - 1 \,.$$

The corresponding Bayesian power is given by

$$\Pr\left(S_\epsilon^B\right) = \Phi\left[\sqrt{t_o^2 - z_\epsilon^2}\right] .$$

We can again identify three cases. When $t_o > |z_\epsilon|$, the minimum Bayesian power is $\Phi\left[\sqrt{t_o^2 - z_\epsilon^2}\right]$ and corresponds to $c = t_o^2/z_\epsilon^2 - 1$. Remarkably, the minimum Bayesian power increases for increasing evidence in the original study (increasing original test statistic t_o). In contrast, when $t_o < |z_\epsilon|$, the relative sample size c corresponding to the minimum Bayesian power is negative. In this instance, the power curve is monotonically increasing for every $c \geq 0$. When $t_o = |z_\epsilon|$, the minimum power is 50% and corresponds to a relative sample size of zero.

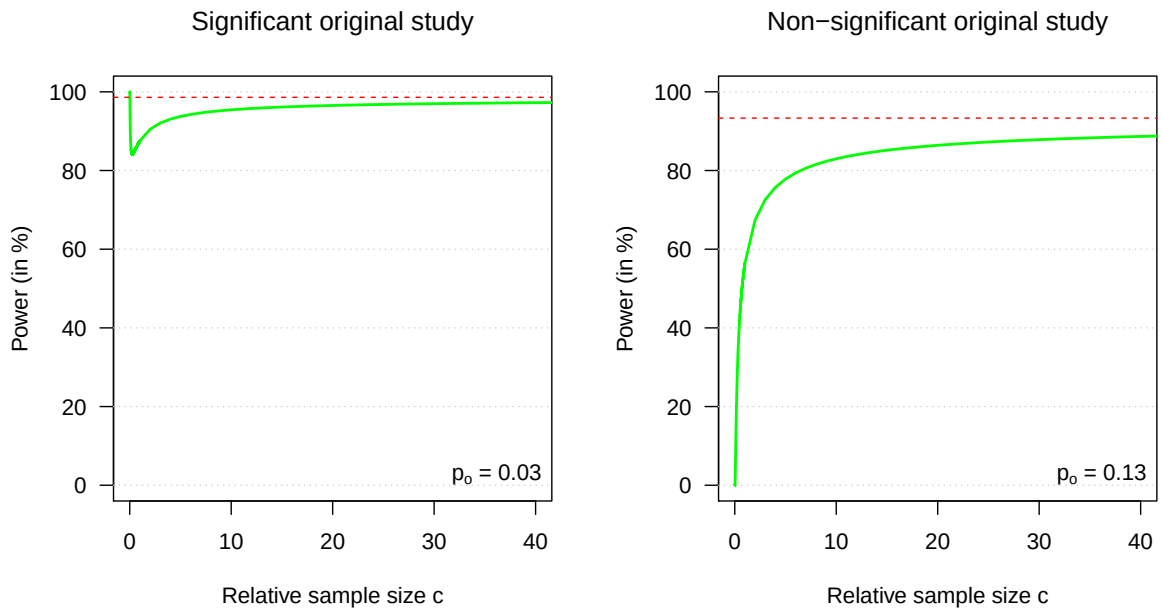


Figure 2.7: Bayesian power as a function of the relative sample size c for one significant and one non-significant original study at the traditional 5% level. In each study the horizontal red line indicates $1 - p_o/2$.

Altogether, our findings concerning the Bayesian power can be summarized in two main categories illustrated in Figure 2.7.

For a significant original study, the Bayesian replication power as a function of the relative sample size c starts at 100%, decreases to $\Phi\left[\sqrt{t_o^2 - z_\epsilon^2}\right]$ and increases up to $1 - p_o/2$. The range of values the Bayesian power can reach gets narrower with decreasing original p-values p_o. Notably, the replication power of a very convincing original study will always be 100% as the minimum and the limiting power converge to 100% when the original p-value p_o decreases.

We now understand why a power of 90% could not be reached with the Bayesian method when the evidence from the original study was very convincing in Figure 2.3. With such low original p-values p_o, any replication sample size results in a Bayesian power larger than 90%.

For a non-significant original study, the Bayesian replication power as a function of the relative sample size c begins at 0% and then monotonically increases up to the limiting value of $1 - p_o/2$. An original study with p-value p_o equal to the significance level α can be considered as a special case of the non-significant original studies. In such an instance, the power begins at 50% and then monotonically increases to $1 - p_o/2$.

Intersection of the power curves

It is interesting to calculate under which circumstances a given method gives a larger power than another for the same replication sample size. As a first step, we compare the standard and the hybrid methods. We want to know from which replication sample size the standard power is larger than the hybrid power. Once again, as the function $\Phi[\cdot]$ is monotonically increasing, we can focus on its argument in the calculation. We want to know when $\Pr\left(S_\epsilon^C \,|\, \theta = \hat{\theta}_o\right) \geq \Pr\left(S_\epsilon^C\right)$, which is equivalent to

$$\begin{aligned}
& \Phi\left[\frac{\hat{\theta}_o\sqrt{n_r}}{\sigma} + z_\epsilon\right] && \geq && \Phi\left[\sqrt{\frac{n_o}{n_o+n_r}}\left(\frac{\hat{\theta}_o\sqrt{n_r}}{\sigma} + z_\epsilon\right)\right] \\
\Leftrightarrow \quad & \frac{\hat{\theta}_o\sqrt{n_r}}{\sigma} + z_\epsilon && \geq && \sqrt{\frac{n_o}{n_o+n_r}}\left(\frac{\hat{\theta}_o\sqrt{n_r}}{\sigma} + z_\epsilon\right) \\
\Leftrightarrow \quad & n_r \leq 0 && \text{or} && n_r \geq \frac{\sigma^2 z_\epsilon^2}{\hat{\theta}_o^2} \\
\Leftrightarrow \quad & c \leq 0 && \text{or} && c \geq \frac{z_\epsilon^2}{t_o^2}\,.
\end{aligned}$$

These results show that for a replication sample size below 0 or above $\sigma^2 z_\epsilon^2/\hat{\theta}_o$, the standard power is larger than the hybrid power. However, as a negative sample size is not of interest, we focus on the second case. By plugging $\sigma^2 z_\epsilon^2/\hat{\theta}_o$ in the standard (or hybrid) power formula, we obtain the replication power corresponding to the intersection of both power curves,

$$\begin{aligned}
\Phi\left[\frac{\hat{\theta}_o\sqrt{n_r}}{\sigma} + z_\epsilon\right] &= \Phi\left[\frac{\hat{\theta}_o\sqrt{\sigma^2 z_\epsilon^2/\hat{\theta}_o^2}}{\sigma} + z_\epsilon\right] \\
&= \Phi\left[\frac{\hat{\theta}_o\left|\sigma z_\epsilon/\hat{\theta}_o\right|}{\sigma} + z_\epsilon\right] \\
&= \Phi\left[|z_\epsilon| + z_\epsilon\right] \\
&= 0.5\,. \qquad (2.17)
\end{aligned}$$

Because σ is always positive, z_ϵ is always negative and $\hat{\theta}_o$ is assumed to be positive, the formula can be simplified. This result formally shows that the standard method will return a larger power than the hybrid method for the same replication sample size provided the standard power

is larger than 50%. Both power curves cross at power of 50% for $n_r = \sigma^2 z_\epsilon^2/\hat{\theta}_o^2$ or equivalently $c = z_\epsilon^2/t_o^2$. This means that the larger the evidence in the original study, the earlier the two curves cross. If we now think in terms of sample size calculation, as the target power is usually above 50%, the hybrid method will in principle require a larger sample size in the replication study than the standard method.

The same procedure is carried out for the intersection of the Bayesian and the conditional Bayesian power curves. We want to know when $\Pr\left(S_\epsilon^B \,|\, \theta = \hat{\theta}_o\right) \geq \Pr\left(S_\epsilon^B\right)$, which is equivalent to

$$\begin{aligned}
& \Phi\left[\sqrt{\frac{n_o+n_r}{n_r}}z_\epsilon + \frac{\hat{\theta}_o(n_o+n_r)}{\sigma\sqrt{n_r}}\right] \geq \Phi\left[\frac{\hat{\theta}_o\sqrt{n_o}\sqrt{n_o+n_r}}{\sigma\sqrt{n_r}} + \sqrt{\frac{n_o}{n_r}}z_\epsilon\right] \\
\Leftrightarrow \quad & \sqrt{\frac{n_o+n_r}{n_r}}z_\epsilon + \frac{\hat{\theta}_o(n_o+n_r)}{\sigma\sqrt{n_r}} \geq \frac{\hat{\theta}_o\sqrt{n_o}\sqrt{n_o+n_r}}{\sigma\sqrt{n_r}} + \sqrt{\frac{n_o}{n_r}}z_\epsilon \\
\Leftrightarrow \quad & n_r \geq \frac{\sigma^2 z_\epsilon^2 - \hat{\theta}_o^2 n_o}{\hat{\theta}_o^2} \\
\Leftrightarrow \quad & c \geq \frac{z_\epsilon^2}{t_o^2} - 1\,. \qquad (2.18)
\end{aligned}$$

By plugging the replication sample size corresponding to (2.18) in the Bayesian power formula (2.13), we once again retrieve a power of 50%. This result states that the conditional Bayesian power is larger than the Bayesian power for the same relative sample size c provided that the Bayesian power is larger than 50%. However, in the case of a significant original study ($t_o > |z_\epsilon|$), the sample size c where the Bayesian and the conditional Bayesian power curves cross is negative and the power is above 50% for every replication sample size larger than 0. As a result, the conditional Bayesian power is always larger than the Bayesian power in this situation. Remarkably, the Bayesian and the conditional Bayesian power attain 50% one unit of relative sample size before the standard and the hybrid power.

Illustration Figure 2.8 presents the power of a replication study based on a non-significant original study ($p_o = 0.13$) as a function of the relative sample size c with the standard, hybrid, Bayesian and conditional Bayesian methods. As expected, the standard and the hybrid power curves and the Bayesian and the conditional Bayesian power curves cross at a power of 50% and at $c = z_\epsilon^2/t_o^2$ and $c = z_\epsilon^2/t_o^2 - 1$, respectively.

2.3.4 Predictive power at an interim analysis

In sequential trials, the data are regularly analyzed at interim and the study is stopped if sufficiently convincing results are obtained (Spiegelhalter *et al.*, 2004). In the previous sections, we were in the setting of two non-sequential trials. We had an original study and in order to assess its reliability we intended to conduct a replication study. The aim of this very short section is to show how the Bayesian power calculation for replication studies is related to power calculation at an interim analysis.

Suppose we are conducting a clinical trial and after collecting the data of a certain number of subjects we decide to perform an interim analysis. Let m denote the number of subjects at interim, n the number of additional subjects to be collected after the interim analysis, y_m

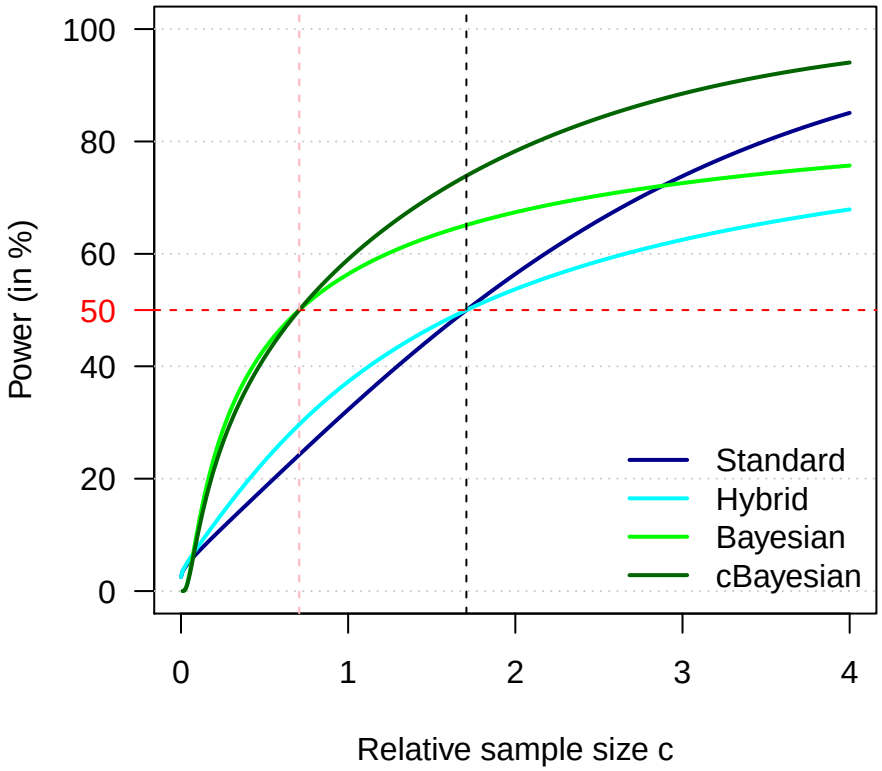


Figure 2.8: Power of a replication study based on a non-significant original study ($p_o = 0.13$) as a function of the relative sample size c with the four methods and at the traditional 5% level. The horizontal red line indicates a power of 50% and the two vertical lines indicate $c = z_\epsilon^2/t_o^2$ (black) and $c = z_\epsilon^2/t_o^2 - 1$ (pink).

the effect size at interim, σ the common standard deviation of one observation and θ the true unknown effect size. We want to know the power of the study, given the data so far. In order to calculate the predictive power, we need to select a prior distribution for θ, independent of the data. When using an uninformative prior, the predictive power is

$$\Pr\left(S_\epsilon^C \,|\, y_m\right) = \Phi\left[\frac{\sqrt{m+n}}{\sqrt{n}}\frac{\sqrt{m}y_m}{\sigma} + \sqrt{\frac{m}{n}}z_\epsilon\right]. \tag{2.19}$$

This formula is the classical predictive power in Spiegelhalter *et al.* (2004). Remarkably, if we assume that the current sample size m is the sample size of the original study n_o, the number of additional subjects n is the sample size of the replication study n_r and the effect estimate at interim y_m is the original effect estimate $\hat{\theta}_o$, the predictive power formula at interim (2.19) is identical to the Bayesian power formula (2.9). This last statement indicates that the Bayesian method does not consider the original and replication study as two independent studies, but as one pooled study including an interim analysis.

2.4 Application

In the following, we illustrate the methods discussed in the previous sections using real data from a large replication project.

The Open Science Collaboration (2015) conducted a large-scale, multi-year project on the replicability of psychological science. They replicated 100 experimental and correlational studies published in 2008 in three major psychological journals. Replication teams were formed and matched with studies according to their interests, resources and expertise. Each team conducted their study, analyzed their data and wrote a summary report. The results confirm the concerns of the replication crisis (Ioannidis, 2005). The replication effects were half the magnitude of the original effects. Moreover, only 36% of the replication studies were statistically significant against 97% of the original studies. For 73 studies, it was possible to transform the effect sizes to the correlation scale, forming the so-called Meta-Analytic (MA) subset (Johnson *et al.*, 2017). After application of Fisher's z-transformation $\theta = \tanh^{-1}(r)$ to the estimated correlation coefficient $\hat{r}$, a normal assumption is justified and the standard error is a function of the nominal study sample size n only, $\text{se}(\theta) = 1/\sqrt{n-3}$. Because the standard errors of the 27 remaining studies are not available, we do not include them in this work. The effective sample sizes $n-3$ are used in our calculation and we computed the two-sided p-values p_o and p_r taking advantage of the normality of the test statistics $t_o = \hat{\theta}_o / \text{se}(\hat{\theta}_o)$ and $t_r = \hat{\theta}_r / \text{se}(\hat{\theta}_r)$, respectively. However, the p-values reported in the original paper by the Open Science Collaboration (2015) are one-sided. Replication reports, as well as data, were made available on the Open Science Framework (`http://osf.io/ezcuj`). In this application, we focus on the dataset called `final` which can be extracted from the `MASTER` file.

2.4.1 Exploratory Data Analysis

First, we briefly present the data and results of the studies of the MA subset. The original and replication effect sizes are compared, as well as the original and replication sample sizes.

Figure 2.9 is a scatterplot of the original and replication effect estimates. It shows that most of the original studies were significant (82%), whereas this is the case for less than one-third of the replication studies. The difference between these percentages and the percentages in the paper comes from the fact that we used two-sided instead of one-sided p-values. Although these results are of great interest and we could discuss different ways of assessing if the replication was a success, this is not the aim of this thesis. We focus on an earlier phase of the process, namely the determination of the replication sample size.

Figure 2.10 shows the original sample sizes versus the replication sample sizes of the studies of the MA subset. No significant trend can be detected: the replication sample sizes are sometimes larger (46 studies), sometimes smaller (18 studies) and sometimes the same (9 studies) as the original sample sizes.

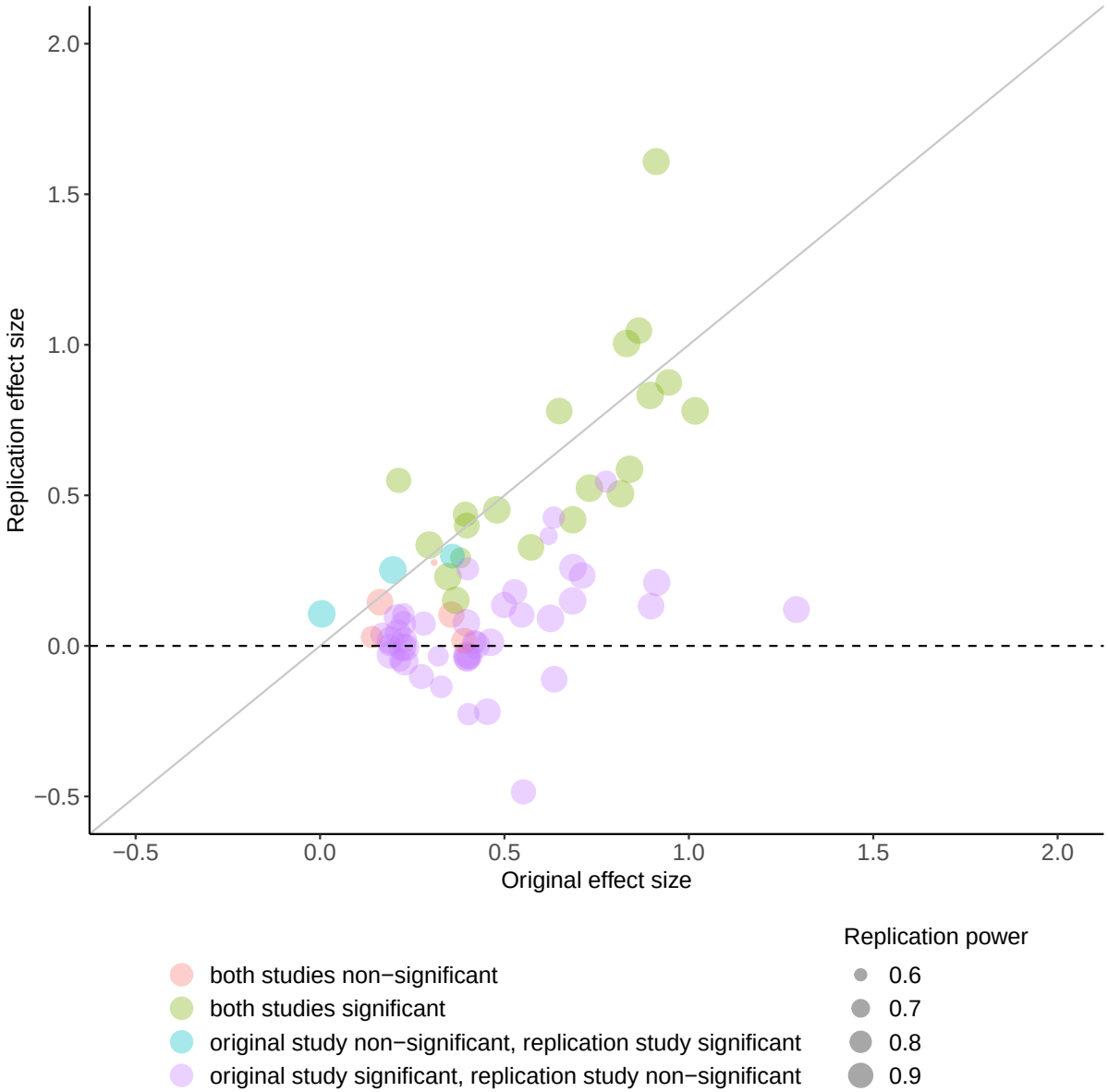


Figure 2.9: Original study effect estimate versus replication study effect estimate (Fisher z-transformed correlation coefficient). Each circle represents one study and the circle size indicates the power at the traditional 5% level. The diagonal line indicates a replication effect estimate equal to the original effect estimate. The horizontal dashed line shows a replication effect estimate of 0. The colors represent the significance of the original and replication studies.

2.4.2 Power calculation

We now apply the different methods of power calculation to the studies of the MA subset. We assume a 5% significance level and a one-sample design. In a first part, we compare the power calculated by the replication teams, which we call the nominal power, with the power calculated by the standard method. We then compute the power with the hybrid, Bayesian and conditional Bayesian methods and compare it to the standard power. In a third part, we investigate the behavior of the different power curves as a function of the relative sample size c for four selected studies.

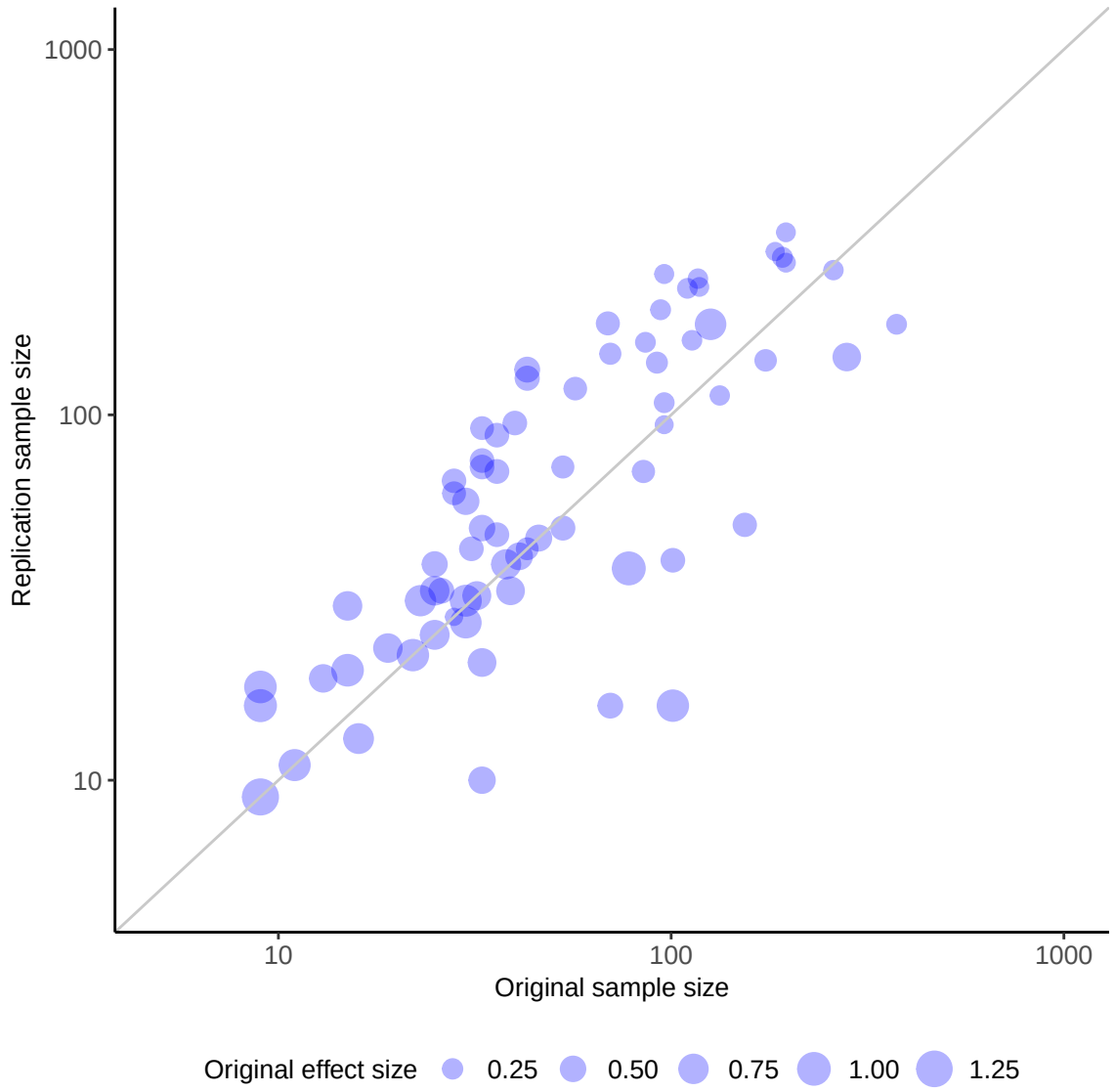


Figure 2.10: Original sample size versus replication sample size. Each circle represents a study and the circle size is proportional to the original effect estimate. The diagonal line indicates a replication sample size equal to the original sample size.

Nominal versus standard power

Each replication team calculated and reported the power of their replication study using the G^* Power software (Erdfelder *et al.*, 1996). The primary original effect estimates were used in their calculation, including Cohen's d, Cohen's f and η^2 for example. In contrast, we used the transformed correlation coefficient $\hat{\theta}_o$ in our power calculation. Figure 2.11 shows the nominal power and the power with the standard method of the 73 studies. Although they are supposed to correspond for every particular study, the standard power tends to be smaller than the nominal power. Some studies show a considerable discrepancy between the nominal and the standard power. Table 2.4 presents the eight studies where this discrepancy exceeds 20%.

In order to investigate these studies in greater detail, we examined the replication reports. For each study, we identified the primary original effect estimate and transformed it using Fisher's

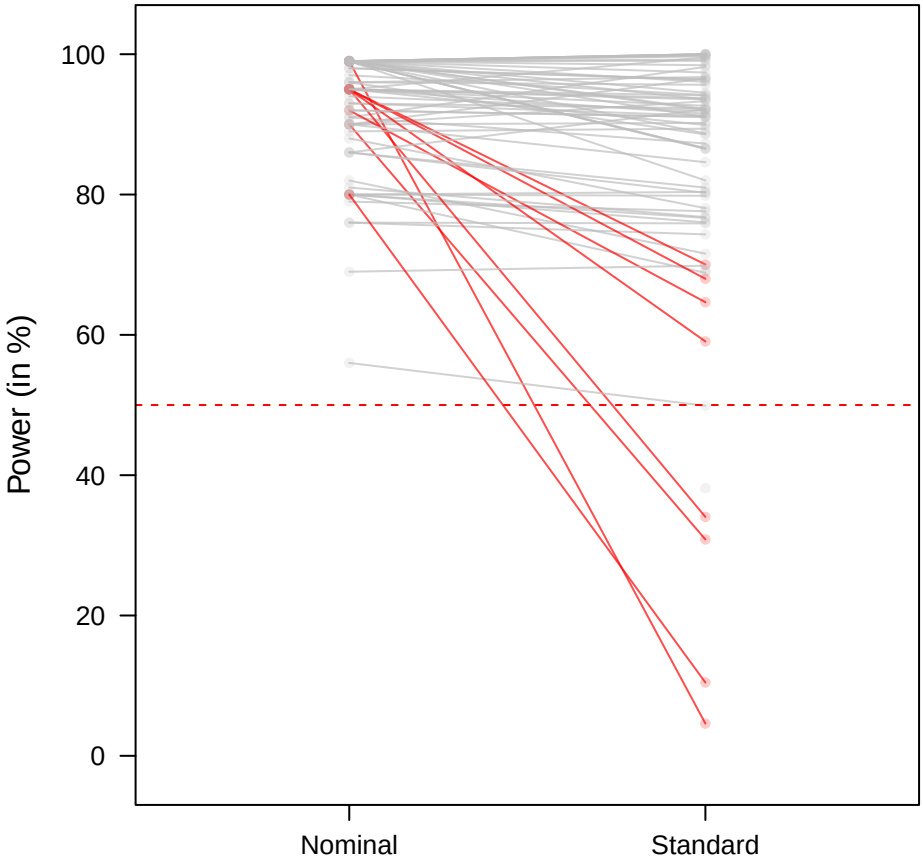


Figure 2.11: Nominal power and power with the standard method for the 73 studies of the MA subset. Each circle represents a study and the lines link the same studies. Studies where the discrepancy in power exceeds 20% are in red. The horizontal red line indicates a power of 50%.

Study	Nominal power	Standard power	Power difference	n_r	$\hat{\theta}_o$
3	95	70	25	33	0.454
20	92	65	27	108	0.228
26	95	34	61	94	0.162
52	95	59	36	113	0.209
56	95	68	27	40	0.399
89	80	10	70	28	0.141
115	90	31	59	10	0.552
135	99	5	94	3513.1	0.005

Table 2.4: Study number, nominal power, standard power, power difference, replication sample size and original effect estimate of the studies with the largest discrepancy in power.

z-transformation. For five studies, this recalculated original effect estimate corresponds to the effect estimate reported in the `final` dataset. Hence the difference in power for those five studies is not a consequence of an inaccurate conversion of effect estimates. For the remaining studies, we compute the power with the recalculated effect estimate. The newly calculated

power corresponds to the nominal power for two studies. For the last study, however, the power still does not correspond even after using the recalculated original effect estimate. Figure 2.12 recapitulates the procedure.

In summary, we could find an explanation of the large discrepancy for only two of the studies. For the other six studies, the reason for this large difference is still unclear and may come from the use of the G^* Power software.

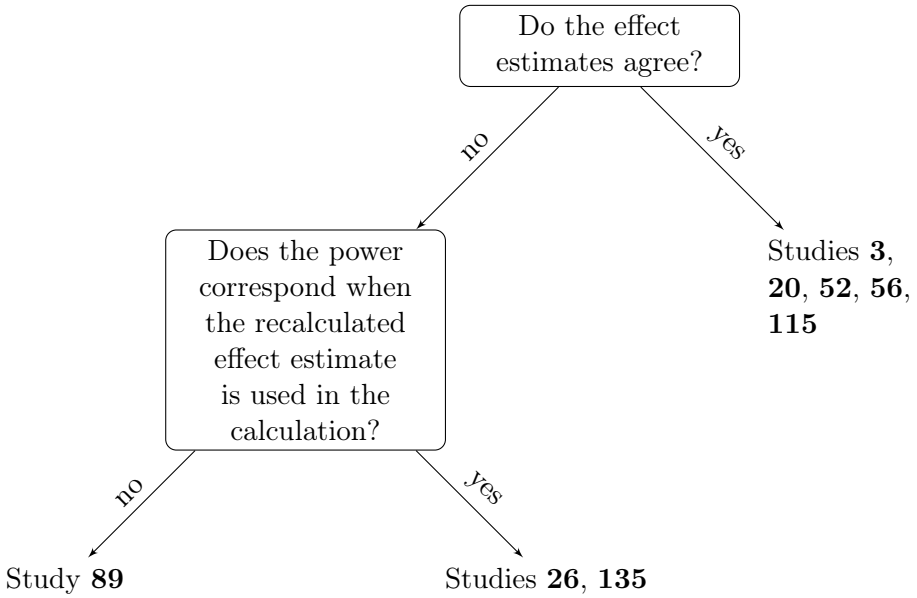


Figure 2.12: Investigation of the discrepancy between the nominal and the standard power.

Power calculation with the four methods

The four methods of power calculation are now used to obtain the power of the 73 replication studies belonging to the MA subset, see Figure 2.13, using the sample size that was chosen by the replication teams. As expected, the replication studies reach a larger power with the standard than with the hybrid method when the power is larger than 50%. This means that a larger number of subjects are required to reach the same level of power if the uncertainty of $\hat{\theta}_o$ is incorporated in the design. The power with the Bayesian method is in general larger than with the standard and the hybrid methods. This is not surprising as the Bayesian replication power of significant original studies is ensured to be between 50% (for $p_o = 0.05$) and 100%. The replication studies reach a larger power with the conditional Bayesian method than with the Bayesian method when the power is larger than 50%. The conditional Bayesian method gives rise to an extremely large median power (99.9%).

Power calculation as a function of the relative sample size c

We now select four studies and pretend the respective replication studies have not been conducted yet. Information about these studies can be found in Table 2.5. For each study, we compute the replication power as a function of the relative sample size c, see Figure 2.14. A 5% significance level is considered.

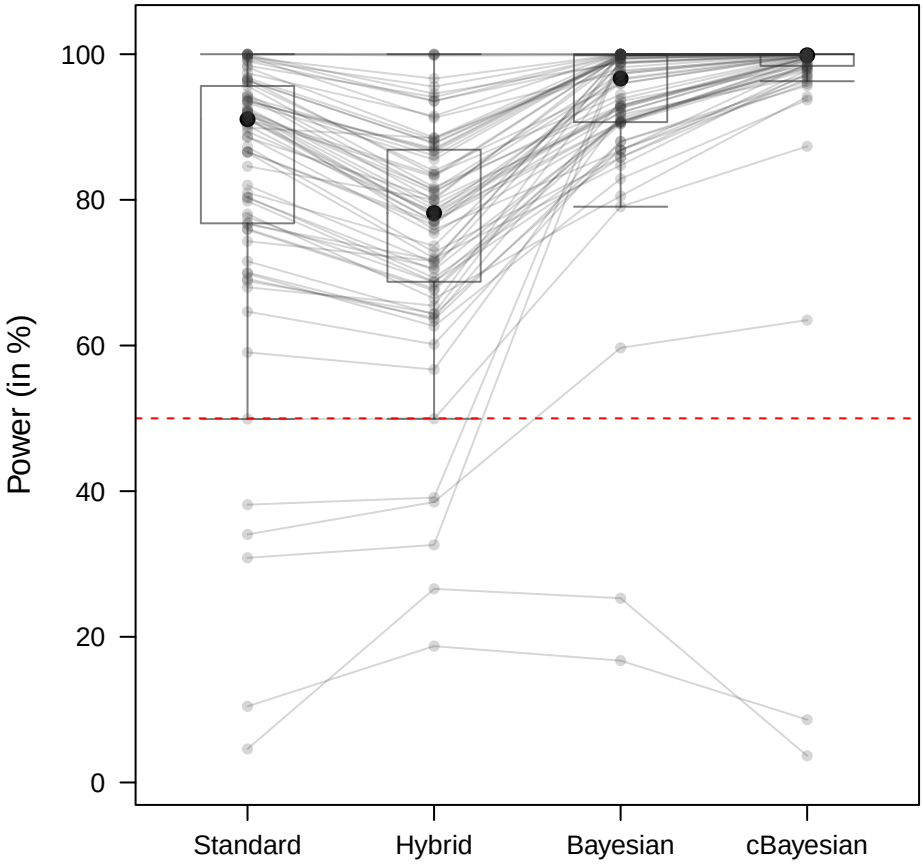


Figure 2.13: Replication power with the standard, the hybrid, the Bayesian and the conditional Bayesian methods for the 73 studies of the MA subset. Each circle represents a study and the lines link the same studies. The horizontal red line indicates a power of 50%.

	Study 24	Study 106	Study 82	Study 26
n_o	154	36	43	96
$\hat{\theta}_o$	0.38	0.40	0.31	0.16
σ	1.01	1.04	1.04	1.02
p_o	< 0.0001	0.021	0.051	0.117

Table 2.5: Description of study 24, 106, 82 and 26.

As a general observation, the standard and the hybrid power curves always cross at a power of 50%, as well as the Bayesian and the conditional Bayesian power curves for non-significant original studies. There is always a difference of one unit between the two intersections. Moreover, the plots make clear that the power for large relative sample size c tends to 100% with the standard method while it is bounded to a lower limit with the hybrid and the Bayesian methods. In the following, we describe the results for each study separately.

Study 24 has a really small p-value p_o. With the standard and the hybrid methods, the replication power rapidly reaches 100% with a small relative sample size c. The power with the Bayesian and the conditional Bayesian methods is very close to 100% for all relative sample sizes.

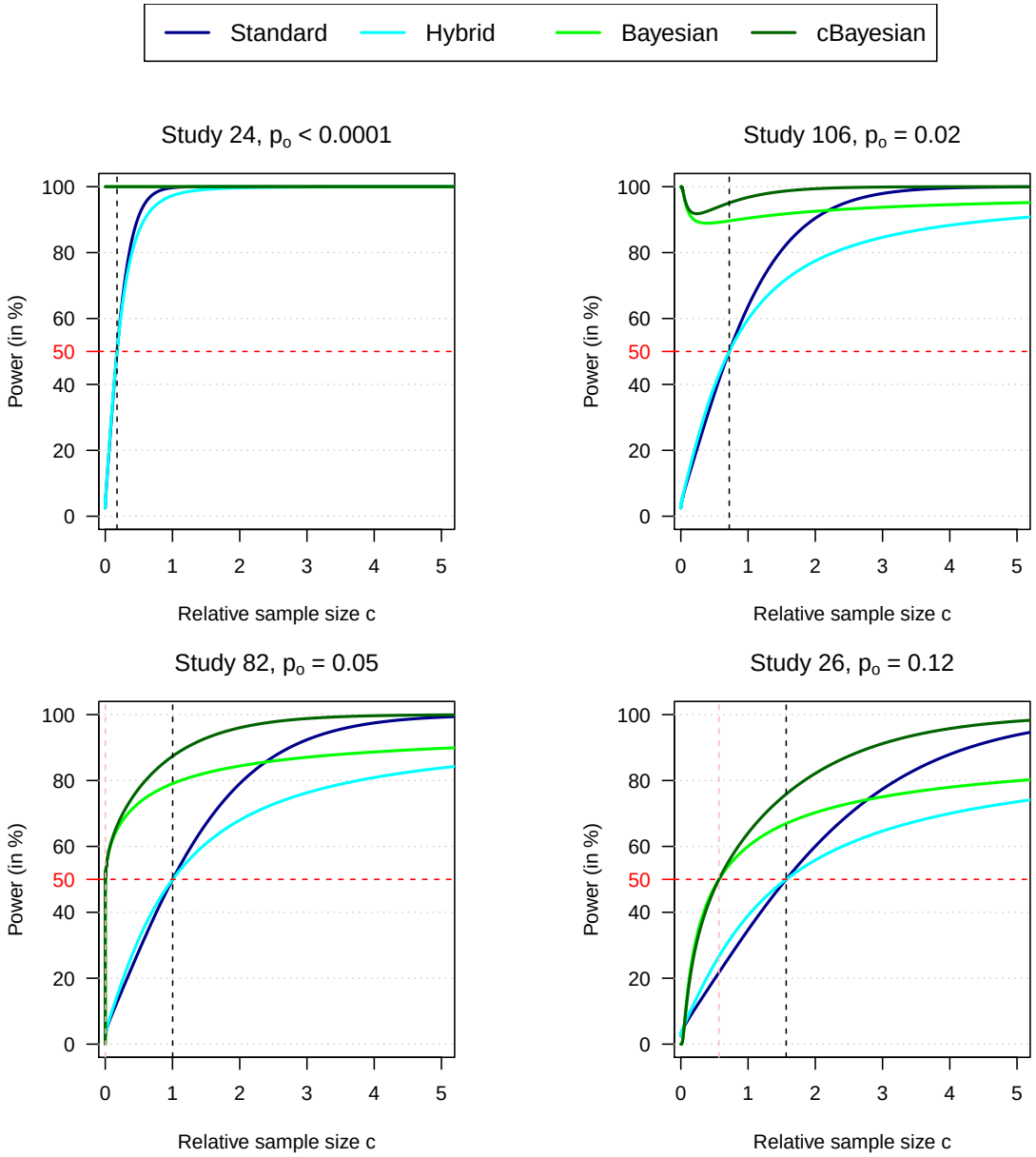


Figure 2.14: Replication power as a function of the relative sample size c for four studies of the MA subset at the 5% level. The vertical lines are plotted at the intersection of the standard and the hybrid power curves (black) and at the intersection of the Bayesian and conditional Bayesian power curves (pink). The horizontal red line indicates a power of 50%.

Study 106 is significant but has a larger p-value p_o than study 24. The power curves behave in a similar way as for the latter study with the difference that the standard and the hybrid power curves are flatter and the range of the Bayesian and the conditional Bayesian power is now larger. This means that more subjects are needed as compared to study 24 to reach the same level of standard or hybrid power and that the power with the Bayesian and the conditional Bayesian methods is not guaranteed to be 100%. However, it is still large for all values of the relative sample size c.

Study 82 is a non-significant study with p-value p_o close to 0.05. Regarding the power with the standard and the hybrid methods, more subjects are necessary to reach the same level of

power than in study 106. Remarkably, we observe a considerable change in behavior for the Bayesian and the conditional Bayesian power curves. The replication study can now reach any level of Bayesian and conditional Bayesian power.

Study 26 is non-significant. The situation is very similar to study 82 with the difference that all the curves are now flatter, meaning that more subjects are required to reach the same level of power than in study 82.

These results can be generalized to other studies with similar original p-values p_o as these four studies.

Chapter 3

Power for Replication Success

Statistical significance and p-values seem to be well-apprehended concepts but are in reality widely misunderstood, misinterpreted and misused (Cohen, 1994; Greenland *et al.*, 2016). In 2016, the American Statistical Association issued a Statement encouraging researchers to steer research into a 'post $p < 0.05$' era (Wasserstein and Lazar, 2016). Some authors suggested to lower the threshold for significance to $p = 0.005$ for claims of new discoveries (Johnson, 2013; Benjamin *et al.*, 2017; Ioannidis, 2018). Held (2019a) provides an additional argument for this new threshold with the $p = 0.0056$ threshold for intrinsic credibility. Intrinsic credibility is a concept proposed by Matthews (2018) in order to assess the credibility of 'out of the blue' findings without any prior support. Matthews found that the threshold for $\alpha = 0.05$ corresponds to the conventional p-value being lower than 0.01266. However, Matthews does not take all the uncertainty into account in his calculation, whereas Held does. In the following, we use the terminology Held's and Matthews' threshold to refer to $p = 0.0056$ and $p = 0.01266$, respectively.

Using standard significance of the replication study to assess replication success has also been questioned and has been shown to easily lead to conclusions opposite to what the evidence warrants (Simonsohn, 2015). As a result and in order to rectify the lack of a unified definition of replicability (Goodman *et al.*, 2016), new standards for evaluating replication success are emerging. In this chapter, we focus on a reverse-Bayes approach proposed by Held (2019b), which combines the Analysis of Credibility (Matthews, 2018) and the Box (1980) prior criticism approach to give rise to a new quantitative measure of replication success, the sceptical p-value p_S. While the main task of Chapter 2 was to design *significant* replication studies, in the following we aim at designing *successful* replication studies. Similarly to significance at a pre-specified level α which was equivalent to $p \leq \alpha$, replication success at level α is equivalent to $p_S \leq \alpha$. The computation of the power or the required sample size to achieve replication success is challenging and no closed form expression exists.

3.1 Theory

In this section, we briefly present the formulas and properties of the sceptical p-value p_S. We refer to Held (2019b) for the detailed derivations. In the following, let us assume that $\hat{\theta}_o$ and $\hat{\theta}_r$ are the effect estimates of the original and replication study, respectively, with the corresponding variance σ_o^2 and σ_r^2. Let $c = \sigma_o^2/\sigma_r^2$ denote the ratio of these variances and $t_o = \hat{\theta}_o/\sigma_o$ and

$t_r = \hat{\theta}_r/\sigma_r$ the test statistics of the original and replication study, respectively.

The sceptical p-value p_S is defined as

$$p_S = 2\left[1 - \Phi(z_S)\right] ,$$

with

$$z_S^2 = \begin{cases} t_H^2/2 & \text{for } c = 1 \text{ and} \\ \frac{1}{c-1}\left\{\sqrt{t_A^2\left[t_A^2 + (c-1)t_H^2\right]} - t_A^2\right\} & \text{for } c \neq 1\,. \end{cases} \tag{3.1}$$

In equation (3.1), $t_A^2 = (t_o^2 + t_r^2)/2$ is the arithmetic and $t_H^2 = 2/(1/t_o^2 + 1/t_r^2)$ the harmonic mean of the squared test statistics t_o^2 and t_r^2.

A central requirement of the sceptical p-value p_S is that $z_S^2 < \min\{t_o^2, t_r^2\}$. Hence the sceptical p-value p_S is always larger than the original and replication p-values p_o and p_r. Moreover, the sceptical p-value p_S takes into account the results from both the original and the replication study.

3.2 Results

In this section, we present the main findings about the power for replication success. While in Chapter 2 we could support our findings with formulas, we mostly base our reasoning on observations here, as no closed form expression exists. Just as the uncertainty of $\hat{\theta}_o$ can be ignored or taken into account in the power calculation for significance, we distinguish here between conditional and predictive power for replication success. The conditional power for replication success uses a point prior at $\theta = \hat{\theta}_o$ and thus does not take the uncertainty of $\hat{\theta}_o$ into account. In contrast, the predictive power for replication success uses a normal prior $\theta \sim \mathrm{N}(\hat{\theta}_o, \sigma_o^2)$ and hence acknowledges the uncertainty surrounding the original effect estimate $\hat{\theta}_o$. Suppose n_o and n_r are the sample sizes of the original and replication studies, respectively, so $\sigma_o^2 = \sigma^2/n_o$ and $\sigma_r^2 = \sigma^2/n_r$ where σ is the common standard deviation of one observation. Hence $c = \sigma_o^2/\sigma_r^2$ becomes $c = n_r/n_o$, the relative sample size.

3.2.1 Dependence on the original p-value p_o and the relative sample size c

For a pre-specified significance level, the power for replication success only depends on the original p-value p_o (or equivalently the original test statistic t_o) and the relative sample size c. Similarly, for a pre-specified significance level, the sample size for replication success only depends on the power and the p-value p_o of the original study. Derivation details are omitted here.

Illustration We reconsider the two studies from Table 2.3 and calculate the conditional and the predictive power for replication success as a function of the relative sample size c for both studies, see Figure 3.1. We see that despite having different effect sizes and standard errors, the two replication studies reach the same conditional and predictive power for the same relative sample size c.

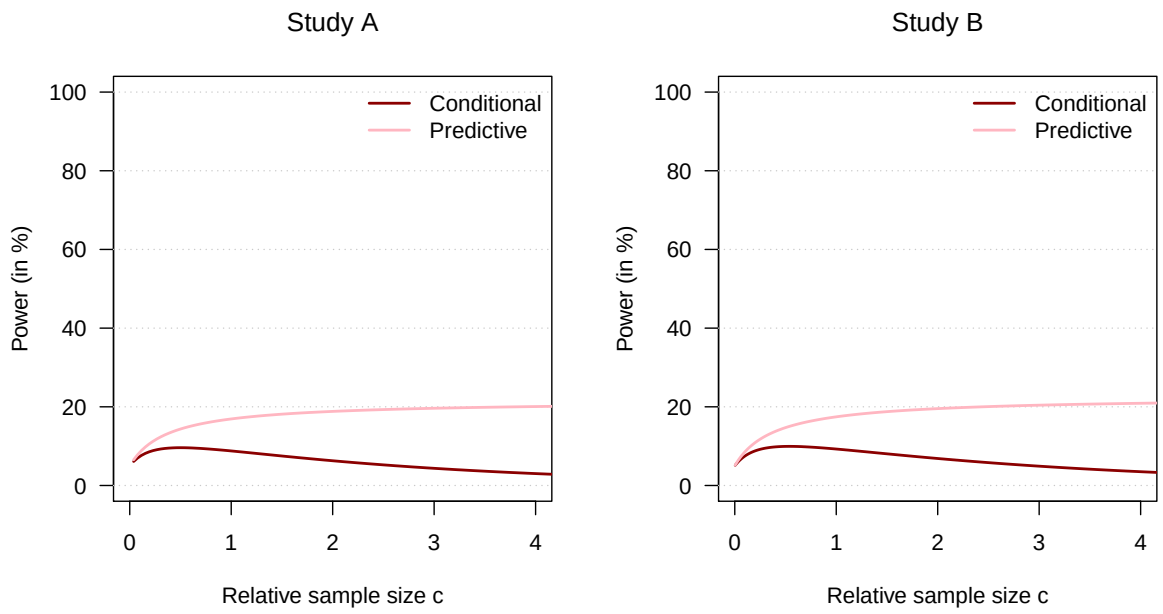


Figure 3.1: Conditional and predictive power for replication success as a function of the relative sample size c for studies A and B from Table 2.3 at the 5% level.

3.2.2 Power for replication success as a function of the original p-value p_o

We calculate the conditional and the predictive power for replication success as a function of the original p-value p_o for a replication study of the same size as the original study ($c = 1$) and compare it to the power for significance, see Figure 3.2. The Bayesian and the conditional Bayesian methods are not considered here. The sceptical p-value p_S is always larger than the original p-value p_o. It thus makes sense that if there is strong evidence in the original study (p_o very small), the power for replication success for a fixed c will be larger than if p_o is relatively large. The conditional and the predictive power for replication success rapidly drops to 0% as the original p-value p_o increases. While an original p-value p_o of 0.05 corresponds to a replication power of 50% with the standard and the hybrid methods, it corresponds to a conditional and predictive power for replication success of 0%. This property makes sense since the significance of the original study is a sine qua non condition to achieve replication success. A study will reach a conditional and predictive power for replication success of 50% if the original p-value p_o is 0.0056, Held's threshold for intrinsic credibility.

Figure 3.3 illustrates the difficulty to reach replication success at the pre-specified 90% power. Remarkably, while it is always possible to reach the pre-specified power with the standard and the hybrid methods with a sufficiently large relative sample size c, some original studies do not allow replication success regardless of the relative sample size c. For example, for intrinsically credible original studies ($p_o = 0.0056$), a relative sample size of slightly above four is required in order to achieve replication success under the point prior, but replication success cannot be achieved under the normal prior.

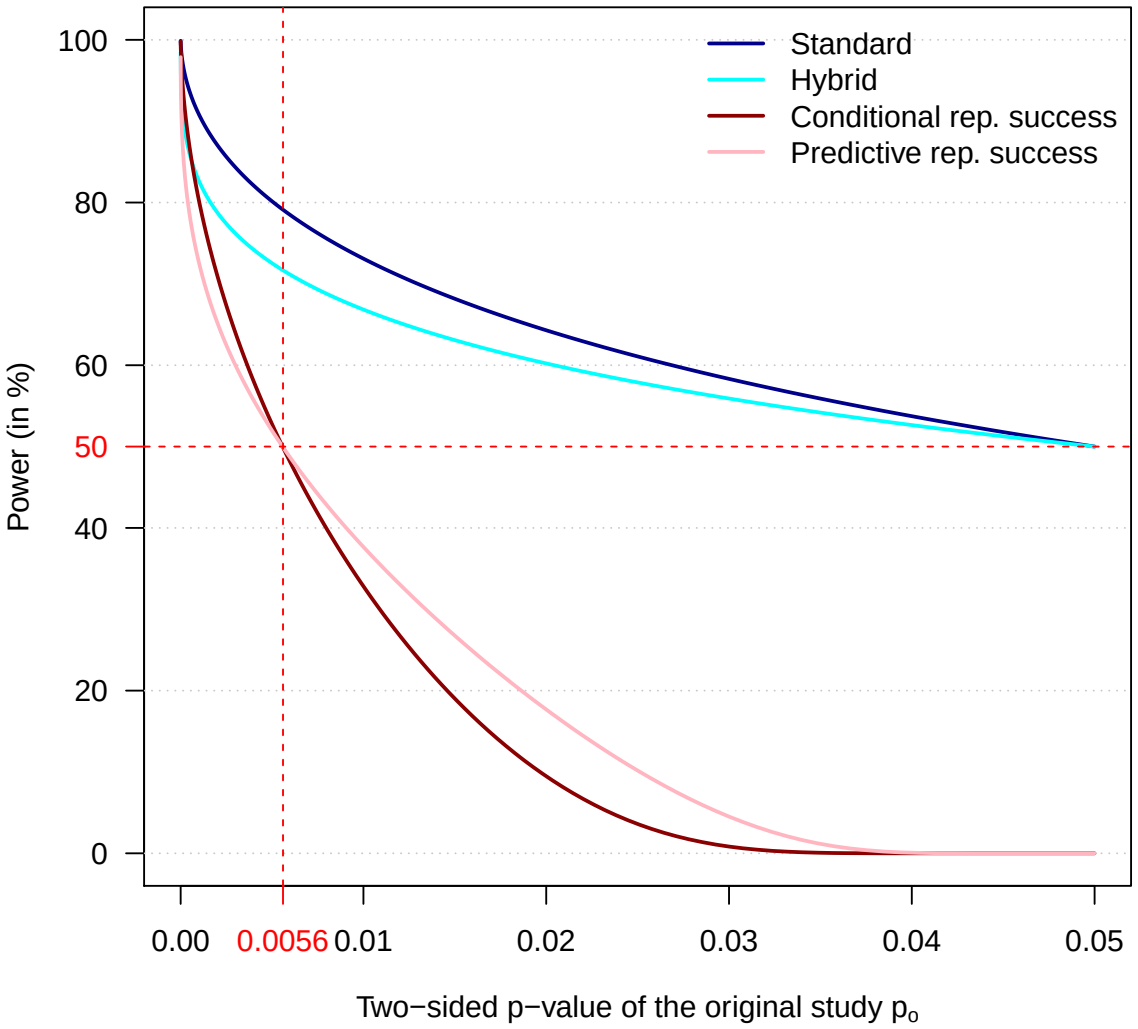


Figure 3.2: Power calculation for a replication study with sample size equal to the original sample size ($c = 1$) as a function of the original p-value p_o with the standard and the hybrid methods and with the conditional and the predictive replication success methods at the 5% level. The horizontal red line indicates a power of 50% and the vertical red line indicates $p_o = 0.0056$.

3.2.3 Power for replication success as a function of the relative sample size c

As explained in Chapter 2, although the behavior of the power with varying p_o might be of great interest, it is not the main focus of this work. We want to make the design of the replication study as efficient as possible and to do so, we investigate the behavior of the power for replication success as a function of the relative sample size c. As the power for replication success only depends on the original p-value p_o and the relative sample size c, by varying p_o and looking at the power as a function of the relative sample size c we are able to observe the most important features. Figure 3.4 shows the power calculations of four studies with original p-values $p_o = 0.03, 0.02, 0.01$

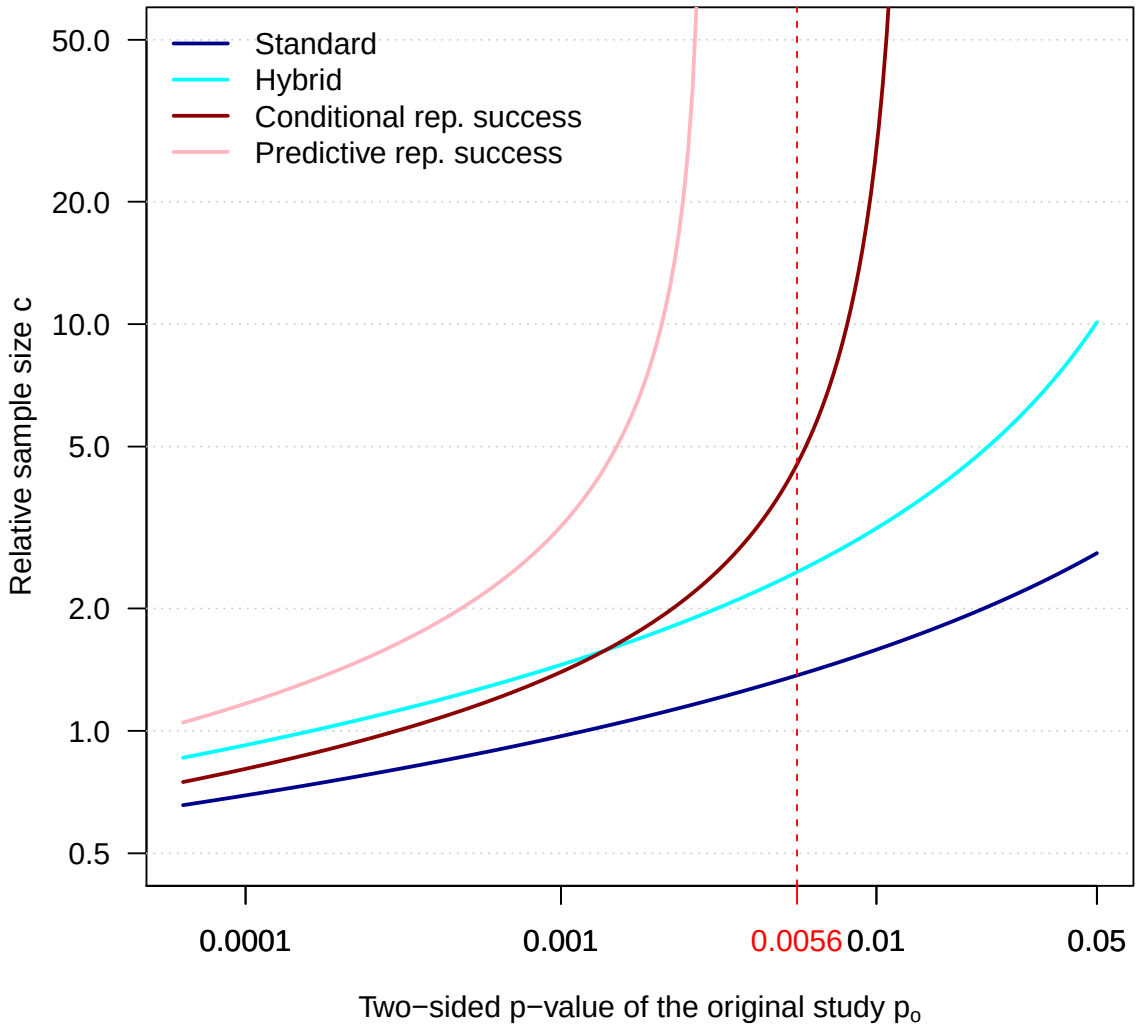


Figure 3.3: Relative sample size c to reach a power of 90% with the standard and the hybrid methods and with the conditional and the predictive replication success methods as a function of the original p-value p_o at the 5% level. The vertical red line indicates $p_o = 0.0056$.

and 0.001. The original p-value p_o seems to dictate the shape of the power curve. The lower the original p-value p_o, the steeper the curve. Moreover, the original p-value p_o appears to influence the monotonicity of the curve. Larger original p-values p_o generate non-monotone replication power curves while smaller p-values p_o give rise to monotonically increasing power curves. In addition, this threshold for monotonicity appears to be different for conditional and predictive power for replication success.

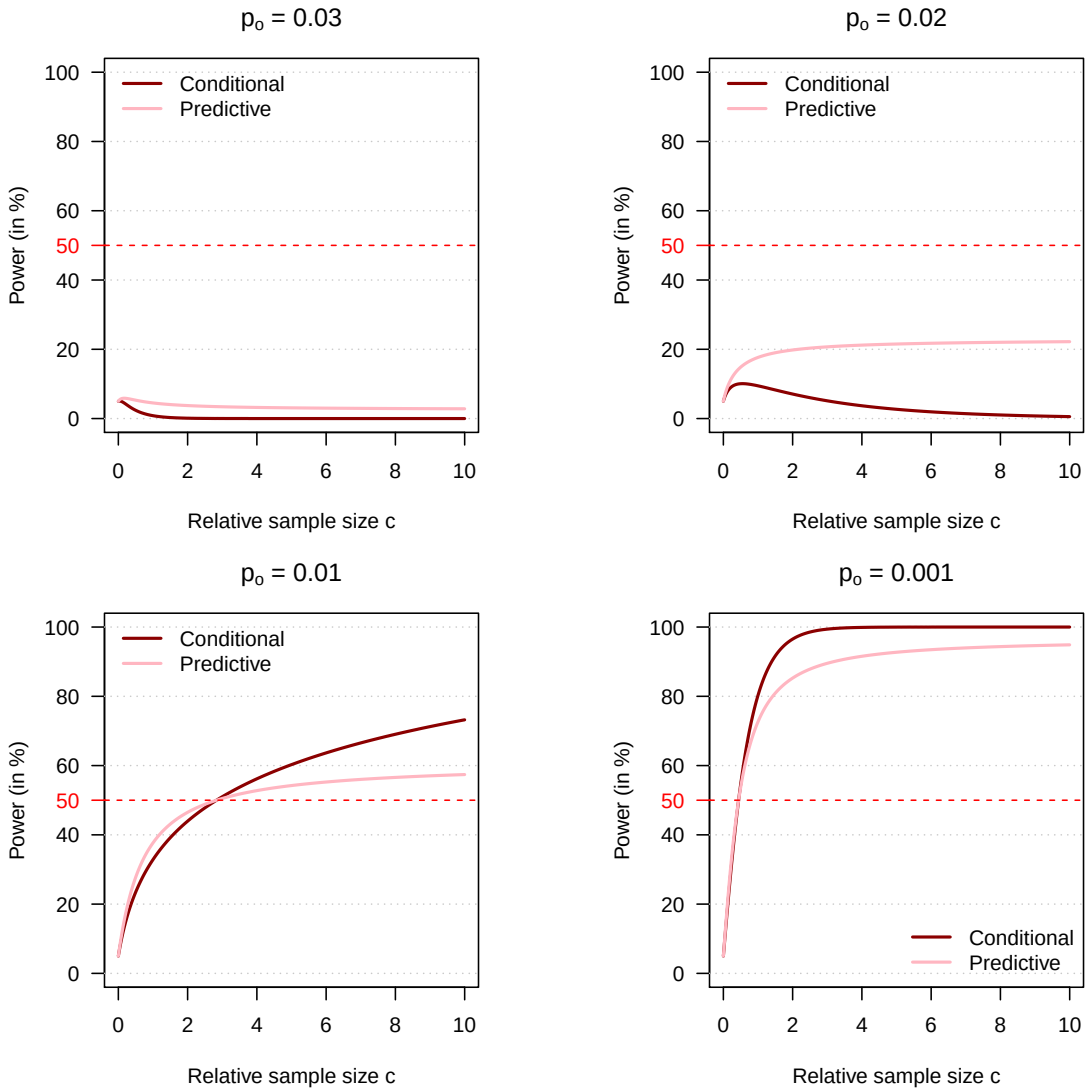


Figure 3.4: Conditional and predictive power for replication success as a function of the relative sample size c for different original p-values p_o at the 5% level. The horizontal red line indicates a power of 50%.

Conditional power for replication success

After examining original studies with different p-values p_o and looking at the (non-)monotonicity of the power curves as a function of the relative sample size c, it appears that original studies with p-values p_o smaller than 0.01266 generate monotonically increasing conditional power curves, whereas original studies with p-values p_o larger than 0.01266 generate non-monotone conditional power curves. Remarkably, the threshold for monotonicity of conditional power for replication success corresponds to Matthews' threshold for intrinsic credibility. We could also notice that the conditional and the predictive power curves cross when the power is 50%. Figure 3.5 presents the power calculations for $p_o = 0.0126$ and $p_o = 0.0127$. For original p-value p_o just above Matthews' threshold, the conditional power for replication success reaches a maximum value

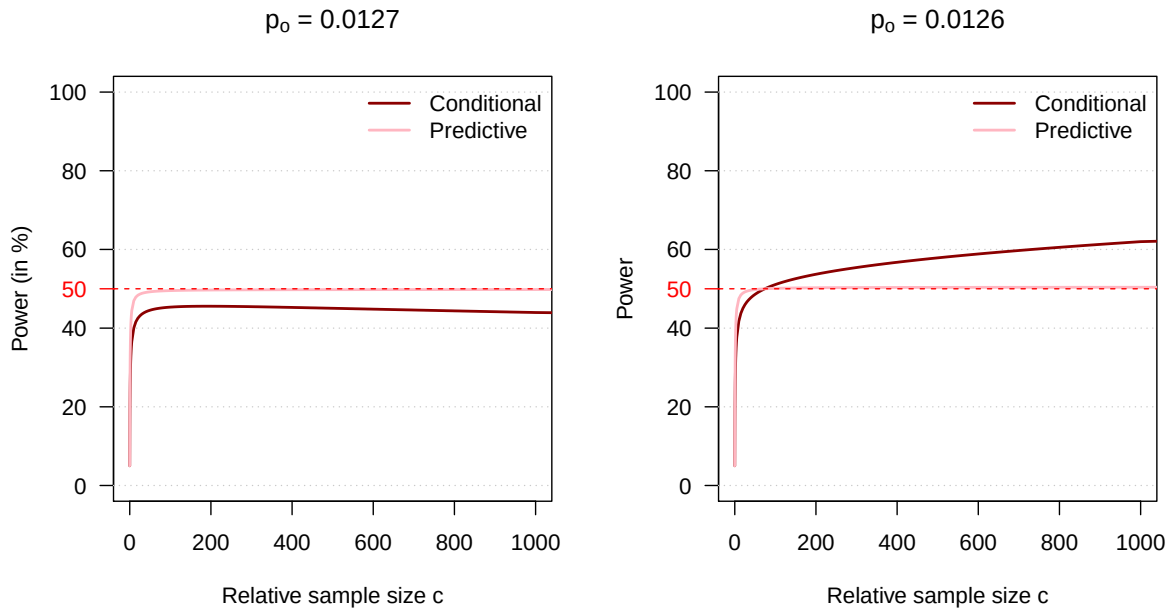


Figure 3.5: Conditional and predictive power for replication success as a function of the relative sample size c for original p-values just above and just below the threshold for monotonicity of the conditional power at the 5% level. The horizontal red line indicates a power of 50%.

slightly below 50% and then slightly decreases. For original p-values p_o just below the threshold, the power for replication success keeps increasing, but at a slower rate for large values of c. For information, the predictive power is also shown and we notice that the limiting predictive power for $p_o = 0.0127$ is 50%.

Predictive power for replication success

The threshold for monotonicity of the predictive power is larger than the threshold for monotonicity of the conditional power. Studies whose original p-value p_o is below 0.023 generate a monotonically increasing power curve. In contrast, studies whose original p-value p_o is above 0.026 generate a non-monotone power curve. Figure 3.6 presents the power calculations for $p_o = 0.026$ and $p_o = 0.023$. The pattern is different than for conditional power. For original p-values p_o just above the threshold, the power for replication success reaches a maximum value largely below 50% and rapidly decreases to a plateau. For original p-values p_o just below the threshold, the power for replication success is monotone but rapidly reaches a plateau and stops increasing. The conditional power is also shown. One notices that the conditional power rapidly decreases to 0% for increasing c with such large original p-values p_o.

3.3 Application

In this section, we re-use the data from Section 2.4 to illustrate our findings about the power for replication success. Figure 3.7 shows the conditional and the predictive power for replication success of the 73 studies of the MA subset computed with their effective original and replication

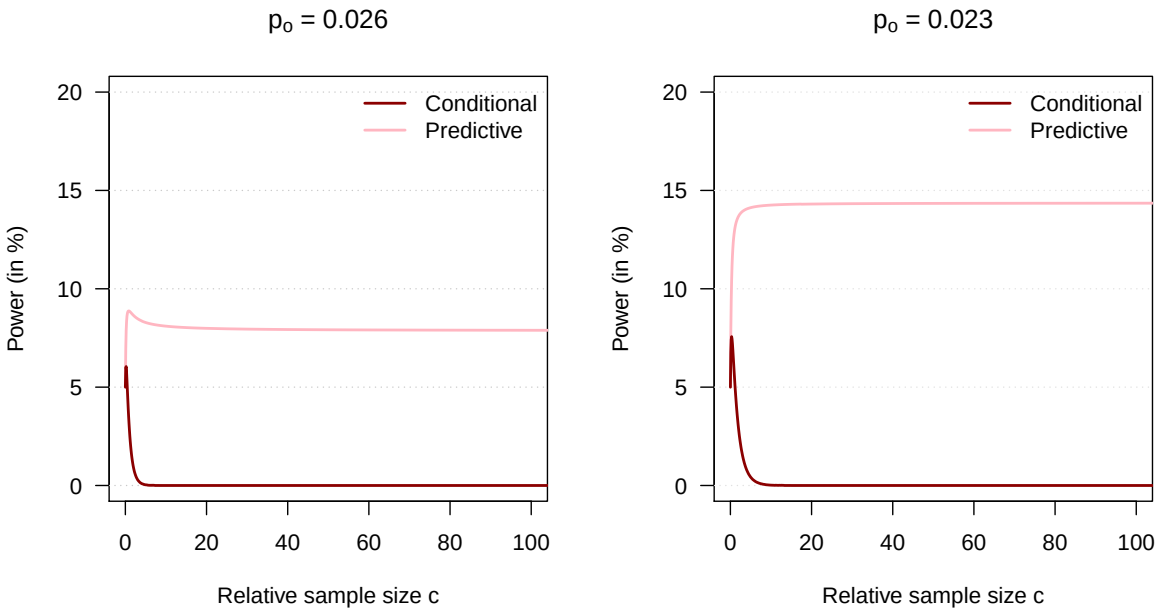


Figure 3.6: Conditional and predictive power for replication success as a function of the relative sample size c for original p-values p_o just above and just below the threshold for monotonicity of the predictive power at the 5% level

sample sizes at the 5% level. In both cases, the median power is below 50% and a considerable number of studies have a power for replication success close to 0%. Furthermore, we observe the same property as in Figure 2.13 with the standard and the hybrid power for replication success: when the power is above 50%, the conditional power is larger than the predictive power and this trend is reversed when the power is below 50%.

We then calculated the power for replication success as a function of the relative sample size c for the 73 studies of the MA subset. We could observe several interesting features which are well illustrated in four studies, see Figure 3.8.

In study 15, the conditional and predictive power for replication success is always 0% as the original study is not significant.

Study 53 is intrinsically credible neither with Matthews' threshold nor with Held's. While the standard and the hybrid methods lead to a large power with sufficiently large relative sample size c, the conditional and predictive power for replication success stays very low and decreases in both cases.

As the p-value p_o of study 1 is between the thresholds for monotonicity of the conditional and of the predictive power for replication success, the predictive power for replication success is monotone while the conditional is not.

Lastly, the original p-value p_o of study 2 is very small and the power curves for replication success behave in a similar way as the standard and the hybrid power curves.

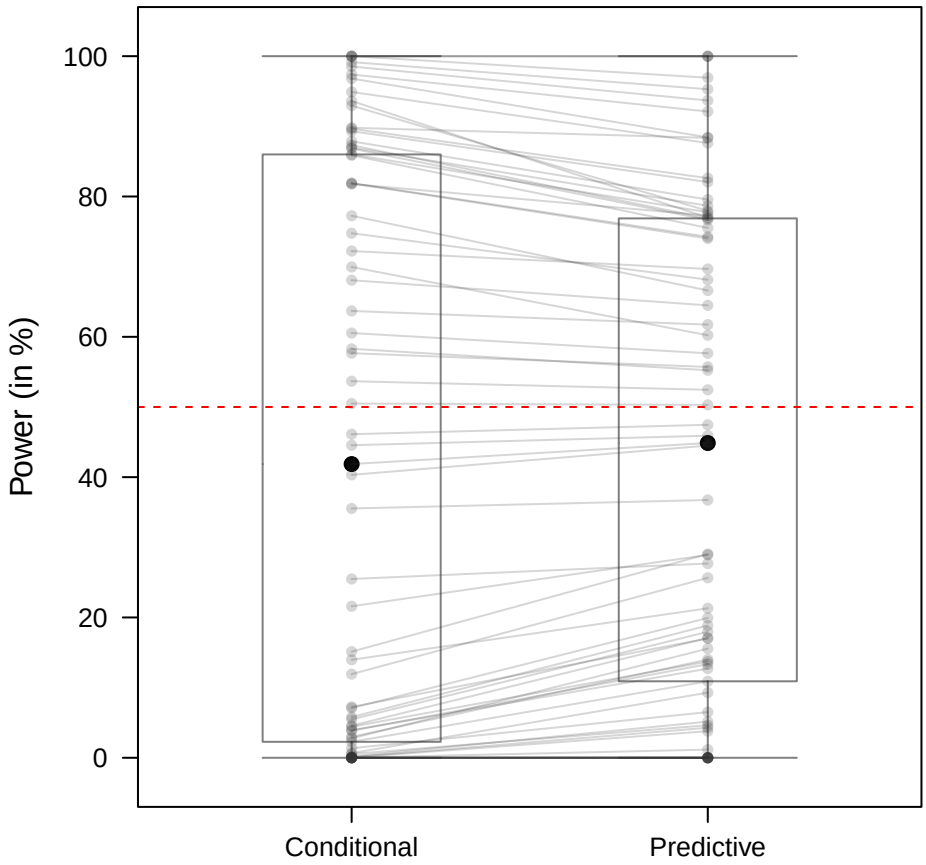


Figure 3.7: Conditional and predictive power for replication success of the 73 studies of the MA subset. The horizontal red line indicates a power of 50%.

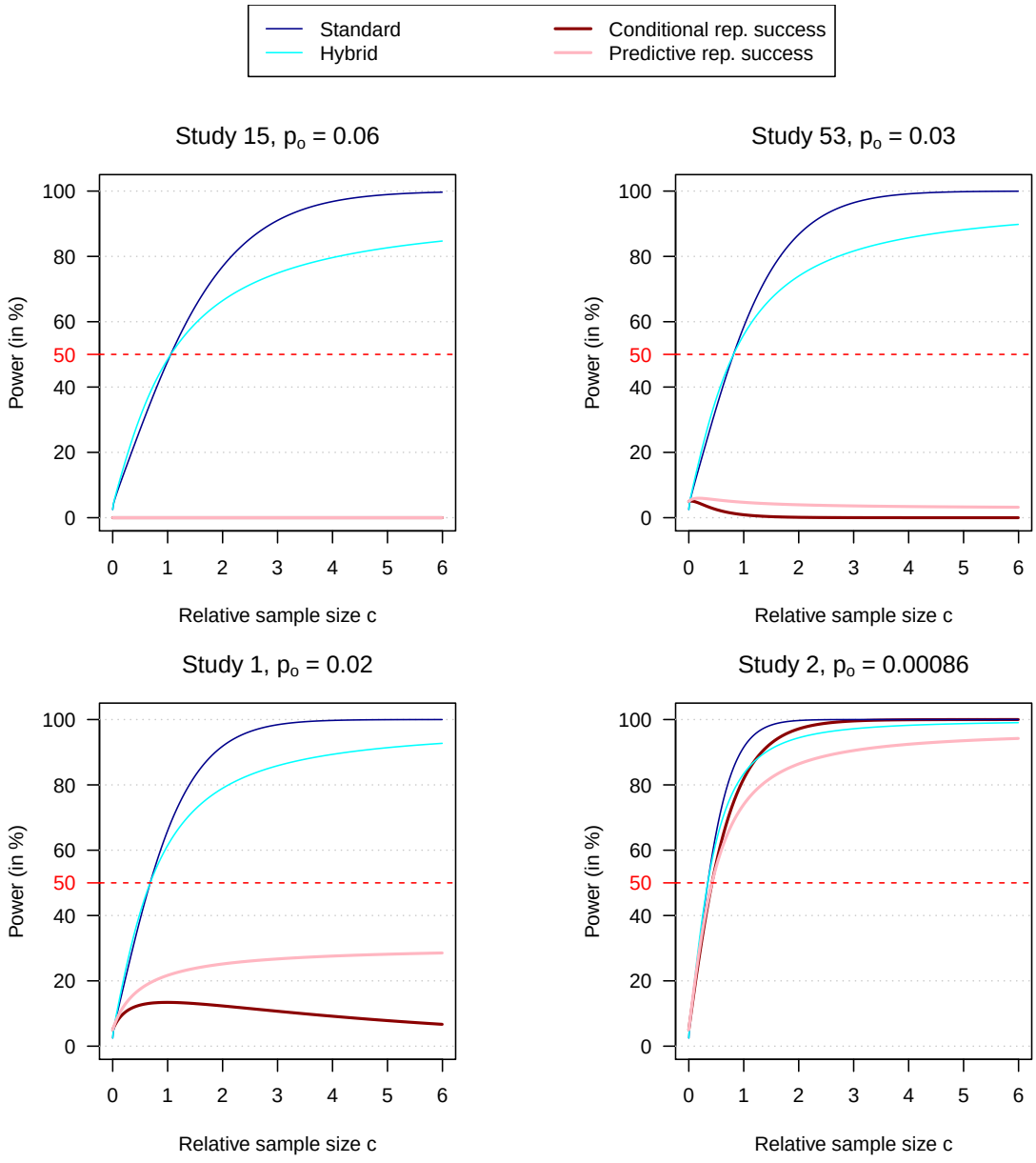


Figure 3.8: Replication power as a function of the relative sample size c for four studies of the MA subset with the standard and the hybrid methods and with the conditional and the predictive replication success methods at the 5% level. The red line indicates a power of 50%.

Chapter 4

Discussion

In this chapter, we summarize and discuss our findings concerning the power for significance and the power for replication success.

4.1 Power for significance

Our findings confirm that using the same sample size in the replication study as in the original study may lead to severely underpowered study designs when the evidence from the original study is only suggestive, even if the original study accurately estimated the true effect size. Moreover, as the true effect size may be smaller than reported by the original study due to publication bias for example, the power is likely to be even lower (Button *et al.*, 2013).

The hybrid method of power calculation acknowledges the uncertainty surrounding the original effect estimate $\hat{\theta}_o$. As a result, larger sample sizes are required to achieve the same level of power than with the standard method. This applies only if the desired power is larger than 50%, which is usually the case. One could argue that if the original effect estimate $\hat{\theta}_o$ is indeed the true effect size, this approach results in a loss of resources. However, this loss of resources is negligible as compared to missing a true, smaller effect because of a low-powered replication study which ignores the uncertainty of the original effect estimate. In addition, the hybrid power is an unconditional power and thus the plausibility of the alternative hypothesis H_1 is taken into account. This prevents the design of studies aiming at detecting implausible effects (Spiegelhalter *et al.*, 2004). The fact that the hybrid power cannot exceed a certain limit dictated by the p-value p_o of the original study indicates that some replication studies are not worth being conducted. This property does not apply to the standard power which goes to 100% with a sufficiently large replication sample size. In summary, the hybrid power gives a more realistic evaluation of the chances of a replication study with n_r subjects to reject the null hypothesis H_0. Grouin *et al.* (2007) investigate the use of predictive power in clinical trials and regret that this method is rarely used to design the trials. We think that it is really important to be aware that predictive power is a different concept than standard power and is better suited for the design of replication studies.

Our investigations demonstrate that the Bayesian method, even if it incorporates the uncertainty of the original effect estimate $\hat{\theta}_o$, is not suited for power calculation in replication studies. The Bayesian replication power of significant original studies is always 100% for a replication

sample size tending to zero and is 100% for any replication sample size if the original study is very convincing. In fact, by planning a Bayesian analysis, the original and replication studies are not considered as two distinct studies, but as one pooled study where an interim analysis is conducted after n_o subjects. This is shown by the Bayesian power formula being the same as the formula of the predictive power at an interim analysis. Dallow and Fina (2011) explain the decreasing property of the predictive power in interim analyses by the potential threat any additional subject represents, able to damage the current results rather than bringing more power to the analysis if the interim analysis is very good. However, although the Bayesian method per se cannot be used to calculate the sample size of replication studies, it could be extended to perform interim analyses during the replication study. Spiegelhalter *et al.* (2004) present the concepts of Bayesian and hybrid predictive power which incorporate an independent prior in the analysis and/or in the design at an interim analysis. By incorporating the knowledge of the original study in the independent prior, one could perform interim analyses within the replication study. This approach can be useful in order to save money and time.

The conditional Bayesian method is actually the worst method to calculate the power of a replication study. It involves the same drawbacks as the Bayesian method but also does not take the uncertainty of the original effect estimate $\hat{\theta}_o$ into account. The replication power of convincing original studies is extremely high with this method, regardless of the sample size.

One common limitation of these methods is that they ignore a possible inflation of the original effect size $\hat{\theta}_o$. This issue could be handled in future research with shrinkage methods for example.

4.2 Power for replication success

As we pointed out the irrelevance of the Bayesian and the conditional Bayesian methods in sample size calculation of replication studies, they are not mentioned in the following and 'power for significance' refers to the standard and the hybrid methods only.

Achieving replication success is more challenging than achieving significance of the replication study. Only original studies with very convincing p-values p_o lead to a reasonable power to achieve replication success. While an original study with p-value $p_o = 0.05$ ensures a power for significance of 50% with the same sample size in the replication as in the original study, this very same original study does not allow the success of the replication study whatever replication sample size is chosen. The power for replication success introduces smaller thresholds than the standard $p \leq 0.05$, namely the thresholds for intrinsic credibility. If the original study is intrinsically credible at Held's threshold ($p_o \leq 0.0056$), the conditional and predictive power for replication success is 50% for a replication study the same size as the original. But if the original study is not intrinsically credible at Matthews' threshold ($p_o > 0.01266$), then the conditional and predictive power for replication success will never be larger than 50% whatever the replication sample size. These findings show that only intrinsically credible original studies lead to replication success at an acceptable power and should encourage researchers to lower the significance threshold for claims of new discoveries.

The incorporation of the uncertainty of the original effect estimate $\hat{\theta}_o$ in the calculation of the power for replication success gives similar results as in the calculation of the power for signif-

icance. First, the predictive power is also smaller than the conditional power provided that the conditional power is larger than 50%. Second, while the conditional power for replication success (assuming the original study is intrinsically credible at Matthews' threshold, $p_o \leq 0.01266$) increases with increasing sample size to reach a power of 100%, the predictive power for replication success is limited and this limit appears to be a function of the p-value p_o. However, this limit is drastically lower than for significance.

The non-monotonicity property of the power for replication success as a function of the relative sample size c reminds of the non-monotonicity property of the Bayesian power curve. However, the two situations differ in several points. First, the Bayesian power curve is non-monotone when p_o is *below* a certain level ($p_o < 0.05$ for the traditional 5% level). A non-monotone Bayesian power curve is a good sign, meaning the original experiment is significant. In contrast, the power for replication success is non-monotone when p_o is *above* a certain level ($p_o > 0.01266$ and $p_o > 0.026$ for conditional and predictive power, respectively). A non-monotone power curve for replication success already indicates that a replication study will probably not be successful. Moreover, the non-monotone Bayesian power curve is convex whereas the non-monotone power for replication success curve is concave.

In this thesis, we investigated the power for replication success in the setting of a superiority study. This approach could be extended in future research to equivalence studies, studies with small sample sizes and multivariate outcomes.

Chapter 5

Software

All analyses were performed in the R system of statistical software (R version 3.5.1 (2018-07-02)), freely available at http://www.r-project.org/. The following packages sampleSize, pCalibrate, ggplot2, reporttools, lattice, xtable and knitr and the base packages stats, graphics, grDevices, utils, datasets, methods and base were used for the analysis of the compilation of this report. The computing environment on the author's personal computer had the following specifications: OSX Mojave, Version 10.14.4 (Operating system), 2,7 GHz Intel Core i5 (Processor) and 8 Go 1867 MHz DDR3 (Memory). This document was generated on May 31, 2019 at 07:52.

Appendix A

Appendix

A.1 Derivation of the Bayesian significance

Bayesian significance S_ϵ^B means that the ϵ-quantile of the posterior distribution $\theta\,|\,\overline{Y}_{n_r}$ given in (2.7) is larger than zero. Let z_ϵ be the ϵ-quantile of the standard normal distribution. By using the equation (2.7), it turns out that

$$\frac{z_\epsilon \sigma}{\sqrt{n_o + n_r}} + \frac{n_o \hat{\theta}_o + n_r Y_{n_r}}{n_o + n_r}$$

is the ϵ-quantile of $\theta\,|\,\overline{Y}_{n_r}$.We then have

$$S_\epsilon^B \Leftrightarrow \frac{z_\epsilon \sigma}{\sqrt{n_o + n_r}} + \frac{n_o \hat{\theta}_o + n_r Y_{n_r}}{n_o + n_r} > 0$$

which can be rearranged as

$$Y_{n_r} > \frac{-\sqrt{n_o + n_r} z_\epsilon \sigma - n_o \hat{\theta}_o}{n_r}\,.$$

A.2 Derivation of the minimum Bayesian power

As the function $\Phi[\cdot]$ is monotonically increasing, considering its argument is sufficient when manipulating it. The first derivative of the Bayesian power formula is

$$\begin{aligned}
&\frac{d}{dn_r}\left(\frac{\hat{\theta}_o \sqrt{n_o}\sqrt{n_o + n_r}}{\sigma\sqrt{n_r}} + \sqrt{\frac{n_o}{n_r}} z_\epsilon\right) = \\
&\frac{\hat{\theta}_o \sqrt{n_o}(\frac{1}{2}(n_o + n_r)^{-1/2}\sigma\sqrt{n_r} - 1/2(n_o + n_r)^{1/2})}{\sigma^2 n_r} + \sqrt{n_o} z_\epsilon n_r^{-3/2}\left[-\frac{1}{2}\right] = \\
&1/2\sqrt{n_o} n_r^{-3/2}\left[\frac{\hat{\theta}_o n_r (n_o + n_r)^{-\frac{1}{2}}}{\sigma} - \frac{\hat{\theta}_o (n_o + n_r)^{1/2}}{\sigma} - z_\epsilon\right].
\end{aligned}$$

By setting it to 0 and solving for n_r, we obtain the replication sample size n_r needed to reach the minimum power, which turns out to be

$$\frac{1}{2}\sqrt{n_o}n_r^{-\frac{3}{2}}\left[\frac{\hat{\theta}_o n_r(n_o+n_r)^{-\frac{1}{2}}}{\sigma}-\frac{\hat{\theta}_o(n_o+n_r)^{\frac{1}{2}}}{\sigma}-z_\epsilon\right]=0 \quad \Leftrightarrow \quad n_r = n_o\left[\frac{\hat{\theta}_o^2 n_o}{\sigma^2 z_\epsilon^2}-1\right]$$
$$\Leftrightarrow \quad c=\frac{t_o^2}{z_\epsilon^2}-1\,. \tag{A.1}$$

By plugging (A.1) in the alternative Bayesian power formula given in (2.13), we find the corresponding minimum Bayesian power which is

$$\Pr(S_\epsilon^B)=\Phi\left[\sqrt{t_o^2-z_\epsilon^2}\right]\,.$$

Bibliography

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B., Wagenmakers, E.-J., Berk, R., Bollen, K., Brembs, B., Brown, L., Camerer, C., Cesarini, D., Chambers, C. D., Clyde, M., Cook, T. D., De Boeck, P., Dienes, Z., Dreber, A., Easwaran, K., Efferson, C., and Johnson, V. (2017). Redefine statistical significance. *Nature Human Behaviour*, **2**, 6–10. 31

Box, G. E. (1980). Sampling and Bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society: Series A (General)*, **143**, 383–430. 31

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., and Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, **14**, 365. 1, 41

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., Heikensten, E., Holzmeister, F., Imai, T., Isaksson, S., Nave, G., Pfeiffer, T., Razen, M., and Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, **351**, 1433–1436. 1

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T., Isaksson, S., Manfredi, D., Rose, J., Wagenmakers, E.-J., and Wu, H. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, **2**, 637–644. 1

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, **49**, 997–1003. 31

Dallow, N. and Fina, P. (2011). The perils with the misuse of predictive power. *Pharmaceutical Statistics*, **10**, 311–317. 42

Erdfelder, E., Faul, F., and Buchner, A. (1996). Gpower: A general power analysis program. *Behavior Research Methods, Instruments, & Computers*, **28**, 1–11. 25

Goodman, S. N. (1992). A comment on replication, p-values and evidence. *Statistics in Medicine*, **11**, 875–879. 1

Goodman, S. N., Fanelli, D., and Ioannidis, J. P. (2016). What does research reproducibility mean? *Science Translational Medicine*, **8**, 341ps12–341ps12. 31

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., and Altman, D. G. (2016). Statistical tests, p-values, confidence intervals, and power: a guide to misinterpretations. *European Journal of Epidemiology*, **31**, 337–350. 31

Grouin, J.-M., Coste, M., Bunouf, P., and Lecoutre, B. (2007). Bayesian sample size determination in non-sequential clinical trials: Statistical aspects and some regulatory considerations. *Statistics in Medicine*, **26**, 4914–4924. 41

Held, L. (2019a). The assessment of intrinsic credibility and a new argument for $p < 0.005$. *Royal Society Open Science*, **6**, 181534. iii, 31

Held, L. (2019b). A new standard for the analysis and design of replication studies. Technical report, arXiv 1811.10287, v2. iii, 1, 31

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Medicine*, **2**, e124. iii, 1, 23

Ioannidis, J. P. (2008). Why most discovered true associations are inflated. *Epidemiology*, **19**, 640–648. 1

Ioannidis, J. P. (2018). The proposal to lower p-value thresholds to .005. *JAMA*, **319**, 1429–1430. 31

Johnson, V. E. (2013). Revised standards for statistical evidence. *Proceedings of the National Academy of Sciences*, **110**, 19313–19317. 31

Johnson, V. E., Payne, R. D., Wang, T., Asher, A., and Mandal, S. (2017). On the reproducibility of psychological science. *Journal of the American Statistical Association*, **112**, 1–10. 1, 23

Matthews, R. A. (2018). Beyond 'significance': principles and practice of the analysis of credibility. *Royal Society Open Science*, **5**, 171047. iii, 31

O'Hagan, A. and Stevens, J. W. (2001). Bayesian assessment of sample size for clinical trials of cost-effectiveness. *Medical Decision Making*, **21**, 219–230. 5

O'Hagan, A., Stevens, J. W., and Campbell, M. J. (2005). Assurance in clinical trial design. *Pharmaceutical Statistics*, **4**, 187–201. 6

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, **349**, aac4716. iii, 1, 2, 3, 23

Sahu, S. and Smith, T. (2006). A Bayesian method of sample size determination with practical applications. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **169**, 235–253. 5

Schönbrodt, F. D. and Wagenmakers, E.-J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review*, **25**, 128–142. 5

Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science*, **26**, 559–569. 31

Spiegelhalter, D. J., Abrams, K. R., and Myles, J. P. (2004). *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*, volume 13. John Wiley & Sons. iii, 1, 3, 8, 21, 22, 41, 42

Spiegelhalter, D. J., Freedman, L. S., and Blackburn, P. R. (1986). Monitoring clinical trials: conditional or predictive power? *Controlled Clinical Trials*, **7**, 8–17. 5, 6

Wang, F. and Gelfand, A. E. (2002). A simulation-based approach to Bayesian sample size determination for performance under a given model and for separating models. *Statistical Science*, **17**, 193–208. 5

Wang, Y., Fu, H., Kulkarni, P., and Kaiser, C. (2013). Evaluating and utilizing probability of study success in clinical development. *Clinical Trials*, **10**, 407–413. 6

Wasserstein, R. L. and Lazar, N. A. (2016). The ASA's statement on p-values: context, process, and purpose. *The American Statistician*, **70**, 129–133. 31