Design and Analysis of Replication Studies with `ReplicationSuccess`

Leonhard Held, Charlotte Micheloud, Samuel Pawel
Epidemiology, Biostatistics and Prevention Institute (EBPI)
Center for Reproducible Science (CRS)
University of Zurich, Switzerland

Abstract

This vignette provides an introduction to the R package `ReplicationSuccess`. The package contains utilities for planning and analysing replication studies. Traditional methods based on statistical significance and confidence intervals, as well as more recently developed methods such as the sceptical p-value (Held, 2020a) are included. The functionality of the package is illustrated using data sets from four large-scale replication projects which come also with the package.

1 Introduction

Over the course of the last decade, the conduct of replication studies has increased substantially. These developments were mainly caused by the so-called "replication crisis" in the social and life-sciences. However, there is no consensus on which statistical analysis approach should be used to assess whether a replication study successfully replicated an original discovery. Moreover, depending on the chosen analysis approach, the statistical considerations in the design of the replication study differ.

The R package `ReplicationSuccess` provides functionality to analyse and plan replication studies in several different ways. Specifically, functions for power and samples size calculations based on statistical significance, as well as based on more recent methods, such as the sceptical p-value (Held, 2020a), are included. In this vignette the usage of the package is illustrated on the data sets from four large-scale replication projects which are also included in the package.

1.1 Statistical framework

`ReplicationSuccess` assumes a simple but general and practically relevant statistical framework for effect sizes. Specifically, after a suitable transformation the effect estimates are assumed to be approximately normally distributed with known variances which do not depend on the effect anymore. The same framework is also common in the meta-analysis literature and can for example be applied to mean differences, odds ratios (log transformation), or correlation coefficients (Fisher z-transformation).

Moreover, most functions in `ReplicationSuccess` take unitless quantities as inputs. In particular, the z-values $z_o = \hat{\theta}_o/\sigma_o$, $z_r = \hat{\theta}_r/\sigma_r$, and the variance ratio $c = \sigma_o^2/\sigma_r^2$ ($\hat{\theta}$ denotes an effect estimate and σ^2 the corresponding variance, the subscripts indicate original or replication). Assuming that the standard errors of the effect estimates only depend on some unit variance κ^2 and inversely on the sample size of the study, *i. e.* $\sigma_o^2 = \kappa^2/n_o$ and $\sigma_r^2 = \kappa^2/n_r$, the variance ratio is also the relative sample size $c = \sigma_o^2/\sigma_r^2 = n_r/n_o$. For this reason, all functions from `ReplicationSuccess` used for sample size computations return c.

2 Data sets

`ReplicationSuccess` includes data from four replication projects, all with a "one-to-one" design (*i. e.* one replication for one original study). They come from the following projects:

- **Reproducibility Project: Psychology:** In the *Reproducibility Project: Psychology* 100 replications of studies from the field of psychology were conducted (Open Science Collaboration, 2015). The original studies were published in three major Psychology journals in the year 2008. Only the study pairs of the "meta-analytic subset" are included here, which consists of 73 studies where the standard error of the Fisher z-transformed effect estimates can be computed (Johnson et al., 2016).

- **Experimental Economics Replication Project:** This project attempted to replicate 18 experimental economics studies published between 2011 and 2015 in two high impact economics journals (Camerer et al., 2016). For this project a *prediction market* was also conducted in order to estimate the peer beliefs about whether a replication will result in a statistically significant result. Prediction markets are a tool to aggregate beliefs of market participants regarding the possibility of an investigated outcome and they have been used successfully in numerous domains, *e. g.* sports and politics (Dreber et al., 2015). The estimated peer beliefs are also included for each study pair.

- **Social Sciences Replication Project:** This project involved 21 replications of studies on the social sciences published in the journals *Nature* and *Science* between 2010 and 2015 (Camerer et al., 2018). As in the experimental economics replication project, a prediction market to estimate peer beliefs about the replicability of the original studies was conducted and the resulting belief estimates are also provided in the package. In this project, the replications were conducted in two stages. In stage 1, the replication studies had 90% power to detect 75% of the original effect estimate. Data collection eas stopped if a two-sided p-value < 0.05 and an effect in the same direction as the original were found. If not, data collection was continued in stage 2 to have 90% power to detect 50% of the original effect size for the first and second data collection pooled.

- **Experimental Philosophy Replicability Project:** In this project, 40 replications of experimental philosophy studies were carried out. The original studies had to be published between 2003 and 2015 in one of 35 journals in which experimental philosophy research is usually published (a list defined by the coordinators of this project) and they had to be listed on the experimental philosophy page of the Yale university (Cova et al., 2018). The data from the subset of 31 study pairs where effect estimates on correlation scale as well as effective sample size for both the original and replication were available are included in the package.

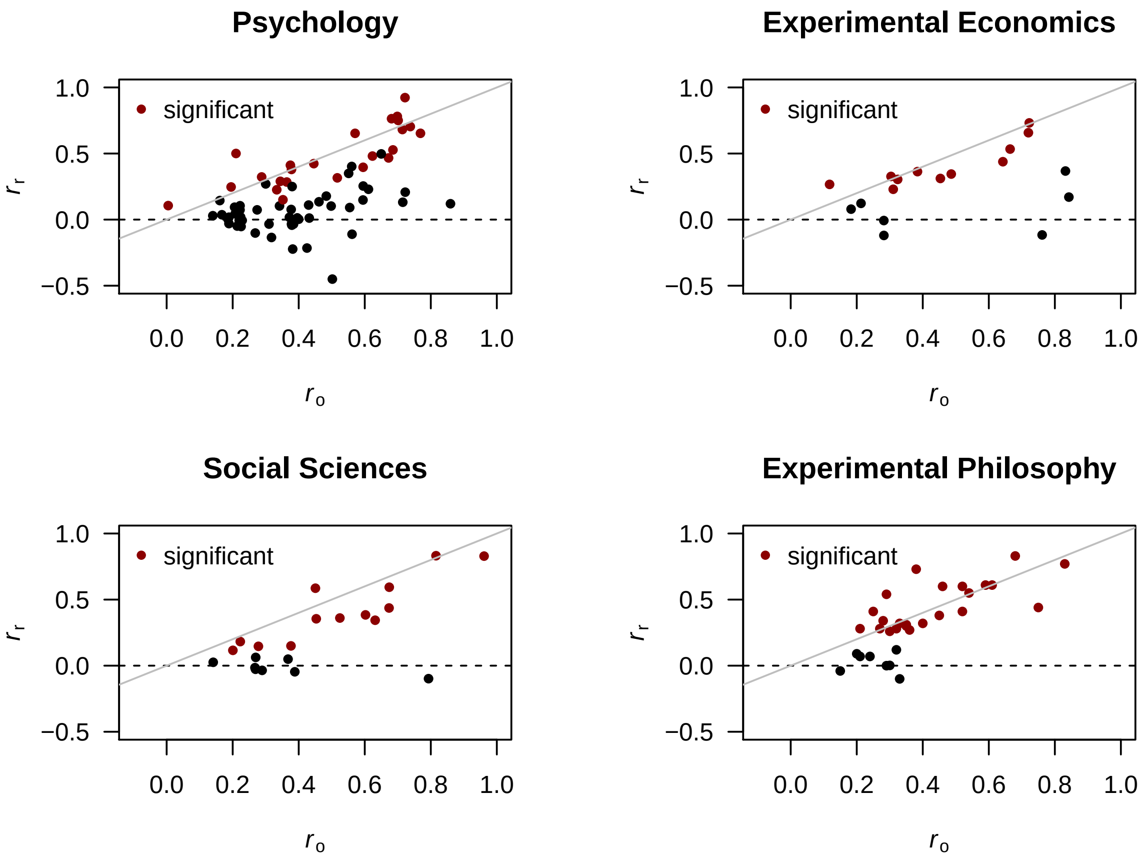
In all data sets, effect estimates are provided as correlation coefficients (r), as well as Fisher z-transformed correlation coefficients ($\hat{\theta} = \tanh^{-1}(r)$). In the descriptive analysis of data from replication projects it has become common practice to transform effect sizes to the correlation scale, because correlations are bounded to the interval between minus one and one and thus easy to compare and interpret. Design and statistical analysis, on the other hand, is then usually carried out on a scale where the estimates are approximately normally distributed. For correlation coefficients this is the case after applying the Fisher z-transformation, which leads to their variance asymptotically being only a function of the study sample size n, *i. e.* $\text{Var}(\hat{\theta}) = 1/(n-3)$ (Fisher, 1921).

The data can be loaded with the command `data("RProjects")`. For a description of the variables see the documentation with `?RProjects`. An extended version of the Social Sciences Replication Project including the details of stages one and two can be loaded with `data("SSRP")`. It is a good idea to first compute the unitless quantities z_o, z_r and c, since most functions of the package use them as input. We also use the function `z2p` to compute the one-sided p-values for original and replication study. As all original estimates are positive, we specify the argument `alternative` to `"greater"`.

```
library(ReplicationSuccess)
data("RProjects")
str(RProjects)

## 'data.frame': 143 obs. of  13 variables:
##  $ study    : chr  "A Roelofs" "AL Morris, ML Still" "B Liefooghe, P Barrouillet, A Vandierendonck,
##  $ project  : chr  "Psychology" "Psychology" "Psychology" "Psychology" ...
##  $ ro       : num  0.595 0.611 0.425 0.229 0.461 ...
##  $ rr       : num  0.14834 0.2296 -0.21524 -0.00611 0.13481 ...
##  $ fiso     : num  0.685 0.711 0.454 0.233 0.499 ...
##  $ fisr     : num  0.14944 0.23377 -0.21866 -0.00611 0.13564 ...
##  $ se_fiso  : num  0.2887 0.2132 0.2085 0.0727 0.1826 ...
##  $ se_fisr  : num  0.1925 0.2132 0.1826 0.0612 0.1474 ...
##  $ po       : num  0.017688 0.000858 0.029546 0.001368 0.006277 ...
##  $ pr       : num  0.437 0.273 0.231 0.92 0.358 ...
##  $ pm_belief: num  NA NA NA NA NA NA NA NA NA NA ...
##  $ nr       : num  30 25 33 270 49 33 16 33 31 31 ...
##  $ no       : num  15 25 26 192 33 25 101 39 30 23 ...

## computing zo, zr, c
RProjects$zo <- with(RProjects, fiso/se_fiso)
RProjects$zr <- with(RProjects, fisr/se_fisr)
RProjects$c <- with(RProjects, se_fiso^2/se_fisr^2)

## computing one-sided p-values for alternative = "greater"
RProjects$po1 <- z2p(z = RProjects$zo, alternative = "greater")
RProjects$pr1 <- z2p(z = RProjects$zr, alternative = "greater")
```

Note that each variable ending with an `o` is associated with the original, while each variable ending with an `r` is associated with the replication. Plotting the original versus the replication effect estimate on the correlation scale gives a good overview of the data.

```
## plots of effect estimates
par(mfrow = c(2, 2), las = 1, mai = rep(0.65, 4))
for (p in unique(RProjects$project)) {
    data_project <- subset(RProjects, project == p)
    significant <- ifelse(data_project$pr < 0.05, "darkred", "black")
    plot(rr ~ ro, data = data_project, ylim = c(-0.5, 1), col = significant,
         xlim = c(-0.1, 1), main = p, xlab = expression(italic(r)[o]),
         cex = 0.7, pch = 19, ylab = expression(italic(r)[r]))
    legend("topleft", legend = "significant", pch = 20, col = "darkred", bty = "n")
    abline(h = 0, lty = 2)
    abline(a = 0, b = 1, col = "grey")
}
```



In most cases the replication estimate is smaller than the corresponding original estimate. Furthermore, a substantial number of the replication estimates do not achieve statistical significance at one-sided 2.5% level, while almost all original estimates did.

3 Design and analysis of replication studies

Although a replication study needs to be planned and conducted before the results can be analysed, we will first discuss the particular analysis approaches. We do this because the chosen analysis strategy substantially influences the design of a replication study. In the design phase of a replication study, we focus only on the sample size determination.

3.1 Statistical significance

Analysis One of the most commonly used approaches to analyse the result of a replication study is to declare a replication study successful if the replication estimate achieves the same statistical significance status as the original estimate and also goes in the same direction. There are some variations of this approach, for example, Camerer et al. (2016) only assessed whether the replication effect is significant in the same direction, but not whether the original effect shows the same significance status.

For the four data sets, we can simply check whether the (two-sided) *p*-values of original and replication are both below the conventional threshold 0.05 and whether the directions of the effects are the same.

```
for (p in unique(RProjects$project)) {
    data_project <- subset(RProjects, project == p)
    significant_O <- data_project$po < 0.05
    significant_R <- data_project$pr < 0.05
    success <- (significant_O == TRUE) & (significant_R == TRUE) &
        (sign(data_project$fiso) == sign(data_project$fisr))
    cat(paste0(p, ": \n"))
    cat(paste0(round(mean(significant_O)*100, 1), "% original studies significant (",
               sum(significant_O), "/", length(significant_O), ")\n"))
    cat(paste0(round(mean(significant_R)*100, 1), "% replications significant (",
               sum(significant_R), "/", length(significant_R), ")\n"))
    cat(paste0(round(mean(success)*100, 1),
               "% both studies significant in the same direction (",
               sum(success), "/", length(success), ")\n \n"))
}

## Psychology:
## 89% original studies significant (65/73)
## 32.9% replications significant (24/73)
## 28.8% both studies significant in the same direction (21/73)
##
## Experimental Economics:
## 88.9% original studies significant (16/18)
## 61.1% replications significant (11/18)
## 55.6% both studies significant in the same direction (10/18)
##
## Social Sciences:
## 100% original studies significant (21/21)
## 61.9% replications significant (13/21)
## 61.9% both studies significant in the same direction (13/21)
##
## Experimental Philosophy:
## 96.8% original studies significant (30/31)
## 74.2% replications significant (23/31)
## 74.2% both studies significant in the same direction (23/31)
##
```

Despite its appealing simplicity, assessing replication success with statistical significance is often criticized. For example, non-significant replication results are expected if the original finding was a false positive (*e. g.* with 95% probability if the two-sided significance level is 5%), on the other hand they are also expected with non-negligible probability if the underlying effect is present (Goodman, 1992). Conversely, when the effect estimate of the replication is much smaller than the estimate from the original study, statistical significance can still be achieved by simply increasing the sample size.

Design 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. To assure that the replication study reliably detects true effects, the studies should be well-powered. In classical sample size planning, usually a clinically relevant effect is specified and the sample size is then determined so that it can be detected with a certain power. Luckily, in the replication setting the clinically relevant effect does not need to be specified but can be replaced with the effect estimate from the original study. However, using the standard sample size calculation approach is not well suited, because the uncertainty of the original effect estimate is ignored.

One way of tackling this issue is to use a Bayesian approach, incorporating the original estimate and its precision into a design prior that is used for power calculations. This corresponds to the concept of “predictive power” and generally leads to larger sample sizes than the standard method. In practice, however, often more ad hoc approaches are used. For instance, the original estimate is just shrunken by an (arbitrary) constant, *e. g.* it was halved in the sociel sciences replication project, and standard sample size calculations are then carried out.

Using the function `sampleSizeSignificance`, it is straightforward to plan the sample size of the replication study with the just mentioned approaches. The argument `designPrior` allows to carry out sample size planning based on classical power ignoring the uncertainty (`"conditional"`) or based on predictive power (`"predictive"`). Moreover, ad hoc shrinkage can be specified with the argument `shrinkage`. It must be noted that the function `sampleSizeSignificance`, as well as most of the functions from the package, takes z-values and no p-values as arguments. The conversion between the two measures can easily be done using the function `p2z`.

The following code shows a few examples. Note that the function returns the required relative sample size $c = n_r/n_o$, *i. e.* by which factor the sample size of the replication needs to be changed compared to the original study.

```
sampleSizeSignificance(zo = 2.5, power = 0.8, level = 0.05, designPrior = "conditional")

## [1] 0.9892092

sampleSizeSignificance(zo = 2.5, power = 0.8, level = 0.05, designPrior = "predictive")

## [1] 1.388114

sampleSizeSignificance(zo = 2.5, power = 0.8, level = 0.05, designPrior = "conditional",
                       shrinkage = 0.25)

## [1] 1.758594
```

Figure 1 shows the power to achieve significance in the replication as a function of either the (two-sided) p-value or the z-value of the original study. If the original estimate was just significant at the 0.05 level, the probability for significance in the replication is just about 0.5 for conditional and predictive power. This result was first mentioned by Goodman (1992) already two decades ago, yet many practitioners of statistics still find it counterintuitive, because they confuse type I error rates with replication probabilities. Thus, for the replication to achieve significance with high probability, the sample size needs to be increased compared to the original if the the evidence for the original discovery was only weak or moderate (Figure 2).

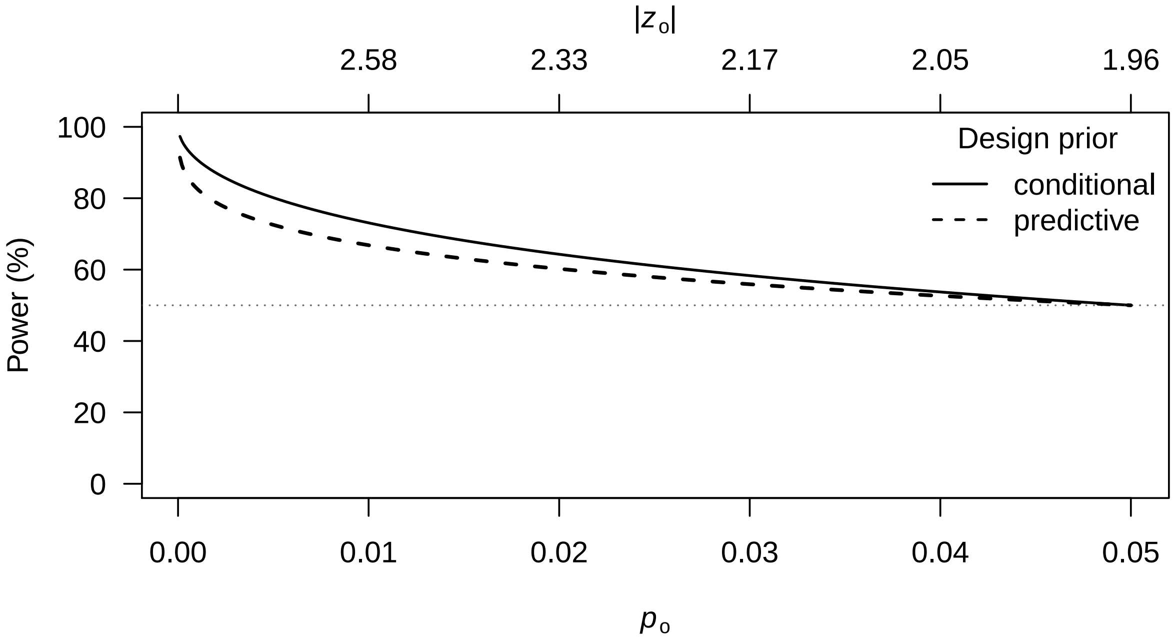


Figure 1: Power to achieve significance at the one-sided 2.5% level in replication as a function of (two-sided) p-value or z-value of original study using the same sample size as in the original study.

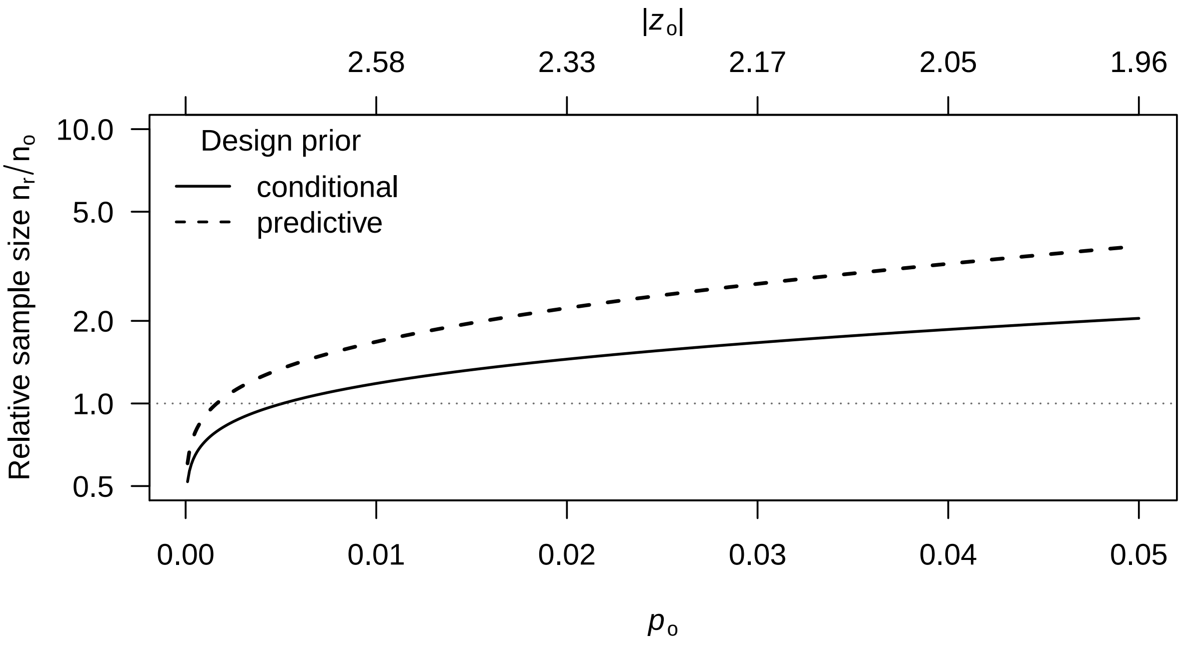


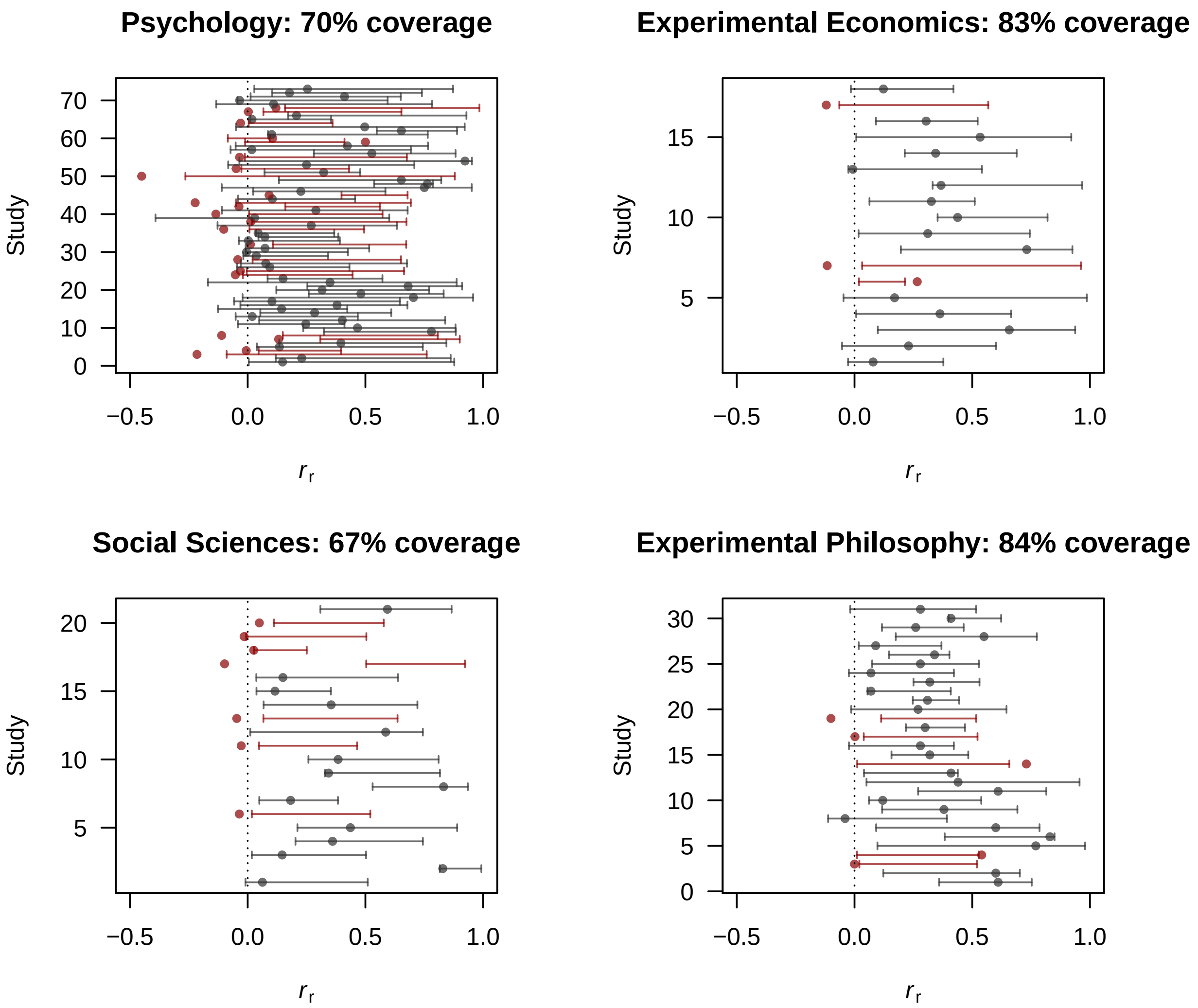
Figure 2: Relative sample size to achieve significance at the one-sided 2.5% level with 80% power as a function of (two-sided) p-value or z-value of original study.

3.2 Compatibility of effect size

Analysis Another analysis approach that has been used is to compare the effect estimates from original and replication study. A reasonable way to assess whether the observed estimates are compatible is to check whether the replication estimate is contained within its prediction interval based on the original estimate (Patil et al., 2016). With the function `predictionInterval`, a prediction interval of the replication effect estimate can be computed under different predictive distributions which depend on the design prior. The default design prior `"predictive"` is likely the choice most people would want to use as it takes into account the uncertainty of the original estimate without shrinking it.

For the four data sets, we can easily compute the prediction intervals and then check whether the replication estimates are contained within them. For easier visual assessment we transform the intervals and estimates back to the correlation scale.

```r
## compute prediction intervals for replication projects
par(mfrow = c(2, 2), las = 1, mai = rep(0.65, 4))
for (p in unique(RProjects$project)) {
    data_project <- subset(RProjects, project == p)
    PI <- predictionInterval(thetao = data_project$fiso,
                             seo = data_project$se_fiso,
                             ser = data_project$se_fisr)
    ## transforming back to correlation scale
    PI <- tanh(PI)
    within <- (data_project$rr < PI$upper) & (data_project$rr > PI$lower)
    coverage <- mean(within)
    color <- ifelse(within == TRUE, "#333333B3", "#8B0000B3")
    study <- seq(1, nrow(data_project))
    plot(data_project$rr, study, col = color, pch = 20,
         xlim = c(-0.5, 1), xlab = expression(italic(r)[r]), ylab = "Study",
         main = paste0(p, ": ", round(coverage*100, 0), "% coverage"))
    arrows(PI$lower, study, PI$upper, study, length = 0.02, angle = 90, code = 3, col = color)
    abline(v = 0, lty = 3)
}
```



The criticism that this approach receives is that for studies which are underpowered, the prediction intervals will become very wide. This in turn can lead to very different effect estimates being compatible, *e. g.* even ones that go in the opposite direction, ultimately providing no information about the effect itself (which actually happens in some cases in the economics and philosophy data sets).

3.3 The sceptical p-value

Analysis The *sceptical p-value*, a new quantitative measure of replication success was recently proposed by Held (2020a). The *sceptical p-value* arises from combining the intrinsic credibility method (Matthews, 2001) with the prior-predictive check (Box, 1980). Specifically, using Bayes theorem in reverse, the prior distribution of the effect size can be determined such that conditional on the original study, the $(1 - \alpha)$ credible interval of the posterior distribution of the effect just includes zero. This prior corresponds to the objection of a sceptic who argues that the original finding is no longer significant if combined with a sufficiently sceptical prior. Replication success at level α is then achieved if the tail probability of the replication estimate under its prior predictive distribution is smaller than α, rendering the objection of the sceptic unrealistic.

The smallest level α at which replication success can be declared corresponds to the sceptical p-value, analogous to the duality of ordinary p-values and confidence intervals (for technical details, see the article).

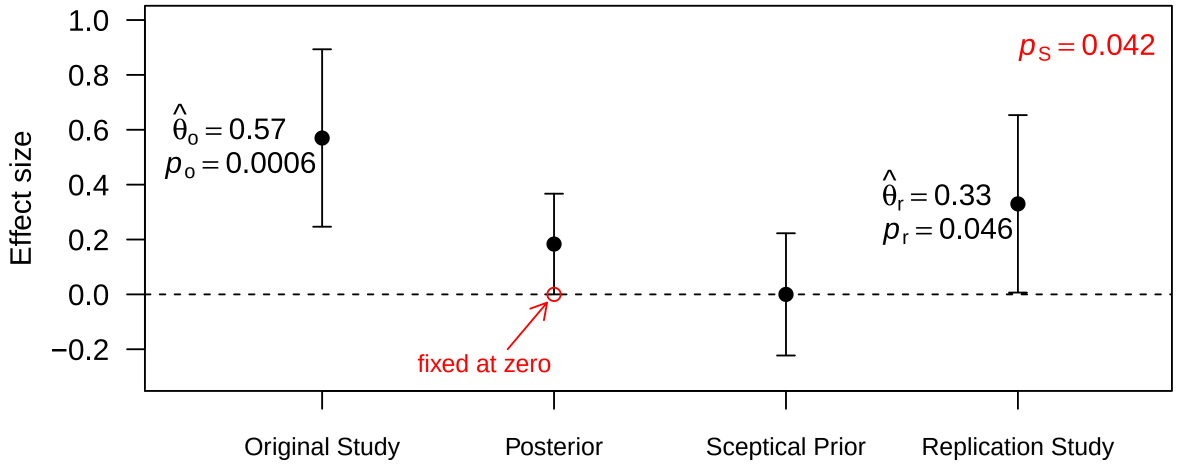


Figure 3: Example of assessment of replication success with one-sided sceptical p-value p_S.

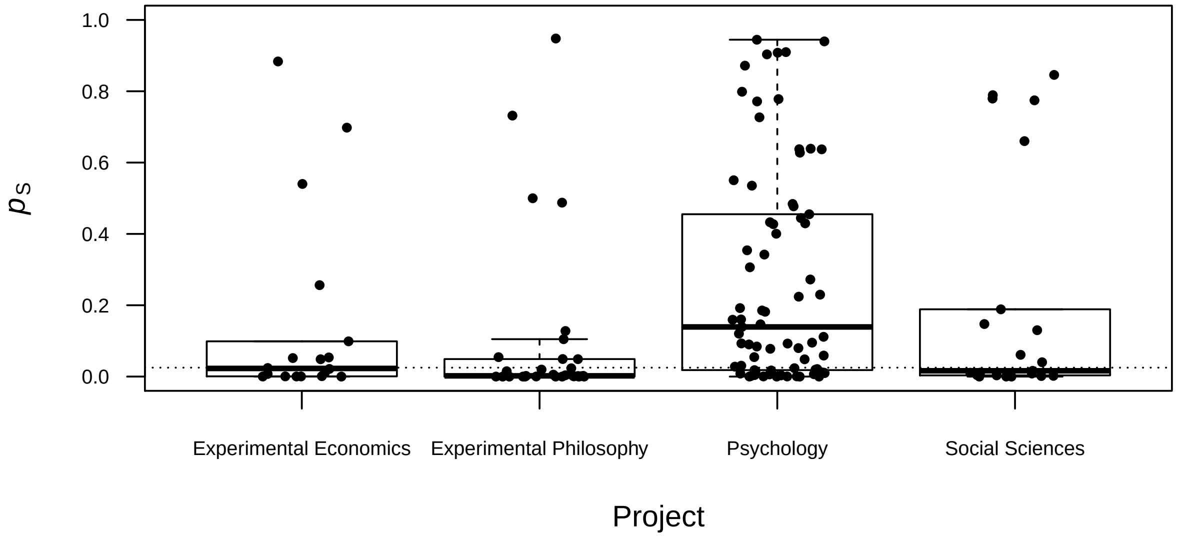
This method provides a theoretically sound approach to quantify replication success and it has attractive properties. In particular, the sceptical p-value is never smaller than the ordinary p-values from both studies and it also takes into account the size of the effect estimates, *i. e.* it becomes larger if the replication estimate is smaller than the original estimate. Held (2020b) further expanded on the calibration of the sceptical p-value. The function `thresholdSceptical` allows to compute different types of thresholds.

```
## computing nominal, controlled, liberal, and golden thresholds for one-sided sceptical p-value
(thresh_nom <- thresholdSceptical(level = 0.025, alternative = "one.sided",
                                  type = "nominal"))

## [1] 0.025

(thresh_contr <- thresholdSceptical(level = 0.025, alternative = "one.sided",
                                    type = "controlled"))

## [1] 0.06530883

(thresh_lib <- thresholdSceptical(level = 0.025, alternative = "one.sided",
                                  type = "liberal"))

## [1] 0.08288814

(thresh_gol <- thresholdSceptical(level = 0.025, alternative = "one.sided",
                                  type = "golden"))

## [1] 0.06167928
```

In particular, $\alpha_S = 0.062$ is the threshold based on the golden level (Held et al., 2020). This ensures that borderline original studies cannot lead to replication success if there is shrinkage of the replication effect estimate. The alternative controlled threshold $\alpha_S = 0.065$ is close in value and ensures one-sided type I error control at $0.025^2 = 0.000625$ if replication and original estimate have equal variances.

The sceptical p-value can be easily computed with the function `pSceptical`. For the analysis of replication studies, it is recommended to report the one-sided sceptical p-value and to apply a recalibration at the golden level (“`type = golden`”). We can then compare the sceptical p-value to the one-sided alpha level (0.025).

```
## computing one.sided sceptical p-value for replication projects
RProjects$ps <- with(RProjects,
                     pSceptical(zo = zo, zr = zr, c = c,
                                alternative = "one.sided", type="golden"))
boxplot(ps ~ project, data = RProjects, las = 1, cex.axis = 0.7, ylim = c(0, 1),
        xlab = "Project", ylab = expression(italic(p)[S]), outline = FALSE)
abline(h = alpha/2, lty = 3)
stripchart(ps ~ project, data = RProjects, vertical = TRUE, add = TRUE,
           pch = 19, method = "jitter", jitter = 0.2, cex = 0.6)
```



```
for (p in unique(RProjects$project)) {
    data_project <- subset(RProjects, project == p)
    cat(paste0(p, ": \n"))
    success_scept <- (data_project$ps < 0.025)
    cat(paste0(round(mean(success_scept)*100, 2),
               "% smaller than 0.025 (one-sided sceptical p-value) \n"))
    success_tradit <- (data_project$po1 < 0.025) & (data_project$pr1 < 0.025)
    cat(paste0(round(mean(success_tradit)*100, 2),
               "% smaller than 0.025 (both one-sided traditional p-values) \n"))
    if(sum(success_scept != success_tradit) > 0){
        discrep <- data_project[(success_scept != success_tradit),
                                c("ro", "rr", "c", "po1", "pr1", "ps")]
        ## print effect estimates, 1sided p-values, and c of discrepant studies
        cat("Discrepant studies: \n")
        print(signif(discrep, 2), row.names = FALSE)
  }
  cat("\n \n")
}

## Psychology:
## 30.14% smaller than 0.025 (one-sided sceptical p-value)
## 28.77% smaller than 0.025 (both one-sided traditional p-values)
## Discrepant studies:
##    ro   rr   c     po1      pr1    ps
##  0.20 0.25 2.6 0.02800 0.000047 0.024
##  0.56 0.40 0.6 0.00026 0.035000 0.017
##  0.35 0.15 2.7 0.00140 0.023000 0.031
##
##
## Experimental Economics:
## 55.56% smaller than 0.025 (one-sided sceptical p-value)
## 55.56% smaller than 0.025 (both one-sided traditional p-values)
##
##
## Social Sciences:
## 52.38% smaller than 0.025 (one-sided sceptical p-value)
## 61.9% smaller than 0.025 (both one-sided traditional p-values)
## Discrepant studies:
##    ro   rr   c    po1    pr1    ps
##  0.28 0.15 3.5 0.0089 0.0110 0.040
##  0.38 0.15 9.2 0.0110 0.0043 0.061
##
##
## Experimental Philosophy:
## 70.97% smaller than 0.025 (one-sided sceptical p-value)
## 74.19% smaller than 0.025 (both one-sided traditional p-values)
## Discrepant studies:
##    ro   rr   c   po1    pr1    ps
##  0.75 0.44 9.4 0.015 0.0006 0.049
##
##
```

We can see some discrepencies between the two approaches. In particular, the sceptical p-value may not indicate replication success when there is substantial shrinkage of the replication effect estimate relative to the original one, even if both estimates are significant.

Design Design works similarly as for the statistical significance analysis strategy; Using the function `sampleSizeReplicationSuccess`, one needs to choose a design prior, a sceptical p-value level, and the desired power to obain the required relative sample size $c = n_r/n_o$. The following code shows a few examples.

```
sampleSizeReplicationSuccess(zo = 2.5, power = 0.8, level = thresh_gol,
                             alternative = "one.sided",
                             designPrior = "conditional")

## [1] 1.377076

sampleSizeReplicationSuccess(zo = 2.5, power = 0.8, level = thresh_gol,
                             alternative = "one.sided",
                             designPrior = "predictive")

## [1] 2.776733
```

Figure 4 shows the power to achieve a one-sided sceptical p-value smaller or equal 0.062 as a function of the p-value or z-value of original study, assuming equal sample sizes in original and replication studies. The probability for replication success if the original study showed only weak evidence ($p_o = 0.05$) is now smaller than 0.5, which is reached for an original p-value of slightly above 0.03.

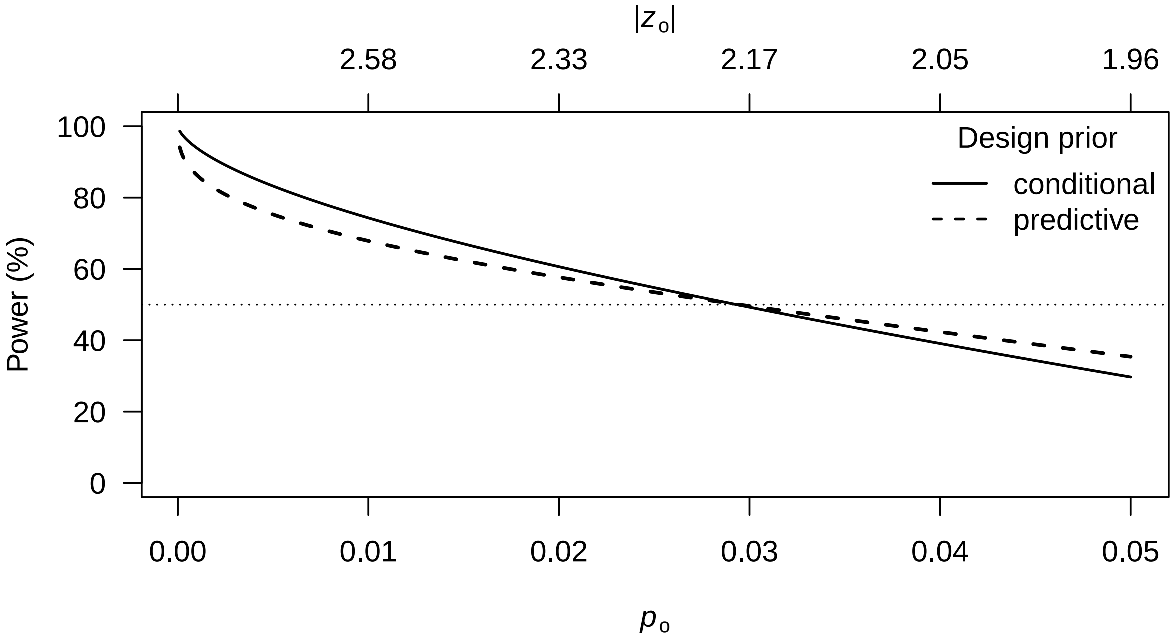


Figure 4: Power to achieve replication success (at the one-sided 0.062 level) as a function of the two-sided p-value or z-value of original study.

Figure 5 shows the required sample size to achieve a one-sided sceptical p-value of 0.062 with 80% power. The relative sample sizes consequently increase with increasing original p-value, with a dramatic increase for p-value larger than 0.023 when the predictive design prior is used.

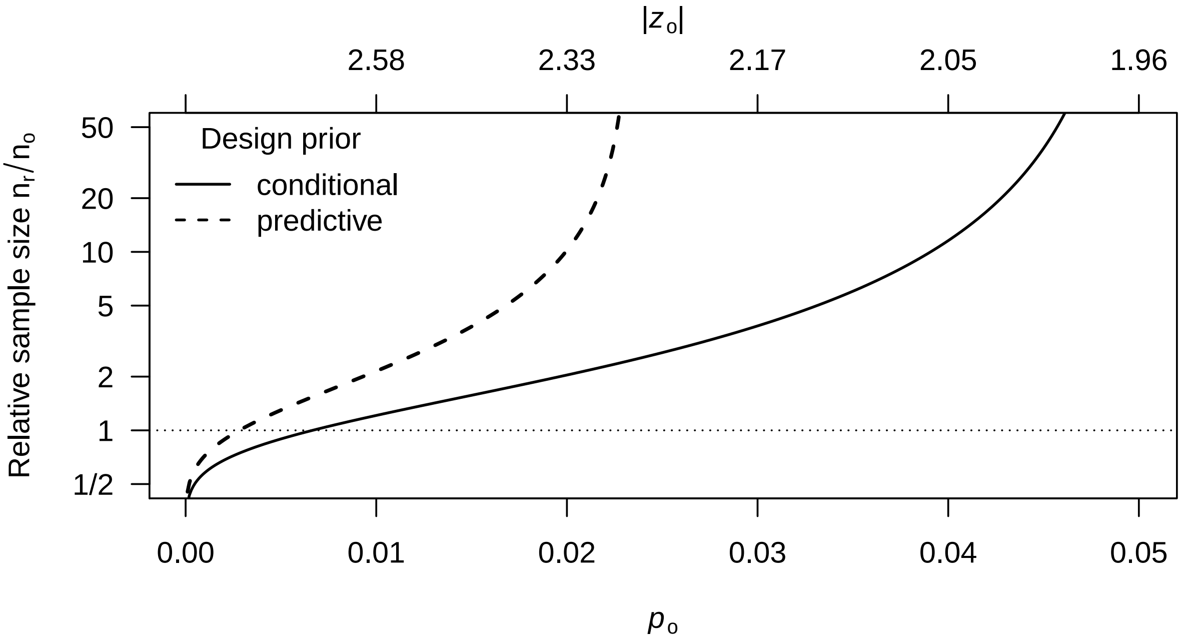


Figure 5: Relative sample size to achieve replication success (at the one-sided 0.062 level) with 80% power as a function of (two-sided) p-value or z-value of original study.

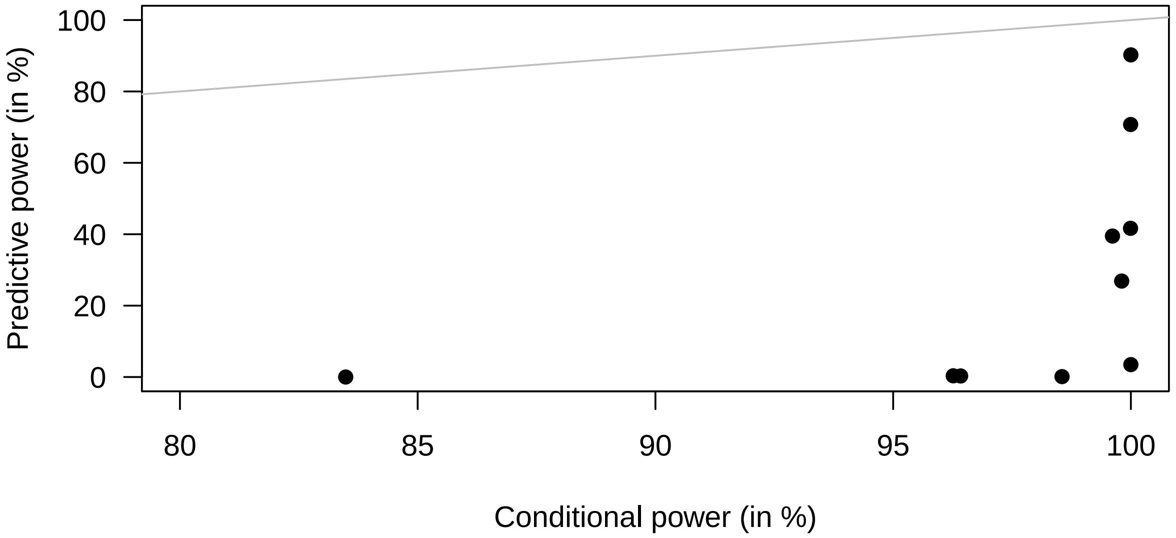


Figure 6: Conditional vs. predictive power at interim of the 10 studies from the social science replication project that were not stopped after stage 1. The grey line indicates the same value for conditional and predictive power.

3.4 Relative effect size

Analysis

Design

4 Special topics

4.1 Interim analysis

Adaptive designs are a type of designs where one or more interim analyses are planned during the course of a study. This topic has extensively been studied and used in clinical trials for example, where continuing a study that should be stopped may lead to serious consequences. However, this type of design has not be covered in the framework of replication studies. `ReplicationSuccess` allows to calculate the power of the replication study after an interim analysis has been performed, taking into account the results from the first part of the study. The function `powerSignificanceInterim` is an extension of `powerSignificance` and requires in addition the specification of `zi`, the z-value at the interim analysis and `f`, the fraction of the replication study already completed. Moreover, the argument `designPrior` can be set to `conditional`, `informed predictive` and `predictive`. Finally, the argument `analysisPrior` allows to also take the original result into account in the analysis of the replication study.

Figure 6 shows the conditional and the predictive power of the replication studies that continued into stage 2. While the condition power is larger than 80% for all the studies, the predictive power is close to 0% for some studies and always smaller than the conditional power.

4.2 Between-study heterogeneity

It is likely that the effect estimates from original and replication studies are not realizations of the exact same underlying effect size, but that there is between-study heterogeneity of effects. This can be caused, for example, if the replication study is conducted in a different laboratory with

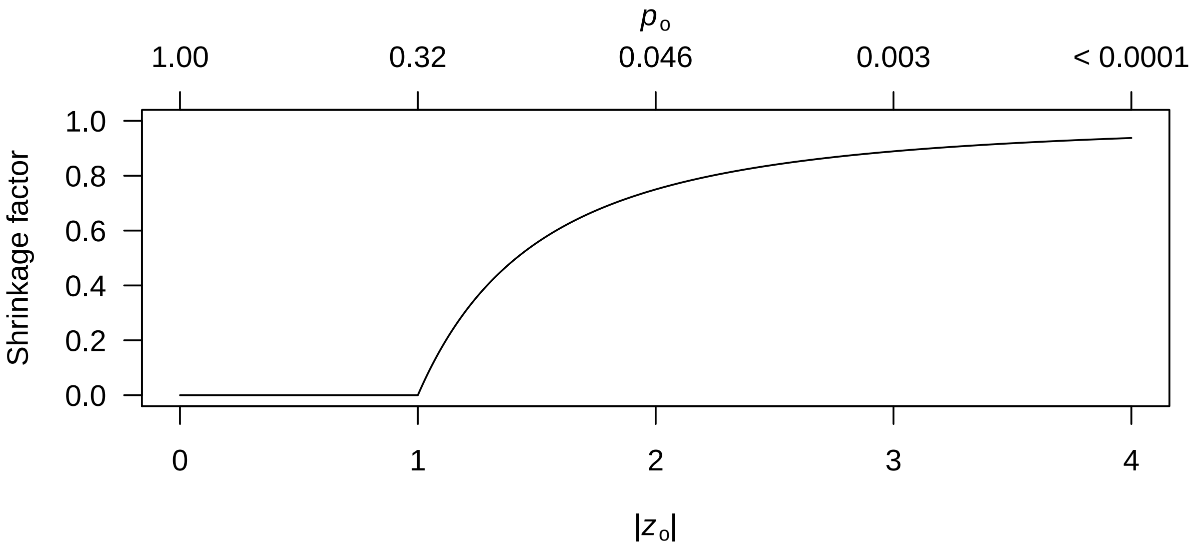


Figure 7: Empirical Bayes shrinkage when there is no between-study heterogeneity.

different equipment. For this reason, many functions in `ReplicationSuccess` allow to incorporate additionally uncertainty due to between-study heterogeneity into the predictive model. For example, `sampleSizeSignificance` or `predictionInterval` allow to specify `d`, the relative between-study heterogeneity variance $d = \tau^2/\sigma^2$, *i. e.* the ratio of the heterogeneity variance to the variance of the original effect estimate. By default, `d` is set to zero, however, if between-study heterogeneity is expected, *e. g.* a different population of study participants is used, this should be considered in the design. For details, see Pawel and Held (2020).

4.3 Data-driven shrinkage with empirical Bayes

As previously mentioned, the functions `sampleSizeSignificance` and `powerSignificance` allow to specify the argument `shrinkage`, in order to shrink the original effect estimate towards zero by a certain (arbitrary) amount. A more principled approach is to use a design prior which induces shrinkage and then estimate the prior variance by empirical Bayes. This leads to "data-driven" shrinkage that is larger when there was only weak evidence for the effect, and smaller when there was strong evidence for the effect (shown in Figure 7). Furthermore, under this prior, the specified between-study heterogeneity will also induce shrinkage towards zero, for details see Pawel and Held (2020). Empirical Bayes shrinkage is currently supported for the functions `sampleSizeSignificance`, `powerSignificance`, and `predictionInterval` by setting the design prior argument to `"EB"`.

References

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

C. F. Camerer, A. Dreber, E. Forsell, T. Ho, J. Huber, M. Johannesson, M. Kirchler, J. Almenberg, A. Altmejd, T. Chan, E. Heikensten, F. Holzmeister, T. Imai, S. Isaksson, G. Nave, T. Pfeiffer, M. Razen, and H. Wu. Evaluating replicability of laboratory experiments in economics. *Science*, 351:1433 – 1436, 2016. doi: 10.1126/science.aaf0918.

C. F. Camerer, A. Dreber, F. Holzmeister, T. Ho, J. Huber, M. Johannesson, M. Kirchler, G. Nave, B. Nosek, T. Pfeiffer, A. Altmejd, N. Buttrick, T. Chan, Y. Chen, E. Forsell, A. Gampa, E. Heikenstein, L. Hummer, T. Imai, S. Isaksson, D. Manfredi, J. Rose, E. Wagenmakers, and H. Wu. Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behavior*, 2:637 – 644, 2018. doi: 10.1038/s41562-018-0399-z.

Florian Cova, Brent Strickland, Angela Abatista, Aurélien Allard, James Andow, Mario Attie, James Beebe, Renatas Berniūnas, Jordane Boudesseul, Matteo Colombo, Fiery Cushman, Rodrigo Diaz, Noah N'Djaye Nikolai van Dongen, Vilius Dranseika, Brian D. Earp, Antonio Gaitán Torres, Ivar Hannikainen, José V. Hernández-Conde, Wenjia Hu, François Jaquet, Kareem Khalifa, Hanna Kim, Markus Kneer, Joshua Knobe, Miklos Kurthy, Anthony Lantian, Shen-yi Liao, Edouard Machery, Tania Moerenhout, Christian Mott, Mark Phelan, Jonathan Phillips, Navin Rambharose, Kevin Reuter, Felipe Romero, Paulo Sousa, Jan Sprenger, Emile Thalabard, Kevin Tobia, Hugo Viciana, Daniel Wilkenfeld, and Xiang Zhou. Estimating the reproducibility of experimental philosophy. *Review of Philosophy and Psychology*, 2018. doi: 10.1007/s13164-018-0400-9.

A. Dreber, T. Pfeiffer, S. Almenberg, Isaksson, J., B. Wilson, Y. Chen, B. A. Nosek, and M. Johannesson. Using prediction markets to estimate the reproducibility of scientific research. *PNAS*, 112:15343 – 15347, 2015. doi: 10.1073/pnas.1516179112.

R. A. Fisher. On the probable error of a coefficient of correlation deduced from a small sample. *Metron*, 1:3 – 32, 1921.

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

L. Held. A new standard for the analysis and design of replication studies (with discussion). *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 183(2):431 – 448, 2020a. doi: 10.1111/rssa.12493. URL https://doi.org/10.1111/rssa.12493.

L. Held, C. Micheloud, and S. Pawel. The assessment of replication studies based on relative effect size., 2020. Preprint.

Leonhard Held. The harmonic mean χ^2 test to substantiate scientific findings. *Journal of the Royal Statistical Society, Series C*, 2020b. https://arxiv.org/abs/1911.10633.

Valen E. Johnson, Richard D. Payne, Tianying Wang, Alex Asher, and Soutrik Mandal. On the reproducibility of psychological science. *Journal of the American Statistical Association*, 112(517): 1 – 10, 2016. doi: 10.1080/01621459.2016.1240079.

R. A. J. Matthews. Methods for assessing the credibility of clinical trial outcomes. *Drug Information Journal*, 35:1469 – 1478, 2001. doi: 10.1177/009286150103500442.

Open Science Collaboration. Estimating the reproducibility of psychological science. *Science*, 349 (6251):aac4716, 2015. doi: 10.1126/science.aac4716.

P. Patil, R. D. Peng, and J. T. Leek. What should researchers expect when they replicate studies? A statistical view of replicability in psychological science. *Perspectives on Psychological Science*, 11:539 – 544, 2016. doi: 10.1177/1745691616646366.

S. Pawel and L. Held. Probabilistic forecasting of replication studies. *PLOS ONE*, 15(4):e0231416, 2020. doi: 10.1371/journal.pone.0231416.