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

Package version 0.1.4

Abstract

This vignette provides an overview of the R package `ReplicationSuccess`. The package contains utilities for the design and analysis of replication studies. Traditional methods based on statistical significance and confidence intervals, as well as the recently developed reverse-Bayes approach from Held (2020b) 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 which statistical approach should be used to assess whether a replication study successfully replicated an original discovery. Moreover, depending on the chosen approach for analysis, 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 ways. Specifically, functions for analysis, power and samples size calculations based on statistical significance and confidence intervals, as well as on more recent methods, such as the sceptical p-value (Held, 2020b), are included. This vignette illustrates the usage of the package on the data sets from four large-scale replication projects which are also included in the package.

`ReplicationSuccess` was first created to provide software for computing the sceptical p-value and related power and sample size calculations (Held, 2020b). Methods to compute the p-value for intrinsic credibility (Held, 2019), the harmonic mean χ^2-test (Held, 2020a), forecasting of replication studies (Pawel and Held, 2020), and interim analyses of replication studies (Micheloud and Held, 2020) were added subsequently. Recently, substantial changes to many of the existing functions were made due to a recalibration of the sceptical p-value approach (Held et al., 2020). Use `news(package = "ReplicationSuccess")` to see a history of the changes. `ReplicationSuccess` has not yet been released on CRAN but can be installed from R-Forge by running the command `install.packages("ReplicationSuccess", repos = "http://R-Forge.R-project.org")` in an R-session.

1.1 Statistical framework

`ReplicationSuccess` assumes a simple but general statistical framework: The (suitably transformed) effect estimates $\hat{\theta}_o$ and $\hat{\theta}_r$ from original (subscript o) and replication study (subscript r) are assumed to be normally distributed around the unknown effect size θ. Their variances are equal to

their squared standard errors σ_o and σ_r which are assumed to be known. The same framework is also common in meta-analysis and can for example be applied to mean differences, odds ratios (log transformation), or correlation coefficients (Fisher z-transformation).

Many of the functions in the package take unitless quantities as input: the z-values $z_o = \hat{\theta}_o/\sigma_o$ and $z_r = \hat{\theta}_r/\sigma_r$, the relative effect size $d = \hat{\theta}_r/\hat{\theta}_o$ (or shrinkage $s = 1 - d$), and the variance ratio $c = \sigma_o^2/\sigma_r^2$. The squared standard errors are usually inversely proportional to the sample size of each study, $\sigma_o^2 = \kappa^2/n_o$ and $\sigma_r^2 = \kappa^2/n_r$ for some unit variance κ^2. The variance ratio can then be identified as the relative sample size $c = n_r/n_o$. This may require some adaptation for certain effect sizes. Computation of these quantities from real data will be illustrated below.

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 was 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

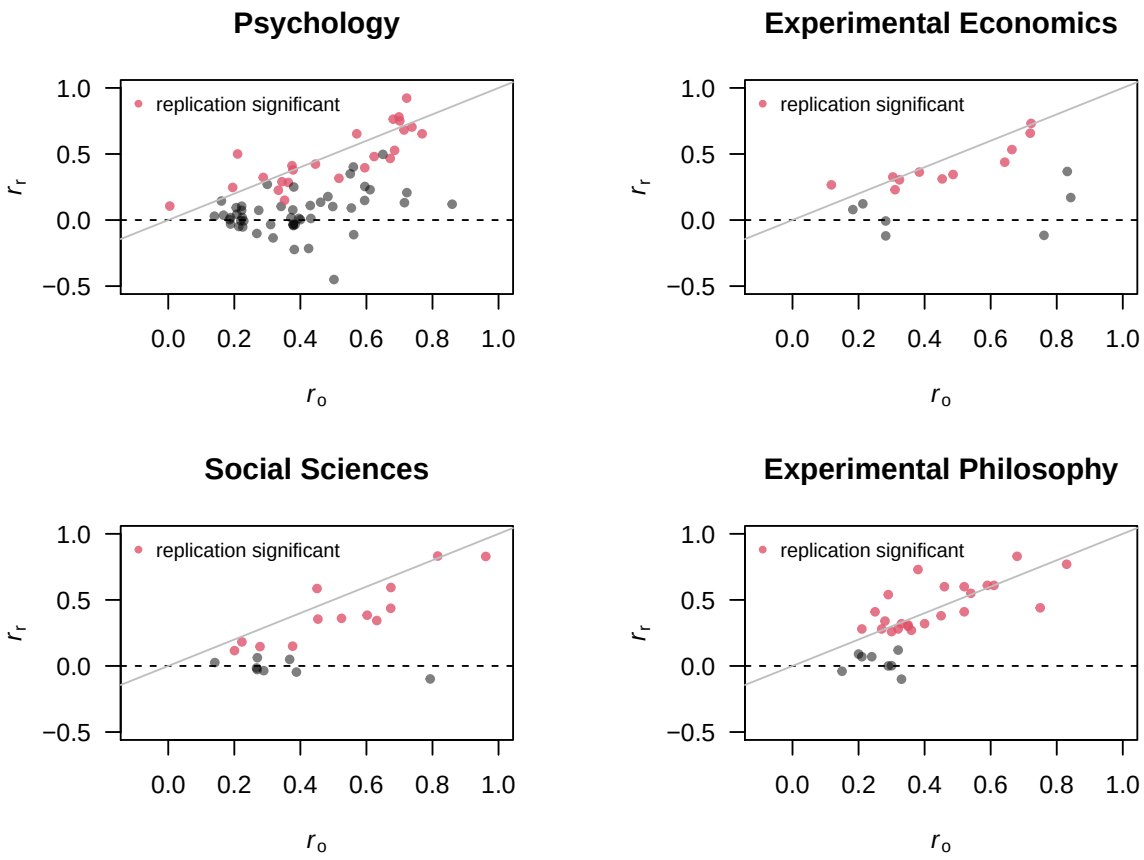
to compare and interpret. Design and statistical analysis, on the other hand, is then usually carried out on a scale where the estimates are approximately normally distributed. For correlation coefficients this is the case after applying the Fisher z-transformation, which leads to their variance asymptotically 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 paints the following picture.

```
## 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, "#DF536BCC", "#00000080")
    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 = "replication significant", cex = 0.8, pch = 20,
           col = "#DF536BCC", 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 has a substantial impact on the design of a replication study. In the design phase of a replication study, we will then focus only on the determination of the sample size.

3.1 Statistical significance

Analysis The most commonly used approach is to declare a replication successful if both, original and replication study, achieve statistical significance (in the same direction). For the four data sets, we can simply check whether the one-sided p-values (in the positive direction) of original and replication are both below the conventional threshold 0.025.

```
for (p in unique(RProjects$project)) {
    data_project <- subset(RProjects, project == p)
    significant_O <- data_project$po1 < 0.025
    significant_R <- data_project$pr1 < 0.025
    success <- significant_O & significant_R
    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, however, they can also be caused due to low power of the replication study (Goodman, 1992). On the other hand, statistical significance can still be achieved for a replication effect estimate which is much smaller than the one from the original study, provided its standard error is small enough (*e. g.* because of a very large replication 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 social 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`. Note that the function `sampleSizeSignificance`, as well as most of the functions from the package, takes z-values (and not p-values) as arguments. The transformation from p- to z-values and vice versa can easily be done using the functions `p2z` and `z2p`.

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

```r
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 (one-sided) p-value or the z-value of the original study. If the original estimate was just significant at the 0.025 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).

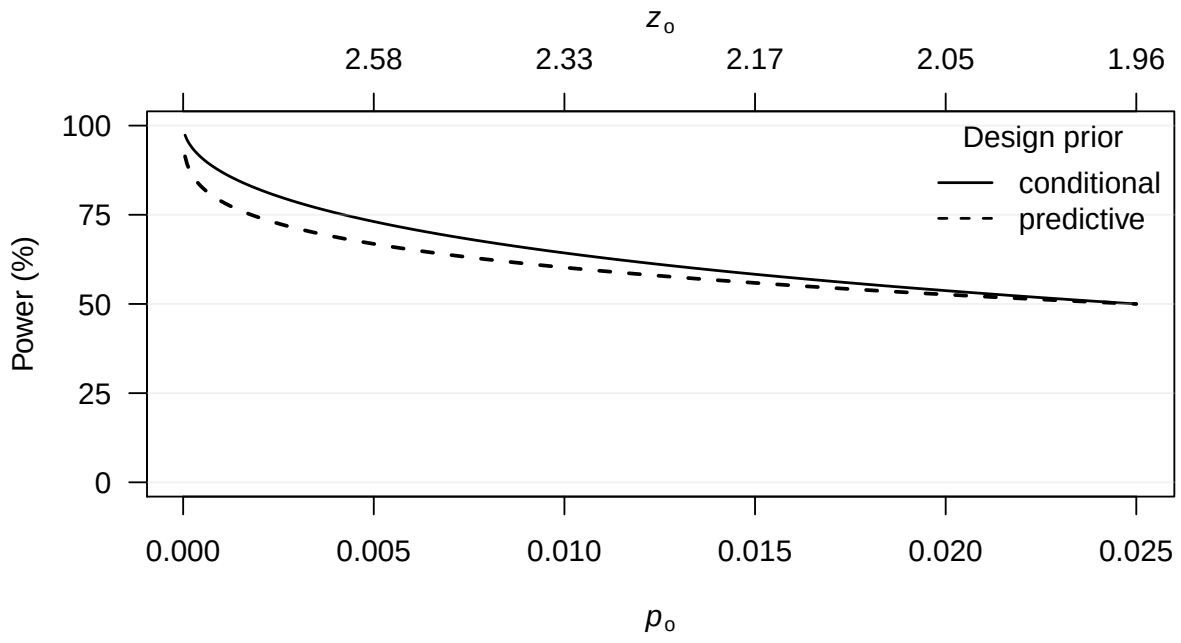


Figure 1: Power to achieve significance at the one-sided 2.5% level in replication as a function of (one-sided) p-value or z-value of the original study using the same sample size as in the original study.

3.2 Compatibility of effect size

Analysis Another way for analysing a replication study is to examine the statistical compatibility between original and replication effect estimate. A popular approach is to check whether the replication estimate is contained within a prediction interval based on the original estimate (Patil et al., 2016; Pawel and Held, 2020). This approach based on a $(1-\alpha)$ prediction interval is in fact equivalent to conducting a meta-analytic Q-test with the two estimates, and rejecting compatibility

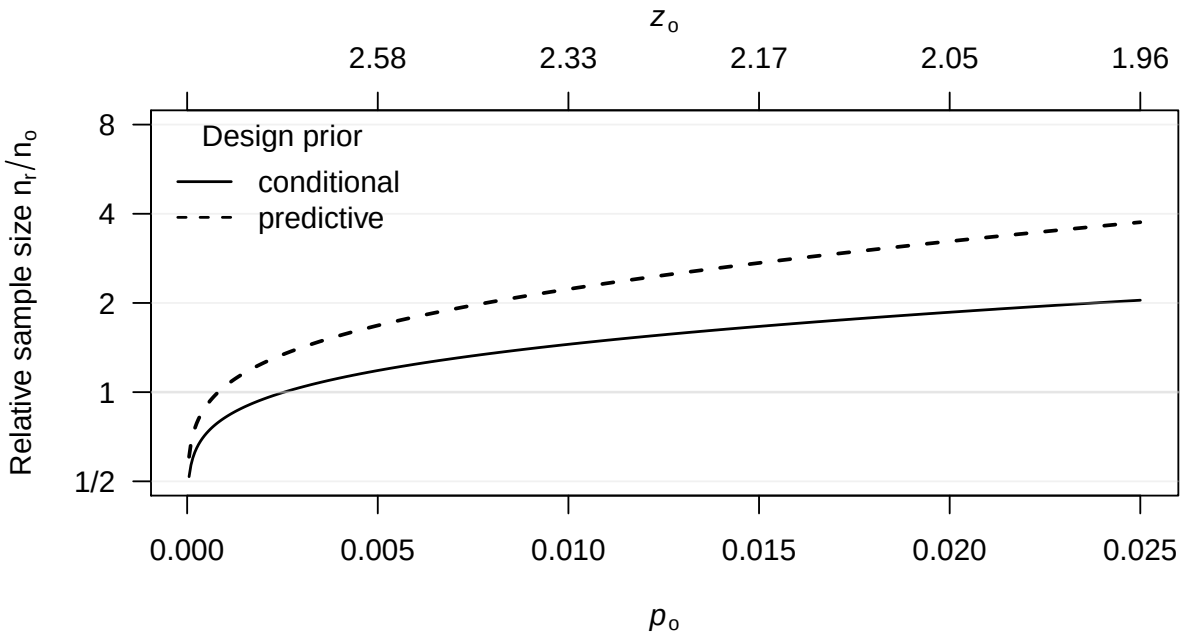


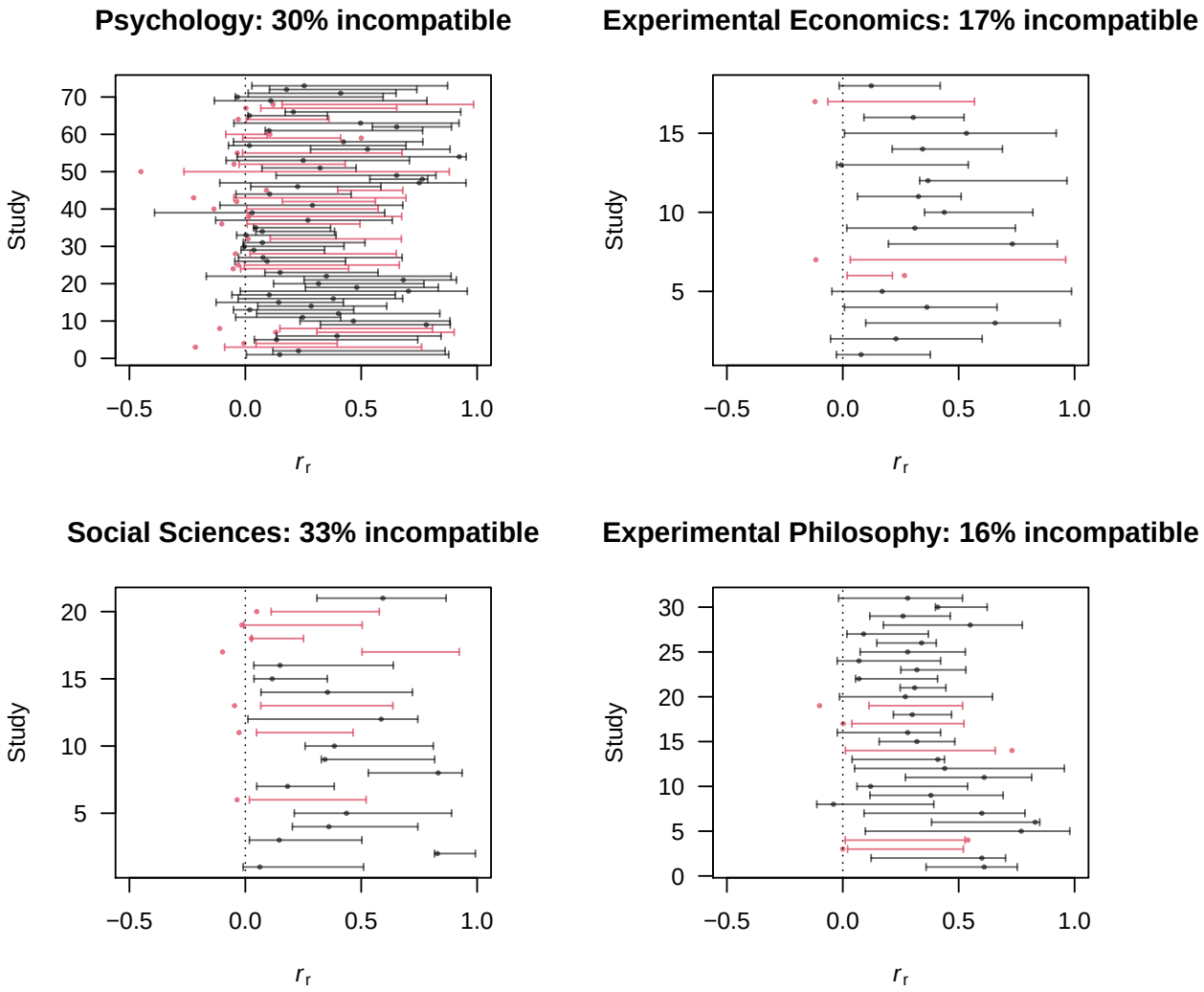
Figure 2: Relative sample size to achieve significance at the one-sided 2.5% level with 80% power as a function of (one-sided) p-value or z-value of original study.

when the corresponding p-value $p_Q < \alpha$ (Hedges and Schauer, 2019). The p-value from the Q-test is usually preferred, since it tells quantitatively how compatible the estimates. In contrast, a prediction interval can give a better idea about the range of plausible replication effect estimates besides the observed one. Both approaches are available in `ReplicationSuccess`: The function `Qtest` returns the p-value from the meta-analytic Q-test, whereas the function `predictionInterval` returns a prediction interval for the replication effect based on the original counterpart (see the documentation of the functions for further details).

For the four data sets, we can easily compute p-values from Q-test, as well as 95% prediction intervals. For easier visual assessment we transform the intervals and estimates back to the correlation scale.

```
## compute prediction intervals for replication projects
par(mfrow = c(2, 2), las = 1, mai = rep(0.65, 4))
for (p in unique(RProjects$project)) {
    data_project <- subset(RProjects, project == p)
    pQ <- Qtest(thetao = data_project$fiso,
                thetar = data_project$fisr,
                seo = data_project$se_fiso,
                ser = data_project$se_fisr)
    PI <- predictionInterval(thetao = data_project$fiso,
                             seo = data_project$se_fiso,
                             ser = data_project$se_fisr)
    ## transforming back to correlation scale
    PI <- tanh(PI)
    incompatible <- pQ < 0.05
    color <- ifelse(incompatible == FALSE, "#00000099", "#DF536BCC")
    study <- seq(1, nrow(data_project))
    plot(data_project$rr, study, col = color, pch = 20, cex = 0.5,
         xlim = c(-0.5, 1), xlab = expression(italic(r)[r]), ylab = "Study",
         main = paste0(p, ": ", round(mean(incompatible)*100, 0), "% incompatible"))
    arrows(PI$lower, study, PI$upper, study, length = 0.02, angle = 90, code = 3, col = color)
```

```
    abline(v = 0, lty = 3)
}
```



While both approaches enable statements about compatibility of original and replication effect estimates, they carry a fundamental problem due to the structure of the underlying hypothesis test: If $p_Q < \alpha$ (or equivalently the $(1-\alpha)$ prediction interval does not contain the replication estimate) one has established incompatibility/non-replication. If, however, one fails to establish incompatibility, it remains unclear whether this is due to small power or true compatibility of both estimates (Hedges and Schauer, 2019).

3.3 The sceptical p-value

Analysis The *sceptical p-value*, a new quantitative measure of replication success was first introduced in Held (2020b). Conceptually, replication success is declared if the replication study is in conflict with a sceptical prior that would render the original study non-significant. 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 is determined such that based 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. See Figure 3 for a practical illustration of the procedure, also see Held (2020b); Held et al. (2020) for technical details.

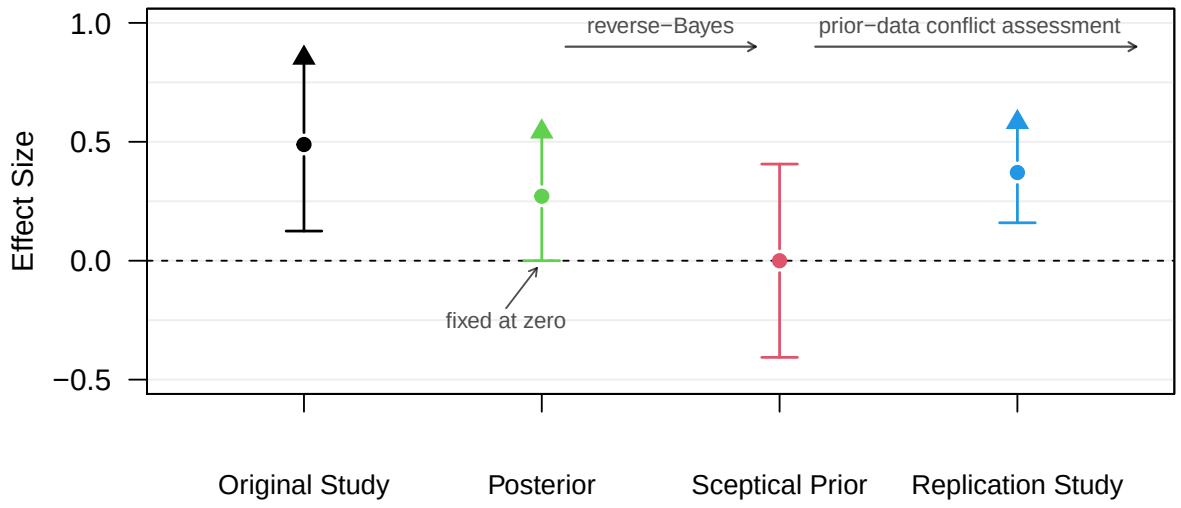


Figure 3: Example of assessment of replication success. The original study from Morewedge et al. (2010) lead to the effect estimate $\hat{\theta}_o = 0.49$ on Fisher's z-scale (95% CI from 0.12 to 0.85) and one-sided p-value $p_o = 0.0043$. The left part of the Figure illustrates the reverse-Bayes derivation of the sufficiently sceptical prior based on the original study result and the posterior with lower credible limit fixed at zero. The right part of the Figure shows the comparison of the sufficiently sceptical prior with the replication study result ($\hat{\theta}_r = 0.37$, 95% CI from 0.16 to 0.58, $p_r = 0.0003$) from the SSRP (Camerer et al., 2018). The one-sided sceptical p-value is $p_S = 0.036$ for this example.

The method has attractive properties: The sceptical p-value is never smaller than the ordinary p-values from both studies which ensures that both studies have to be sufficiently convincing on their own such that replication success is possible at a reasonable level. 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, which guarantees that shrinkage of the replication effect estimate is penalized.

The sceptical p-value can be easily computed with the function `pSceptical`. It is recommended to report the one-sided sceptical p-value because it guarantees that replication success is only possible if the directions of original and replication effect estimates are the same. For the replication study of Morewedge et al. (2010) in Figure 3, the one-sided sceptical p-value turns out to be $p_S = 0.036$.

Interpretation
Interpretation of the sceptical p-value as a continuous measure of replication success is recommended. However, if an answer to "did it replicate?" is needed, several types of levels to threshold the sceptical p-value are available and can be computed with the function `levelSceptical`.

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

## [1] 0.06167928

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

```
## [1] 0.06530883

(thresh_nom <- levelSceptical(level = 0.025, alternative = "one.sided",
                              type = "nominal"))

## [1] 0.025

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

## [1] 0.08288814
```

The four types can be summarized as follows:

- *golden*: Ensures that for original studies, which are just significant at the specified significance level α, replication success is only possible if the replication effect estimate is larger than the original one (*i. e.* if there is no shrinkage).
- *controlled*: When original and replication effect estimate have the same variance ($c = 1$), this type ensures type-I error control at α^2 for `alternative = "two.sided"` and at $2 \times \alpha^2$ for `alternative = "one.sided"` (the conventional level for type I error control of two independent experiments with two-sided testing at level $2 \times \alpha$ in the first and one-sided testing at level α in the second). When the variance of the replication effect estimate is smaller ($c > 1$), the type-I error decreases further.
- *nominal*: Ensures that the type-I error is always smaller than α^2. Significance of both the original and replication study at level α is then a necessary but not sufficient requirement for replication success.
- *liberal*: Ensures that significance of both studies at level α is a sufficient requirement for replication success if original and replication effect estimate have the same variance ($c = 1$).

Held (2020b) only used the nominal level and it led to relatively low rates of replication success in the case studies considered. Held et al. (2020) further studied properties of the different levels and concluded that the nominal level may be too stringent for most realistic scenarios, instead the golden level was recommended as default choice to assess replication success. It provides an attractive balance between significance testing and effect size comparison: If a replication study is not significant, it can still achieve replication success, provided the replication effect estimate does not shrink. For the one-sided sceptical p-value and $\alpha = 0.025$, the golden level turns out to be 0.062. In the Morewedge et al. (2010) replication example, we have $p_S = 0.036 < 0.062$, so the replication is successful at the golden level. A side-remark: The controlled threshold of 0.065 is slightly smaller than the value of the golden level 0.062, which means that the golden level controls the type-I error rate for $c \geq 1$.

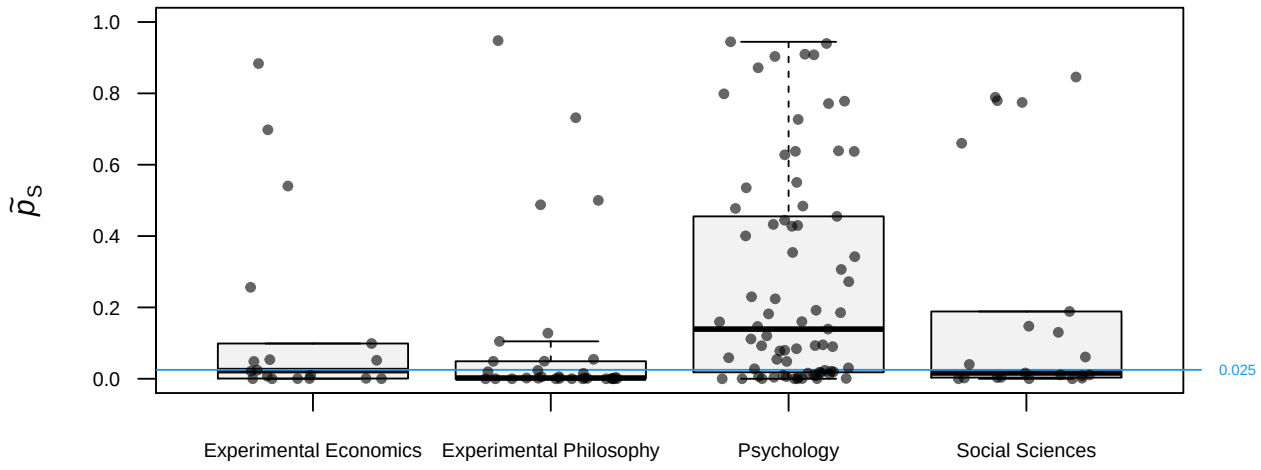
Recalibration

In practice it is easier to compute a recalibrated sceptical p-value $\tilde{p}_S$ and compare it to the significance level (*e. g.* $\tilde{p}_S < 0.025$) instead of comparing the uncalibrated p_S to a level computed by the `levelSceptical` function (*e. g.* $p_S < 0.062$). This can be accomplished with the argument `type` in `pSceptical`, for instance with `type = "golden"`. For the Morewedge et al. (2010) example, the sceptical p-value recalibrated at the golden level turns out to be $\tilde{p}_S = 0.011 < 0.025$, from which we can also see that the replication is successful at the golden level.

The one-sided recalibrated sceptical p-values at the golden level are computed for the four projects as follows:

```
## 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(tilde(p))[S]), outline = FALSE,
        col = "#0000000D")
abline(h = alpha/2, lty = 1, col = 4)
axis(side = 4, at = alpha/2, col.axis = 4, col = 4, las = 1, cex.axis = 0.5)
stripchart(ps ~ project, data = RProjects, vertical = TRUE, add = TRUE,
           pch = 20, method = "jitter", jitter = 0.3, cex = 1, col = "#00000099")
```




```
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
##
##
```

For many studies, replication success is declared both based on significance and on the sceptical p-value. However, six replications show discrepant results. 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 (this is the case for one study in the psychology project, two studies in the social sciences project, one study in the philosophy project). On the other hand, provided there is not much shrinkage, it may indicate replication success for non-significant original or replication studies (this is the case for two studies in the psychology project).

Design Sample size calculations work in a similar manner as when they are based on statistical significance: Using the function `sampleSizeReplicationSuccess`, we need to choose a design prior, a replication success level, recalibration type, and the desired power to obtain the required relative sample size $c = n_r/n_o$. The following code shows two examples.

```
sampleSizeReplicationSuccess(zo = 2.5, power = 0.8, level = 0.025,
                             alternative = "one.sided",
                             designPrior = "conditional",
                             type = "golden")
```

```
## [1] 1.377036
```

```
sampleSizeReplicationSuccess(zo = 2.5, power = 0.8, level = 0.025,
                             alternative = "one.sided",
                             designPrior = "predictive",
                             type = "golden")
```

```
## [1] 2.776701
```

Figure 4 shows the power to achieve replication success at the golden level as a function of the one-sided p-value (or z-value) of the original study, assuming equal sample sizes in original and replication studies. The probability for replication success if the original study showed only weak evidence ($p_o = 0.025$) is now smaller than 0.5, which is reached for an original p-value of around 0.015. Figure 5 shows the required sample size to achieve replication success at the golden level

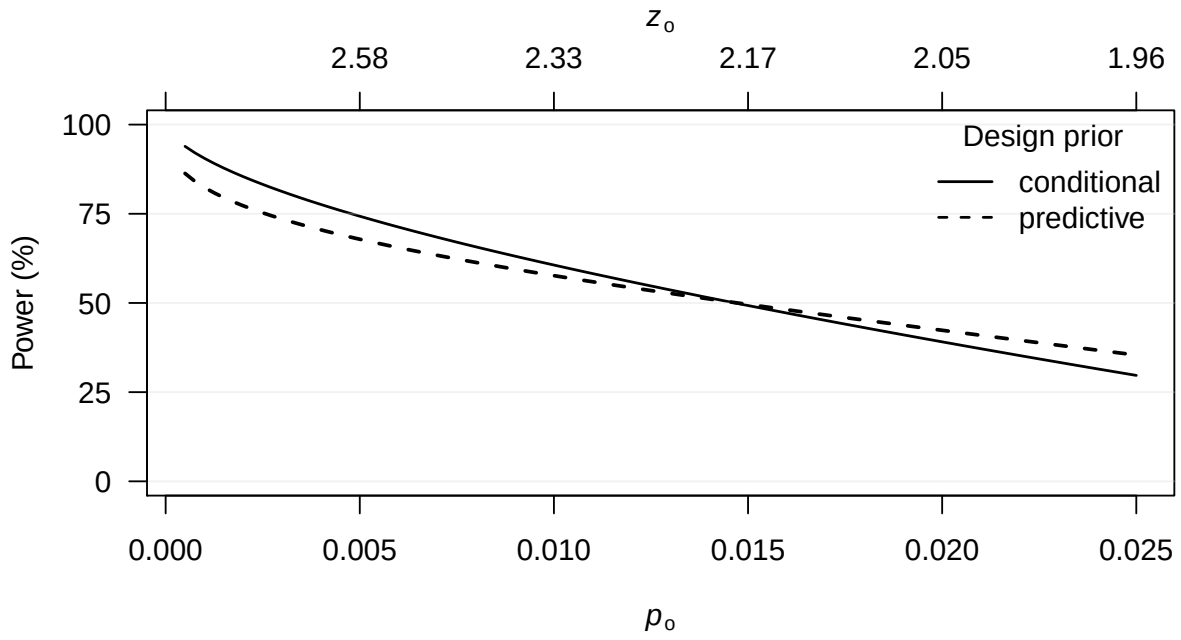


Figure 4: Power to achieve replication success (at the golden level with $\alpha = 0.025$) as a function of the one-sided p-value or z-value of the original study.

with 80% power. The relative sample sizes consequently increase with increasing original p-value, with a dramatic increase for p-value larger than 0.012 when the predictive design prior is used.

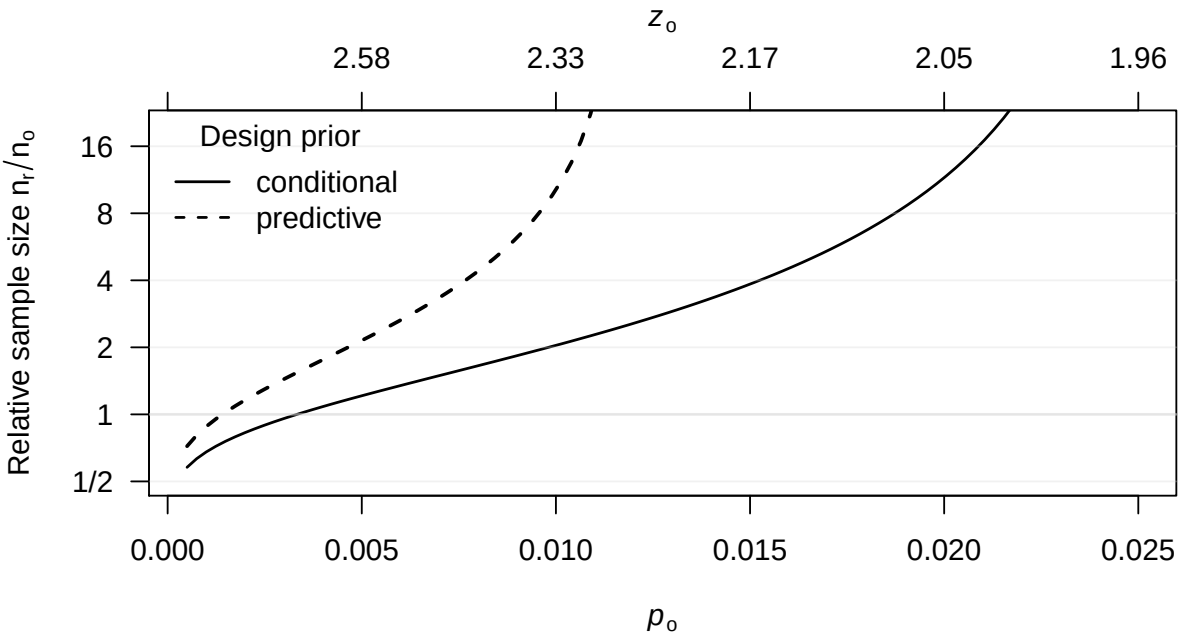


Figure 5: Relative sample size to achieve replication success (at the golden level with $\alpha = 0.025$) with 80% power as a function of the (one-sided) p-value or z-value of the original study.

3.4 Relative effect size

Analysis The requirements on the p-value of statistical significance of the replication study ($p_r < \alpha$) and of replication success ($p_S < \alpha_S$) can both be transformed to requirements on the relative effect size $d = \hat{\theta}_r/\hat{\theta}_o$ (Held et al., 2020). In short, significance of the replication study (or replication success) is guaranteed if the relative effect size is larger than a certain bound (the minimum relative effect size $d_{\min}$). The minimum relative effect size $d_{\min}$ can be computed based on the result from the original study, the relative sample size and the level (and additionally the type of recalibration for replication success). The bound $d_{\min}$ on the relative effect size d can be calculated using `effectSizeSignificance` and `effectSizeReplicationSuccess` for significance and replication success, respectively.

For the Morewedge et al. (2010) replication, the minimum relative effect size to achieve replication success at the golden level turns out to be $d_{\min} = 0.54$. Since the observed relative effect size is $d = 0.76 > 0.54$, the replication is deemed successful.

Design The methodology described above can be inverted such that the relative sample size is computed based on the specification of the minimum relative effect size. Using the functions `sampleSizeSignificance` and `sampleSizeReplicationSuccess`, it is straightforward to plan the sample size of the replication study with the approaches described above. The only thing that changes is that we need to supply the minimum relative effect size $d_{\min}$ to achieve replication success (instead of the power).

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 rarely been covered in the framework of replication studies. An adaptive design was adopted in the social sciences replication project, but without a power (re)calculation at interim. `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 power at interim is a useful tool to decide whether a replication study should be stopped prematurely for futility (Micheloud and Held, 2020). 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. The function `sampleSizeSignificanceInterim` can be used to re-estimate the sample size of the replication study when results from an interim analysis are available.

4.2 Between-study heterogeneity

It is often more realistic to assume that the effects underlying the estimates from original and replication studies are not exactly the same but that there is between-study heterogeneity. This can be caused, for example, if the replication study is conducted in a different laboratory with 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 `h`, the relative between-study heterogeneity variance $h = \tau^2/\sigma^2$, *i. e.* the ratio of the heterogeneity variance to the variance of the original effect estimate. By default, `h` 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. See also Pawel and Held (2020) for more details.

4.3 Data-driven shrinkage with empirical Bayes

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 6). 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"`.

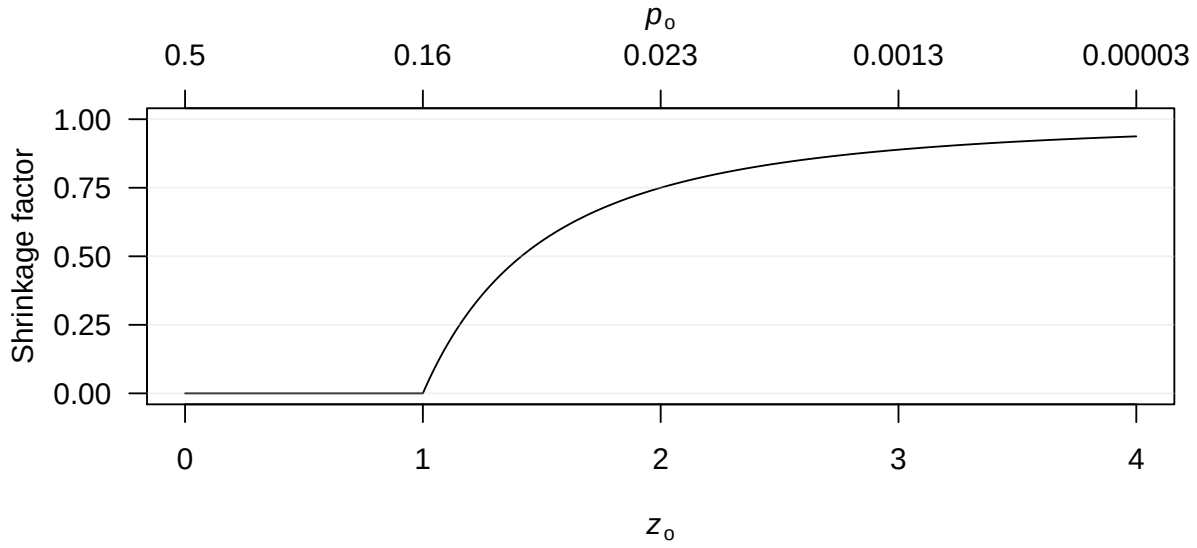


Figure 6: Empirical Bayes shrinkage when there is no between-study heterogeneity.

5 Outlook

Development on `ReplicationSuccess` will continue and we plan to submit a stable version to CRAN at some point in the future. We invite anyone with ideas for new functionality, bug-reports, or other contributions to the package to get in touch with us.

References

Box, G. E. P. (1980). Sampling and Bayes' inference in scientific modelling and robustness (with discussion). *Journal of the Royal Statistical Society, Series A*, 143:383–430. doi:10.2307/2982063.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T., 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. doi:10.1126/science.aaf0918.

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

Cova, F., Strickland, B., Abatista, A., Allard, A., Andow, J., Attie, M., Beebe, J., Berniūnas, R., Boudesseul, J., Colombo, M., Cushman, F., Diaz, R., van Dongen, N. N. N., Dranseika, V., Earp, B. D., Torres, A. G., Hannikainen, I., Hernández-Conde, J. V., Hu, W., Jaquet, F., Khalifa, K., Kim, H., Kneer, M., Knobe, J., Kurthy, M., Lantian, A., yi Liao, S., Machery, E., Moerenhout, T., Mott, C., Phelan, M., Phillips, J., Rambharose, N., Reuter, K., Romero, F., Sousa, P., Sprenger, J., Thalabard, E., Tobia, K., Viciana, H., Wilkenfeld, D., and Zhou, X. (2018). Estimating the reproducibility of experimental philosophy. *Review of Philosophy and Psychology*. doi:10.1007/s13164-018-0400-9.

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

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

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

Hedges, L. V. and Schauer, J. M. (2019). More than one replication study is needed for unambiguous tests of replication. *Journal of Educational and Behavioral Statistics*, 44(5):543–570. doi:10.3102/1076998619852953.

Held, L. (2019). The assessment of intrinsic credibility and a new argument for $p < 0.005$. *Royal Society Open Science*, 6(3):181534. doi:10.1098/rsos.181534.

Held, L. (2020a). The harmonic mean χ^2-test to substantiate scientific findings. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 69(3):697–708. doi:10.1111/rssc.12410.

Held, L. (2020b). 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. doi:10.1111/rssa.12493.

Held, L., Micheloud, C., and Pawel, S. (2020). The assessment of replication success based on relative effect size. URL https://arxiv.org/abs/2009.07782.

Johnson, V. E., Payne, R. D., Wang, T., Asher, A., and Mandal, S. (2016). On the reproducibility of psychological science. *Journal of the American Statistical Association*, 112(517):1–10. doi:10.1080/01621459.2016.1240079.

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

Micheloud, C. and Held, L. (2020). Power calculations for replication studies. URL https://arxiv.org/abs/2004.10814.

Morewedge, C. K., Huh, Y. E., and Vosgerau, J. (2010). Thought for food: Imagined consumption reduces actual consumption. *Science*, 330(6010):1530–1533. doi:10.1126/science.1195701.

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

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

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