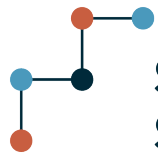
Design and Analysis of Replication Studies

Leonhard Held

University of Zurich



Swiss National
Science Foundation

February 2022

Introduction

Replication studies

Direct replication

- Repeating original study using the same methodology

$\rightarrow$ Tool to assess credibility of scientific discoveries

$\rightarrow$ Regulatory requirement

Replication studies

Direct replication

- Repeating original study using the same methodology

→ Tool to assess credibility of scientific discoveries

→ Regulatory requirement

Replication crisis

- Low replicability of many scientific discoveries

→ Large-scale replication projects

Large-scale replication projects

– 2015: Reproducibility project psychology

Science

Estimating the reproducibility of psychological science

Open Science Collaboration

Science **349** (6251), aac4716.
DOI: 10.1126/science.aac4716

Large-scale replication projects

- 2015: Reproducibility project psychology
- 2016: Experimental economics replication project

Science

REPORTS

Cite as: Camerer *et al.*, *Science*
10.1126/science.aaf0918 (2016).

Evaluating replicability of laboratory experiments in economics

Colin F. Camerer,[1*†] Anna Dreber,[2†] Eskil Forsell,[2†] Teck-Hua Ho,[3,4†] Jürgen Huber,[5†] Magnus Johannesson,[2†] Michael Kirchler,[5,6†] Johan Almenberg,[7] Adam Altmejd,[2] Taizan Chan,[8] Emma Heikensten,[2] Felix Holzmeister,[5] Taisuke Imai,[1] Siri Isaksson,[2] Gideon Nave,[1] Thomas Pfeiffer,[9,10] Michael Razen,[5] Hang Wu[4]

Large-scale replication projects

- 2015: Reproducibility project psychology
- 2016: Experimental economics replication project
- 2018: Experimental philosophy replicability project

Rev.Phil.Psych.
https://doi.org/10.1007/s13164-018-0400-9

CrossMark

Estimating the Reproducibility of Experimental Philosophy

Florian Cova[1,2] · Brent Strickland[3,4] · Angela Abatista[5] · Aurélien Allard[6] · James Andow[7] · Mario Attie[8] · James Beebe[9] · Renatas Berniūnas[10] · Jordane Boudesseul[11] · Matteo Colombo[12] · Fiery Cushman[13] · Rodrigo Diaz[14] · Noah N'Djaye Nikolai van Dongen[15] · Vilius Dranseika[16] · Brian D. Earp[17] · Antonio Gaitán Torres[18] · Ivar Hannikainen[19] · José V. Hernández-Conde[20] · Wenjia Hu[21] · François Jaquet[1] · Kareem Khalifa[22] · Hanna Kim[23] · Markus Kneer[24] · Joshua Knobe[25] · Miklos Kurthy[26] · Anthony Lantian[27] · Shen-yi Liao[28] · Edouard Machery[29] · Tania Moerenhout[30] · Christian Mott[25] · Mark Phelan[21] · Jonathan Phillips[13] · Navin Rambharose[21] · Kevin Reuter[31] · Felipe Romero[15] · Paulo Sousa[32] · Jan Sprenger[33] · Emile Thalabard[34] · Kevin Tobia[25] · Hugo Viciana[35] · Daniel Wilkenfeld[29] · Xiang Zhou[36]

Large-scale replication projects

- 2015: Reproducibility project psychology
- 2016: Experimental economics replication project
- 2018: Experimental philosophy replicability project
- 2018: Social sciences replication project

nature human behaviour

Letter | Published: 27 August 2018

Evaluating the replicability of social science experiments in *Nature* and *Science* between 2010 and 2015

Colin F. Camerer, Anna Dreber, Felix Holzmeister, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Gideon Nave, Brian A. Nosek ✉, Thomas Pfeiffer, Adam Altmejd, Nick Buttrick, Taizan Chan, Yiling Chen, Eskil Forsell, Anup Gampa, Emma Heikensten, Lily Hummer, Taisuke Imai, Siri Isaksson, Dylan Manfredi, Julia Rose, Eric-Jan Wagenmakers & Hang Wu

Large-scale replication projects

- 2015: Reproducibility project psychology
- 2016: Experimental economics replication project
- 2018: Experimental philosophy replicability project
- 2018: Social sciences replication project

nature human behaviour

Letter | Published: 27 August 2018

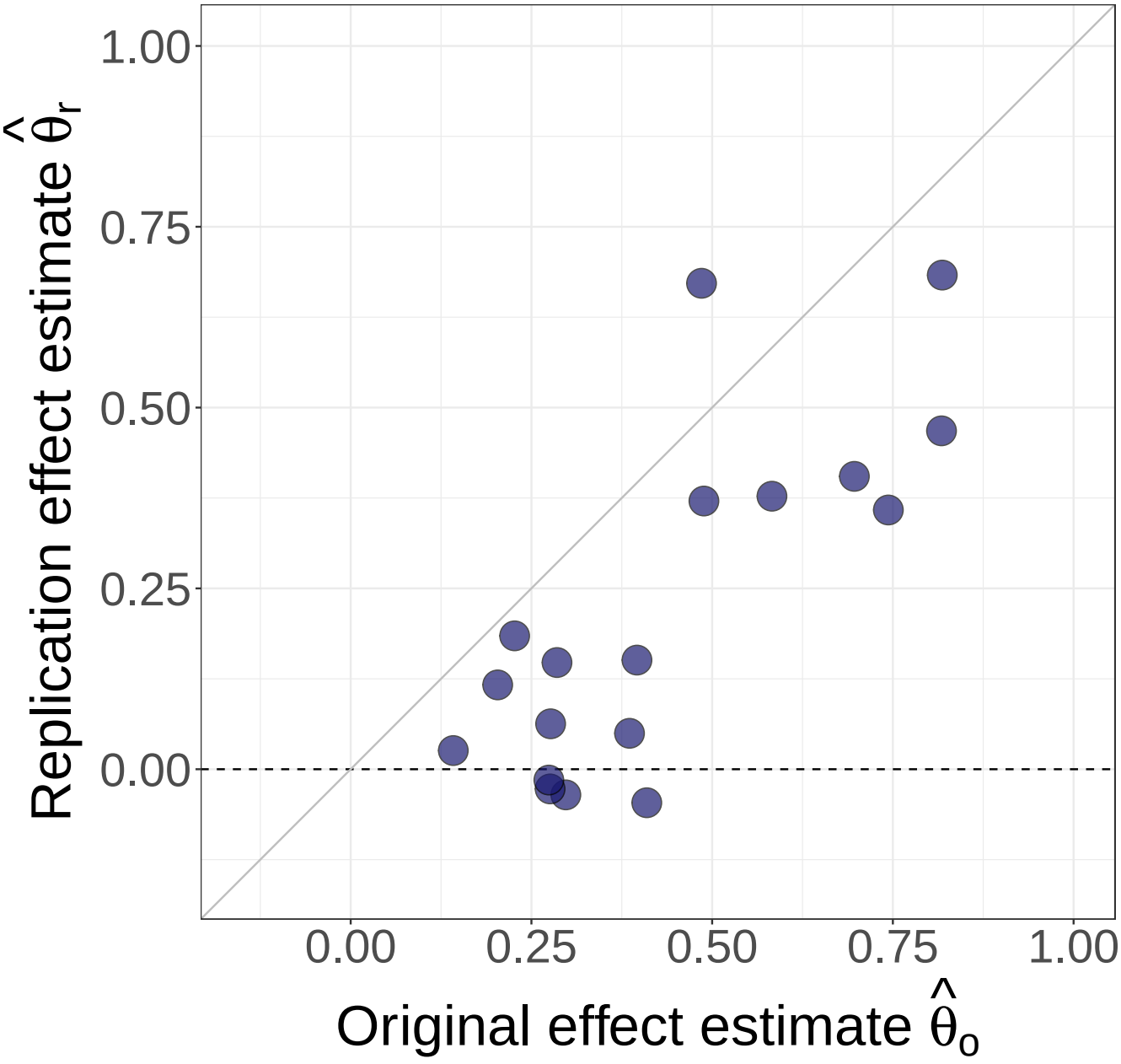
Evaluating the replicability of social science experiments in *Nature* and *Science* between 2010 and 2015

Colin F. Camerer, Anna Dreber, Felix Holzmeister, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Gideon Nave, Brian A. Nosek ✉, Thomas Pfeiffer, Adam Altmejd, Nick Buttrick, Taizan Chan, Yiling Chen, Eskil Forsell, Anup Gampa, Emma Heikensten, Lily Hummer, Taisuke Imai, Siri Isaksson, Dylan Manfredi, Julia Rose, Eric-Jan Wagenmakers & Hang Wu

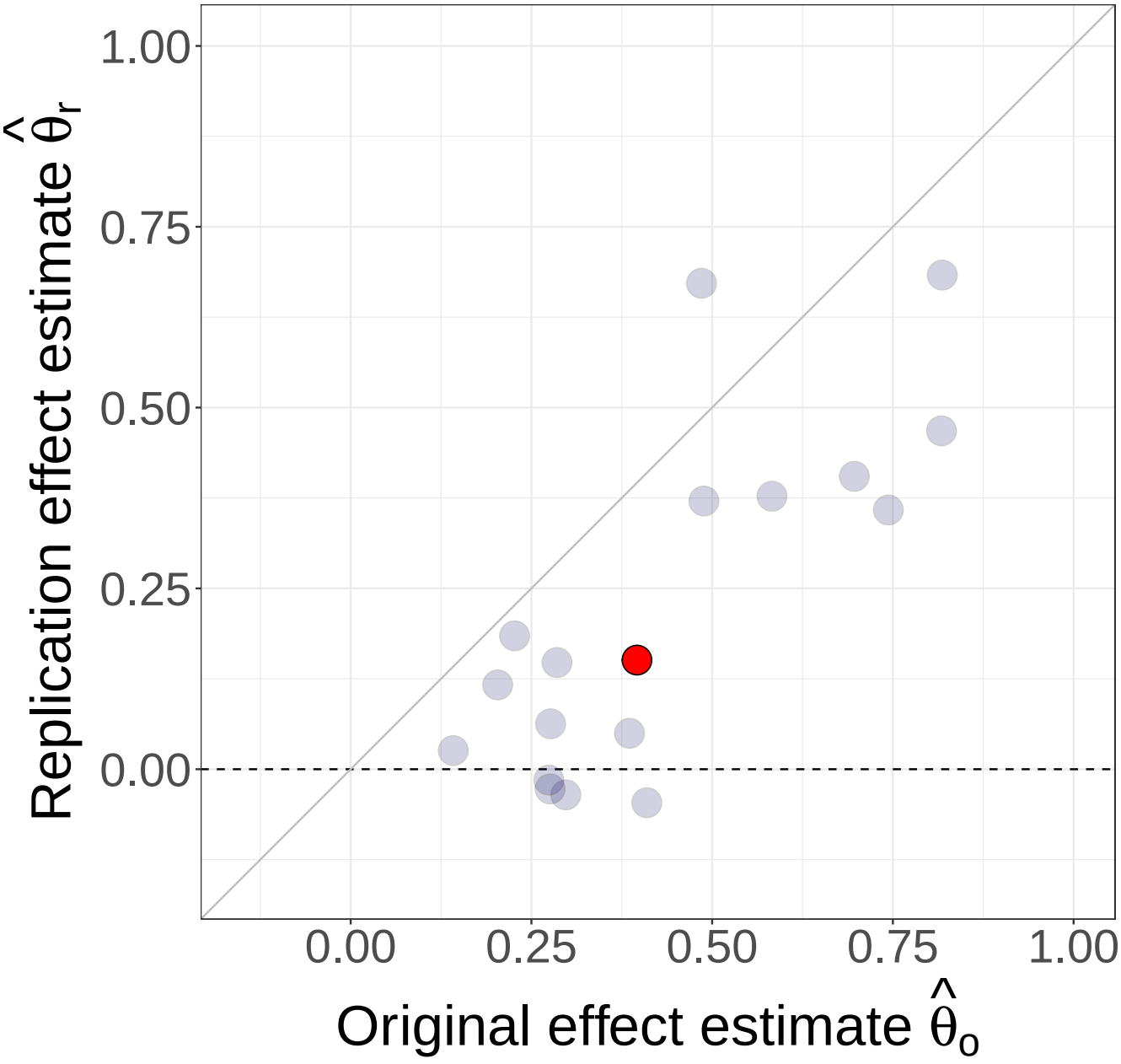
Social sciences replication project

```
library(ReplicationSuccess)
data("RProjects")
social <- subset(RProjects,
                 project == "Social Sciences")
```

Social sciences replication project



Social sciences replication project



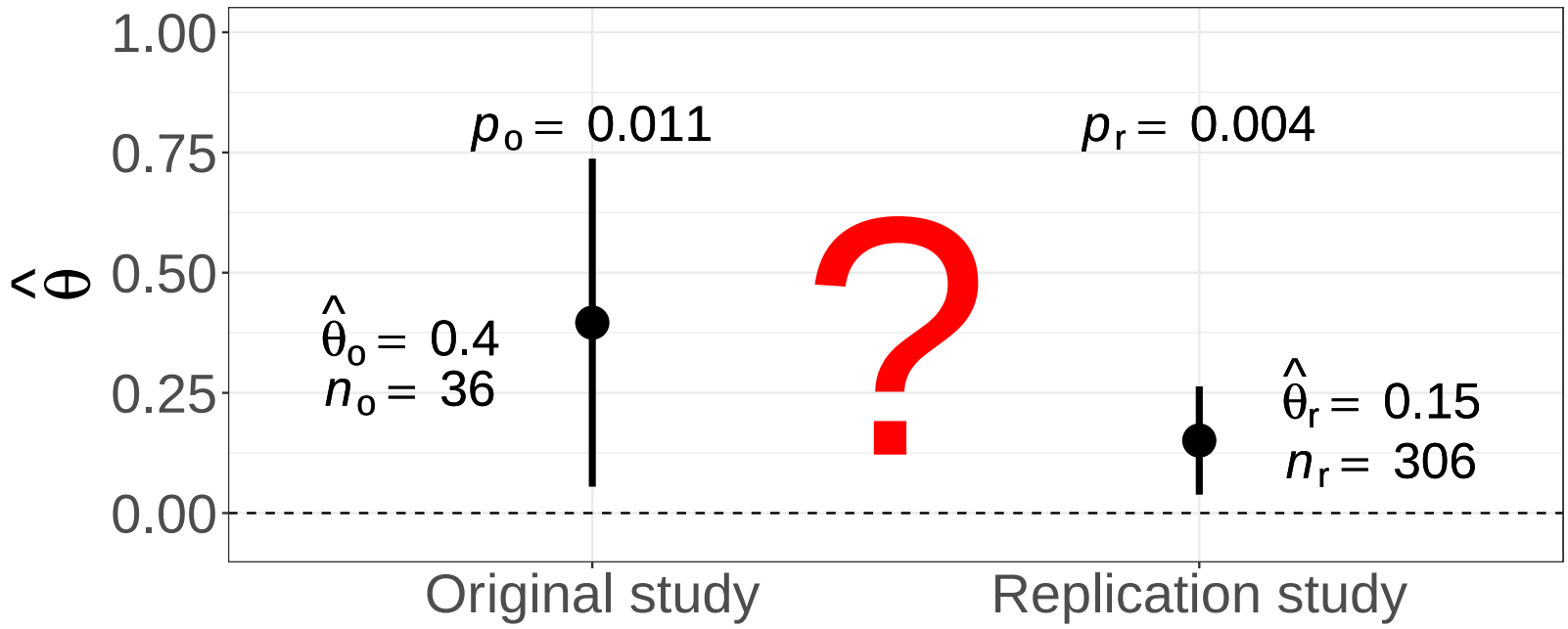
Pyc and Rawson (2010). Science

Original discovery

"Testing improves memory"

Relative effect size $d = \hat{\theta}_r / \hat{\theta}_o = 0.4$

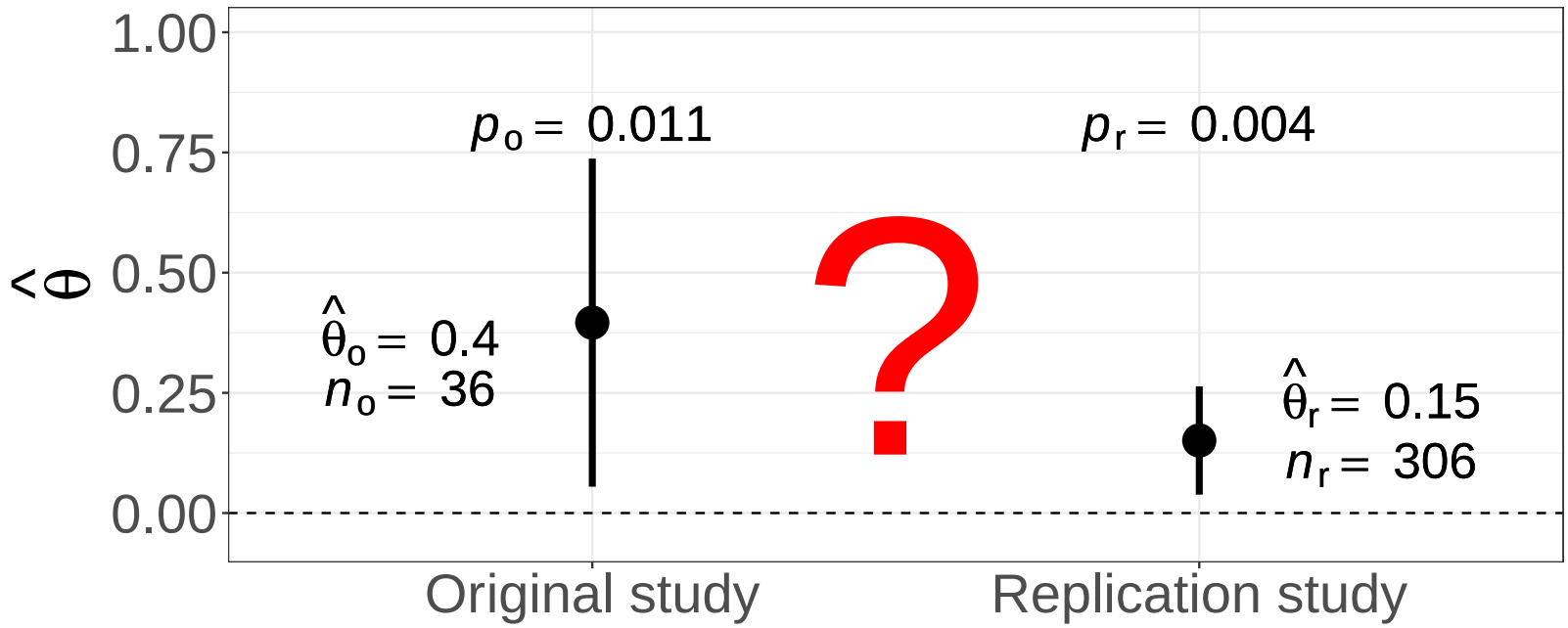
Relative sample size $c = n_r / n_o = 9$



When is a replication successful?

Some proposed criteria

1. Statistical significance
2. Compatibility of effect estimates
3. Meta-analysis of estimates
4. Sceptical *p*-value

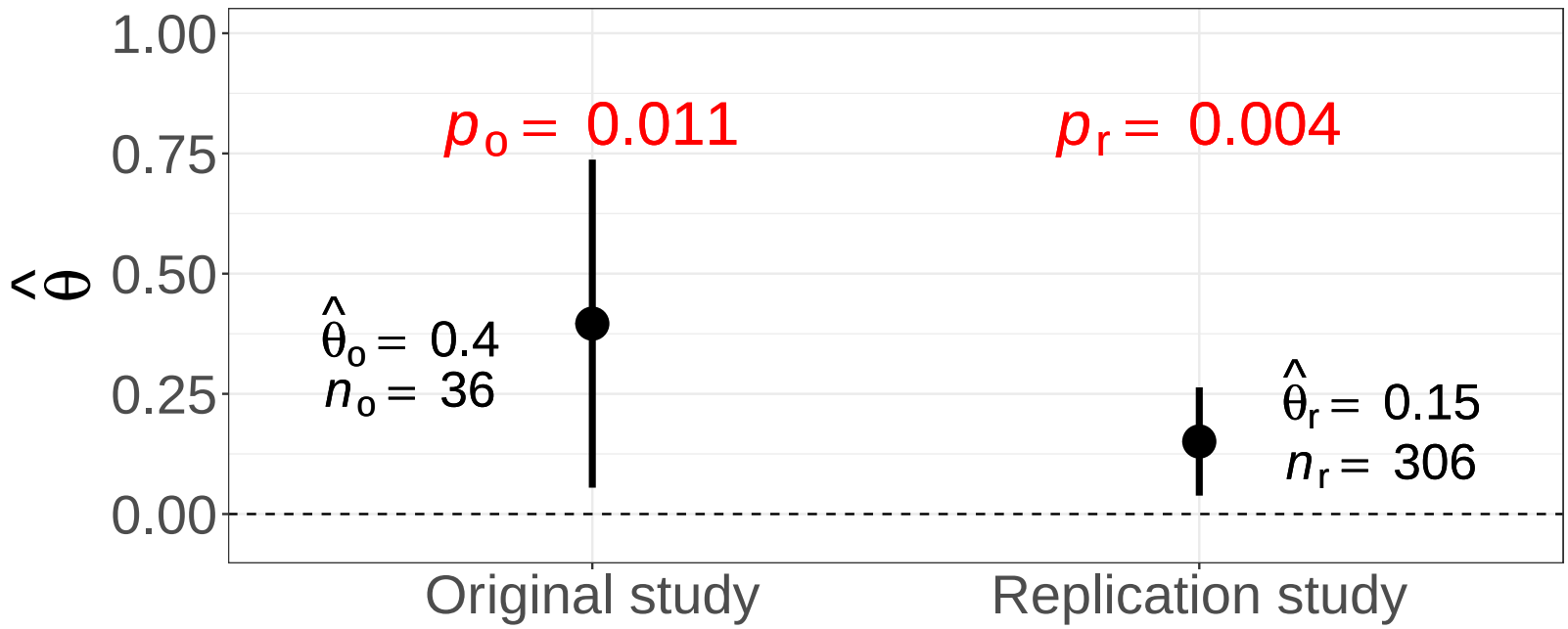


1. Statistical significance

Are both estimates statistically significant in the same direction?

$\rightarrow$ Which threshold?

$\rightarrow$ one-sided $\alpha = 0.025$

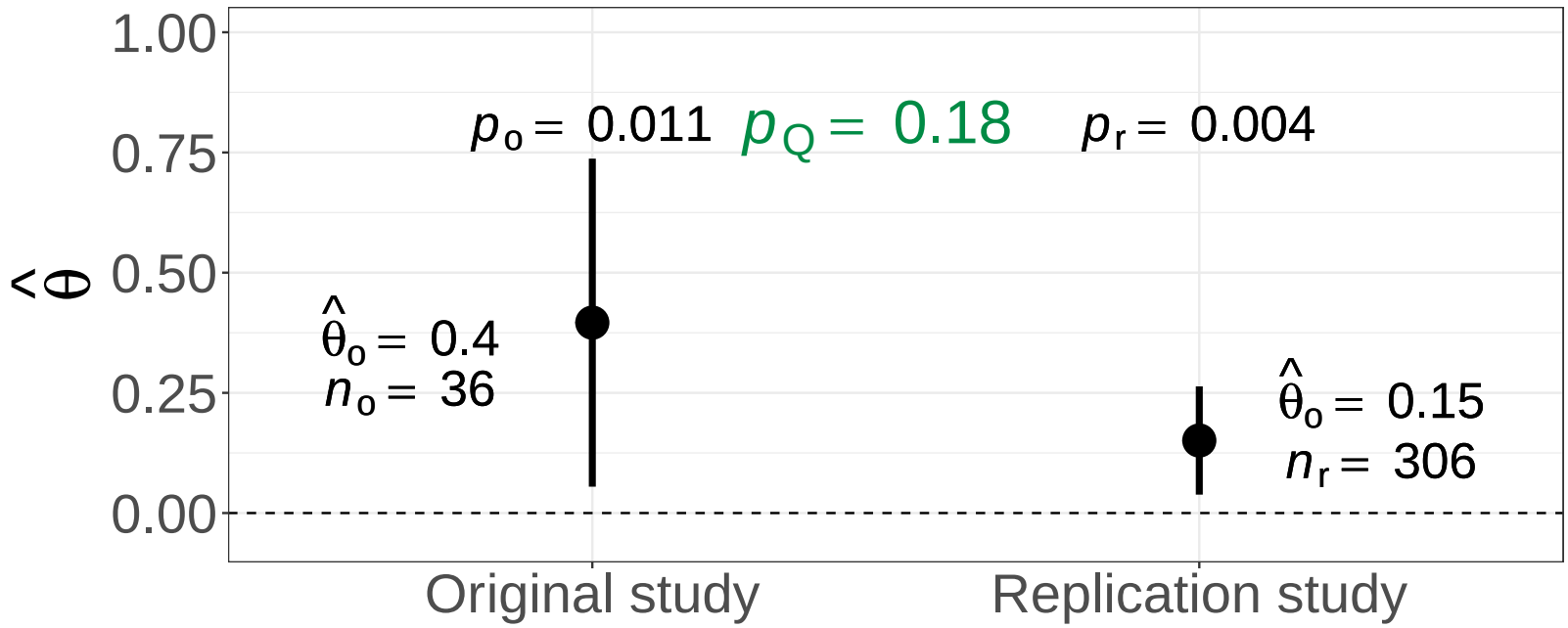


2. Compatibility of effect estimates

Is the meta-analytic Q-test of the estimates statistically significant?

→ Which threshold?

→ two-sided $\alpha = 0.05$

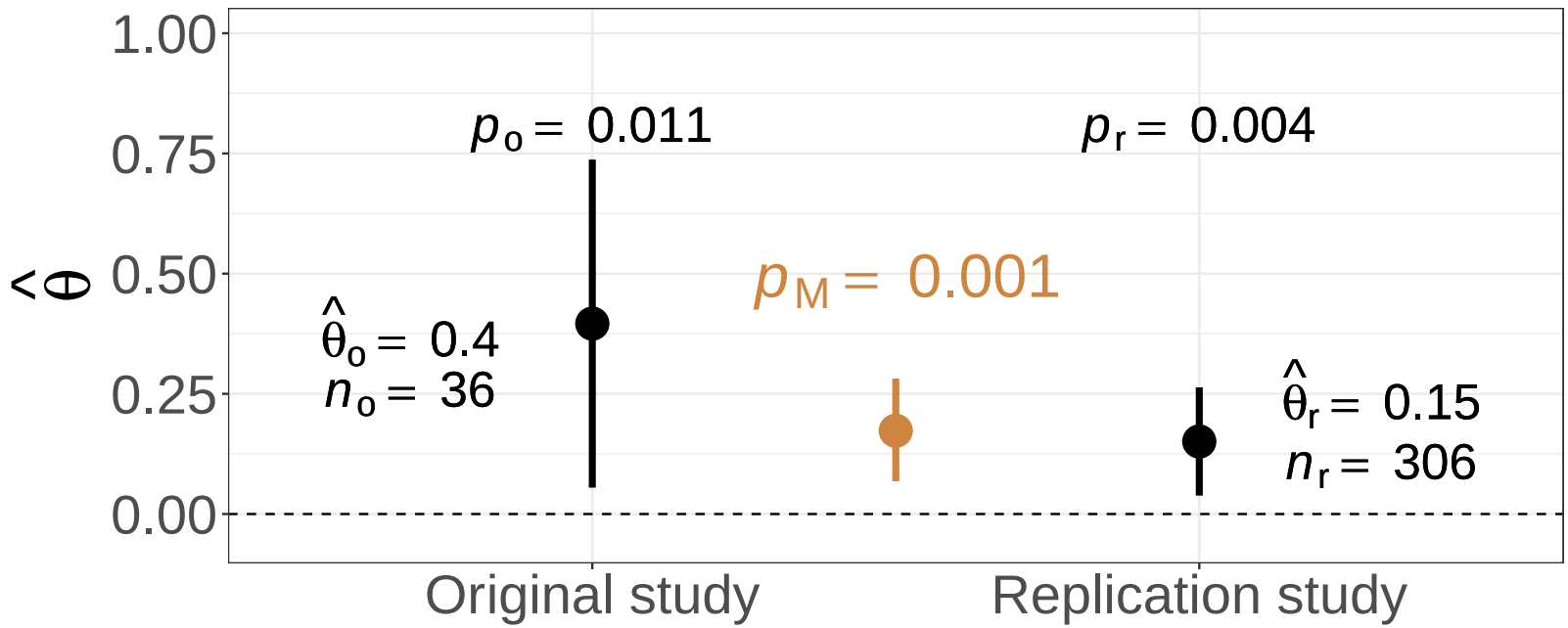


3. Meta-analysis of effect estimates

Is a meta-analytic estimate statistically significant?

$\rightarrow$ Which threshold?

$\rightarrow$ one-sided $\alpha = 0.025^2 = 0.000625$



4. Sceptical *p*-value

New definition of replication success

J. R. Statist. Soc. A (2020)
183, *Part* 2, *pp.* 431–448

A new standard for the analysis and design of replication studies

Leonhard Held

University of Zurich, Switzerland

https://doi.org/10.1111/rssa.12493

THE ASSESSMENT OF REPLICATION SUCCESS BASED ON RELATIVE EFFECT SIZE

By Leonhard Held, Charlotte Micheloud and Samuel Pawel

Epidemiology, Biostatistics and Prevention Institute, Center for Reproducible Science, University of Zurich,
leonhard.held@uzh.ch; charlotte.micheloud@uzh.ch; samuel.pawel@uzh.ch

Replication studies are increasingly conducted in order to confirm original findings. However, there is no established standard how to assess replication success and in practice many different approaches are used. The purpose of this paper is to refine and extend a recently proposed reverse-Bayes approach for the analysis of replication studies. We show how this method is directly related to the relative effect size, the ratio of the replication to the original effect estimate. This perspective leads to a new proposal to recalibrate the assessment of replication success, the golden level. The recalibration ensures that for borderline significant original studies replication success can only be achieved if the replication effect estimate is larger than the original one. Conditional power for replication success can then take any desired value if the original study is significant and the replication sample size is large enough. Compared to the standard approach to require statistical significance of both the original and replication study, replication success at the golden level offers uniform gains in project power and controls the Type-I error rate if the replication sample size is not smaller than the original one. An application to data from four large replication projects shows that the new approach leads to more appropriate inferences, as it penalizes shrinkage of the replication estimate compared to the original one, while ensuring that both effect estimates are sufficiently convincing on their own.

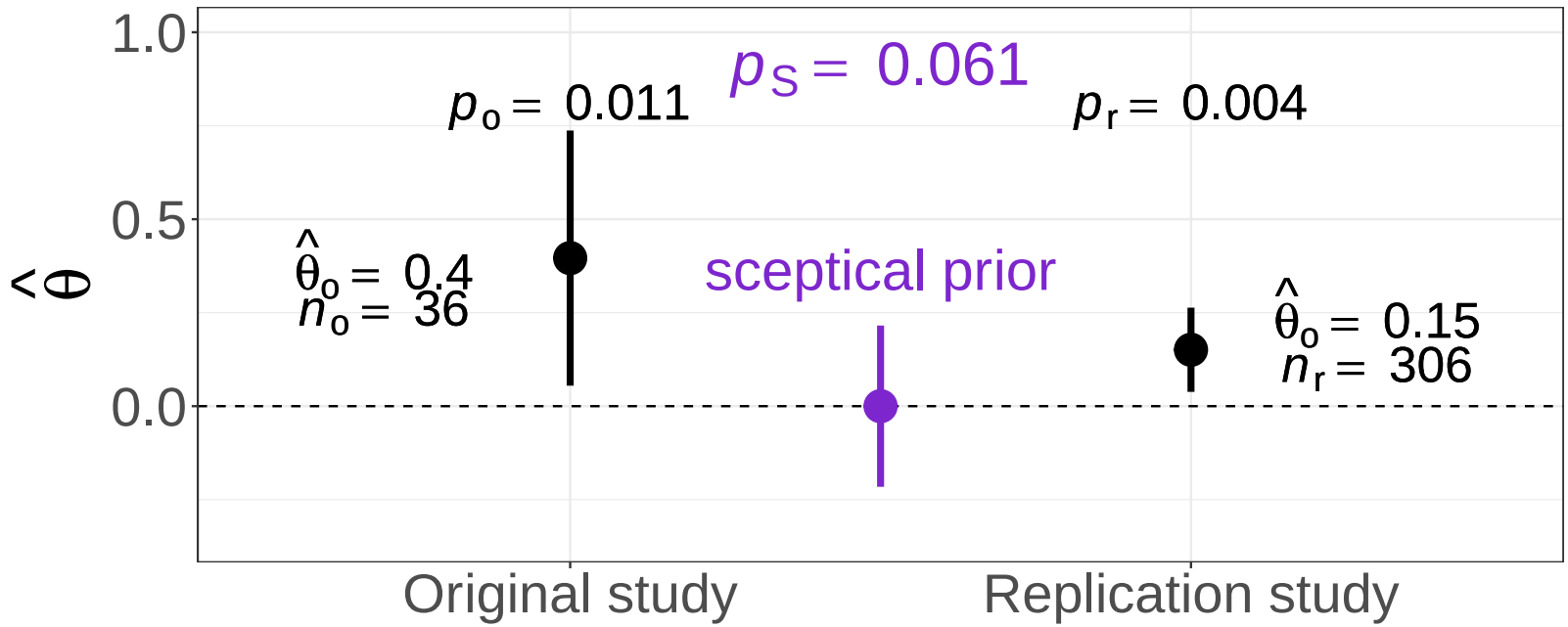
https://arxiv.org/abs/2009.07782

4. Sceptical *p*-value

New definition of replication success

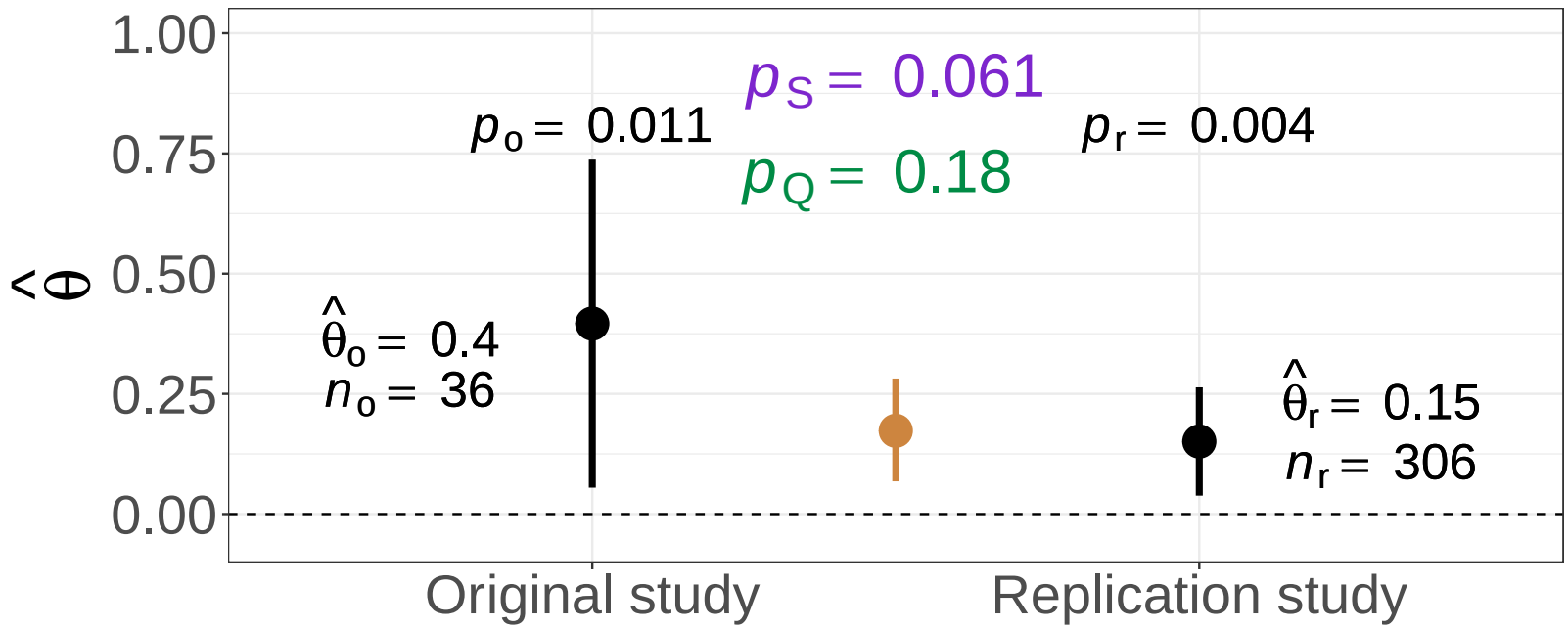
Can we convince a sceptic whose priof beliefs make the original study not significant?

If $p_S \leq \alpha$ we have replication success at level α



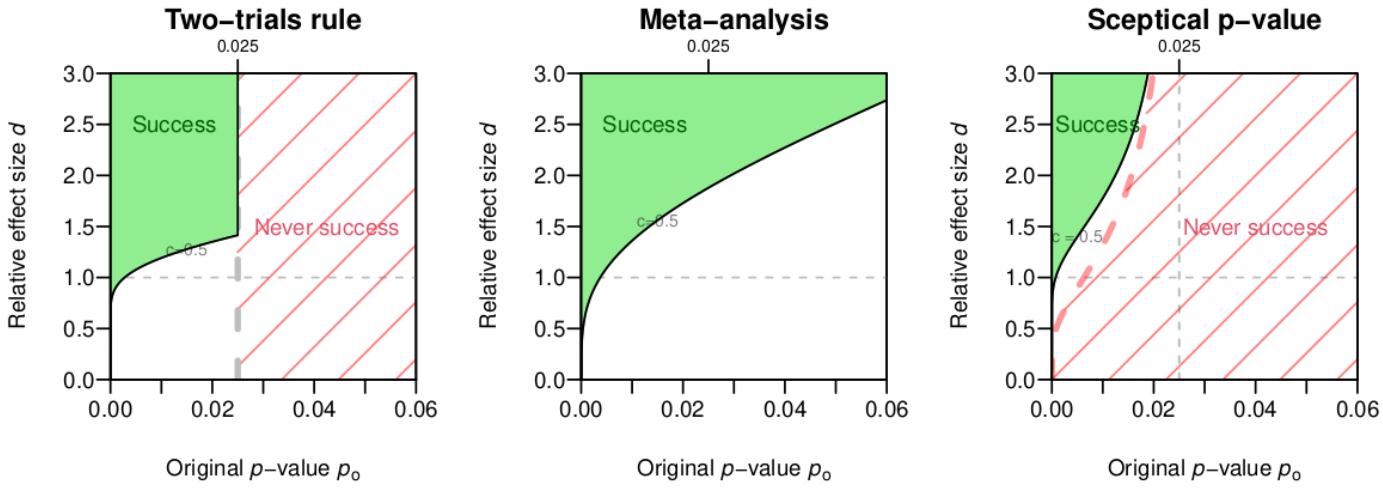
Assessement of replication success

- Significance doesn't take into account effect size
- *Q*-test doesn't take into account significance
- Meta-analysis assumes exchangeability
- Sceptical *p*-value takes into account effect size and significance



Comparison of Success Regions

Relative sample size c = 0.5



CM: Here I would suggest to put the same plots as in the METRICS talk to show the influence of the sample size on the success region. Maybe only for the golden level instead of nominal level. (I unfortunately can't do it as I don't have plotLevel.R)

Exercise Session 1

Analysis of replication studies

February 2022

Package ReplicationSuccess

- Installation

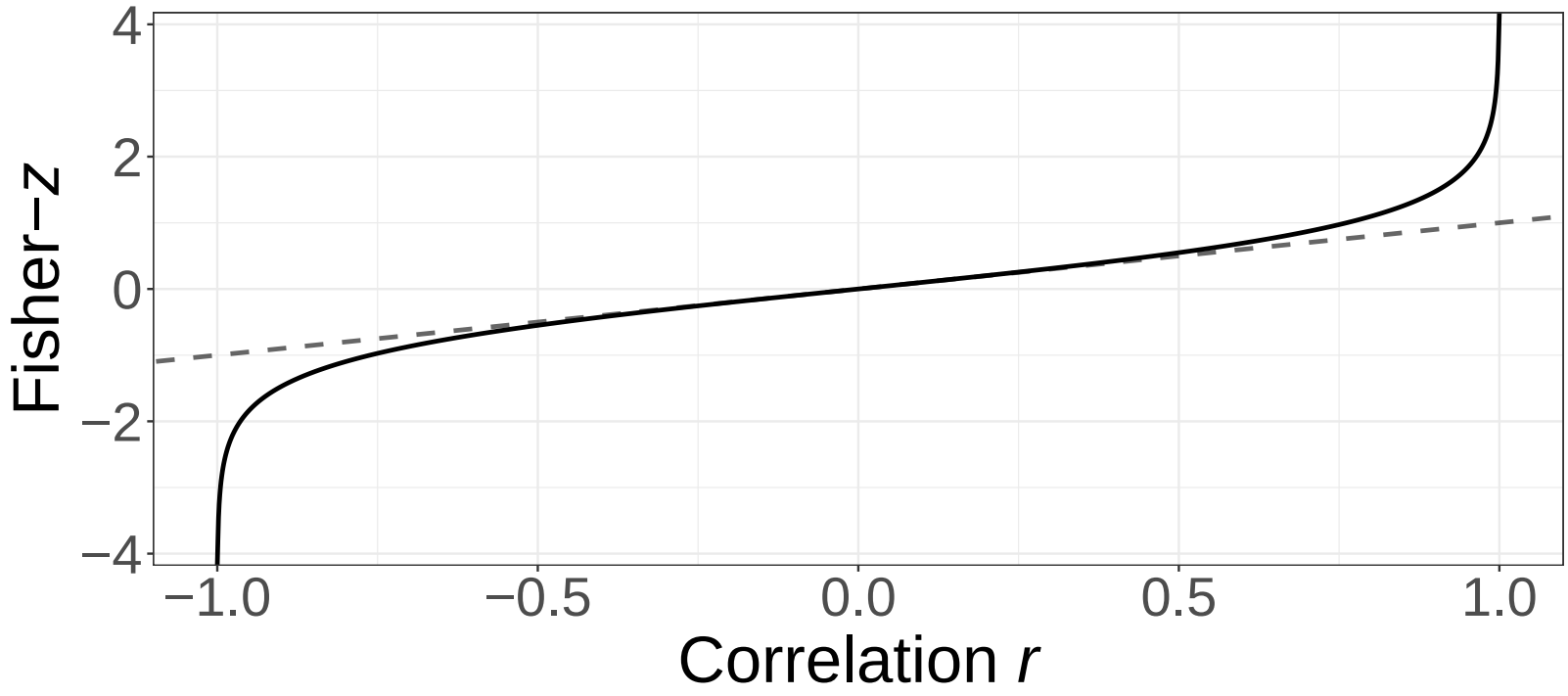
```
# Linux / Windows
install.packages(pkgs = "ReplicationSuccess",
                 repos ="http://R-Forge.R-project.org")
# Mac
install.packages(pkgs = "ReplicationSuccess",
                 repos = "http://R-Forge.R-project.org",
                 type = "source")
```

- Usage

```
library(ReplicationSuccess)
vignette("ReplicationSuccess")
?pSceptical # documentation
news(package = "ReplicationSuccess") # news page
```

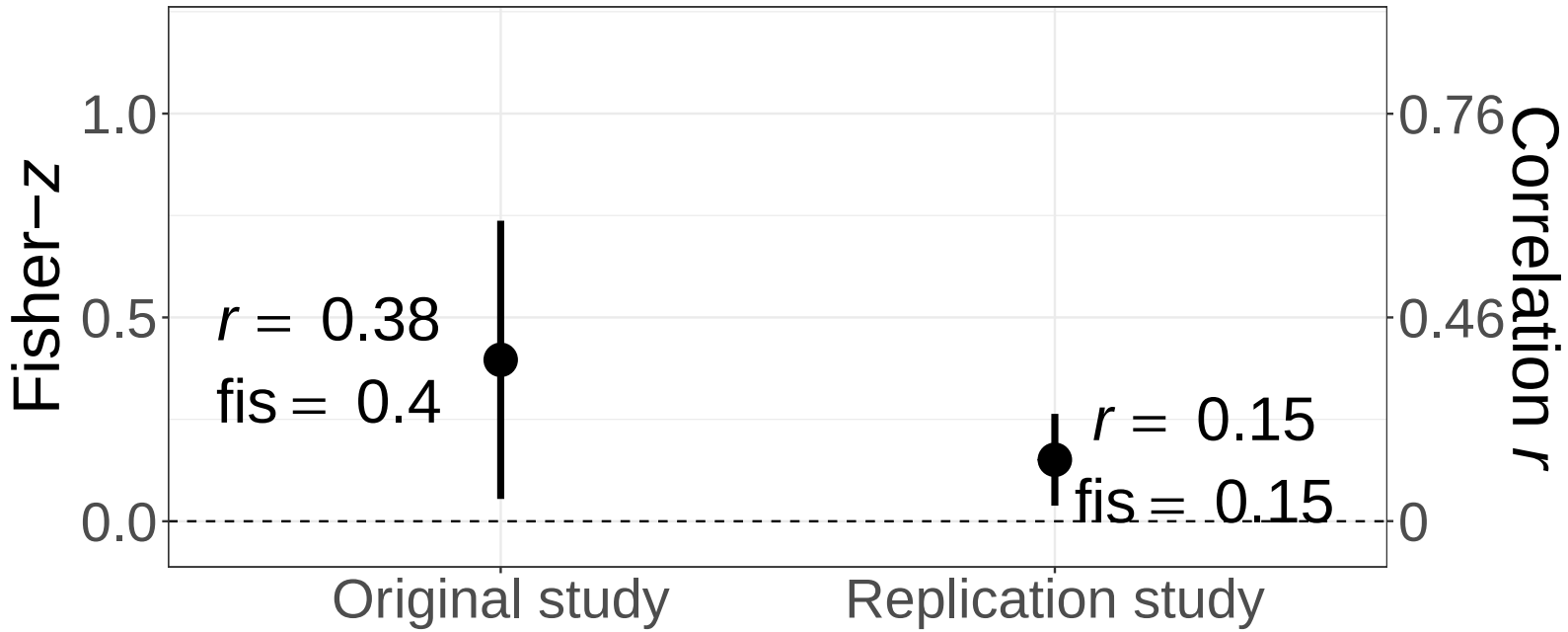
Statistical framework of package

– Effect estimates are assumed to be normally distributed after suitable transformation
→ Fisher's *z*-transformation for correlation coefficients *r* with (effective) sample size $n - 3$



Statistical framework of package

– Effect estimates are assumed to be normally distributed after suitable transformation
→ Fisher's z-transformation for correlation coefficients r with (effective) sample size $n - 3$



Data sets

```
?RProjects # Documentation
```

Most important variables

`project`	Replication project
`ro`	Original effect on correlation scale
`rr`	Replication effect on correlation scale
`fiso`	Original effect on Fisher-*z* scale
`fisr`	Replication effect on Fisher-*z* scale
`se_fiso`	Standard error of `fiso`
`se_fisr`	Standard error of `fisr`

Statistical framework of package

Key quantities

– z-value z_o or (one-sided) p-value p_o of original study

```
RProjects$zo <- RProjects$fiso/RProjects$se_fiso
RProjects$po1 <- z2p(RProjects$zo,
                     alternative = "one.sided")
```

Statistical framework of package

Key quantities

- z-value z_o or (one-sided) p-value p_o of original study

```
RProjects$zo <- RProjects$fiso/RProjects$se_fiso
RProjects$po1 <- z2p(RProjects$zo,
                     alternative = "one.sided")
```

- z-value z_r or (one-sided) p-value p_r of replication study

```
RProjects$zr <- RProjects$fisr/RProjects$se_fisr
RProjects$pr1 <- z2p(RProjects$zr,
                     alternative = "one.sided")
```

Statistical framework of package

Key quantities

- z-value z_o or (one-sided) p-value p_o of original study

```
RProjects$zo <- RProjects$fiso/RProjects$se_fiso
RProjects$po1 <- z2p(RProjects$zo,
                     alternative = "one.sided")
```

- z-value z_r or (one-sided) p-value p_r of replication study

```
RProjects$zr <- RProjects$fisr/RProjects$se_fisr
RProjects$pr1 <- z2p(RProjects$zr,
                     alternative = "one.sided")
```

- relative sample size (or variance ratio)
  $c = \sigma_o^2/\sigma_r^2 = n_r/n_o$

```
RProjects$c <- RProjects$se_fiso^2/RProjects$se_fisr^2
```

Exercises

(Solutions available at https://osf.io/fcrj6/)

Load the package and the data sets with

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

Compute the key quantities z_o, z_r, c, and the one-sided p-values p_o and p_r with

```
RProjects$zo <- RProjects$fiso/RProjects$se_fiso
RProjects$zr <- RProjects$fisr/RProjects$se_fisr
RProjects$c <- RProjects$se_fiso^2/RProjects$se_fisr^2
RProjects$po1 <- z2p(RProjects$zo,
                     alternative = "one.sided")
RProjects$pr1 <- z2p(RProjects$zr,
                     alternative = "one.sided")
```

Exercises

(Solutions available at https://osf.io/fcrj6/)

For all studies from the replication projects investigate

Exercise 1.1

How many study pairs fulfill the **significance** criterion for replication success? Use a threshold of $\alpha = 0.025$ for the one-sided *p*-values.

Exercise 1.2

For how many study pairs do you find evidence for **incompatible** effect estimates (on Fisher *z*-scale)? Use the function `Qtest()` and a threshold of $\alpha = 0.05$ for the resulting *p*-value.

Exercises

(Solutions available at https://osf.io/fcrj6/)

For all studies from the replication projects investigate

Exercise 1.3

Compute the one-sided **sceptical *p*-value**. How many replication studies are successful at 0.025? Use the function `pSceptical()`

Exercise 1.4

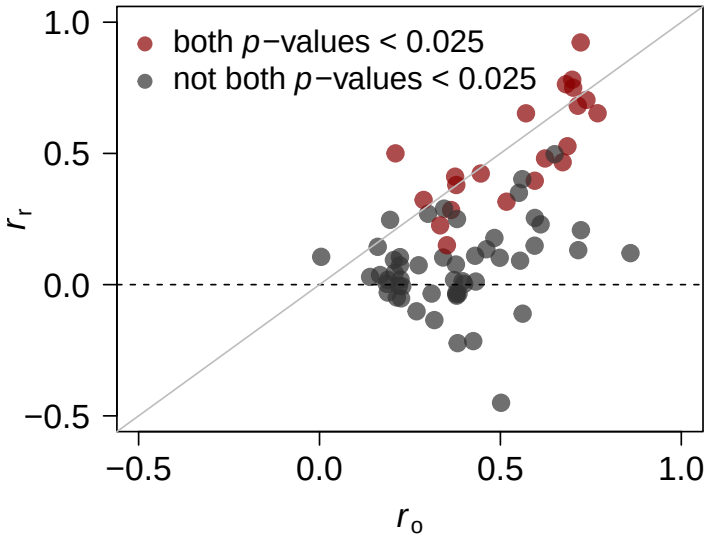
Look closer at the studies which show **discrepancies** in terms of replication success based on significance and the sceptical *p*-value. How do their effect estimates and sample sizes compare?

Solution: Exercise 1.1

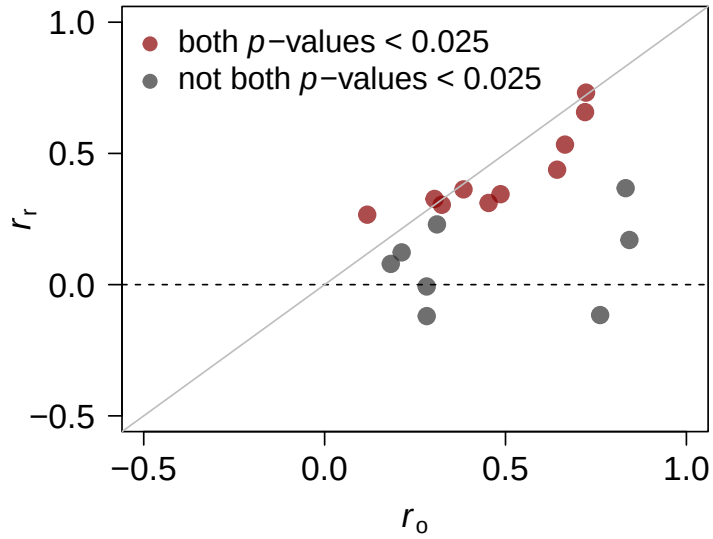
Project	Both p-values < 0.025
Psychology	29% (21/73)
Experimental Economics	56% (10/18)
Social Sciences	62% (13/21)
Experimental Philosophy	74% (23/31)
all	47% (67/143)

Solution: Exercise 1.1

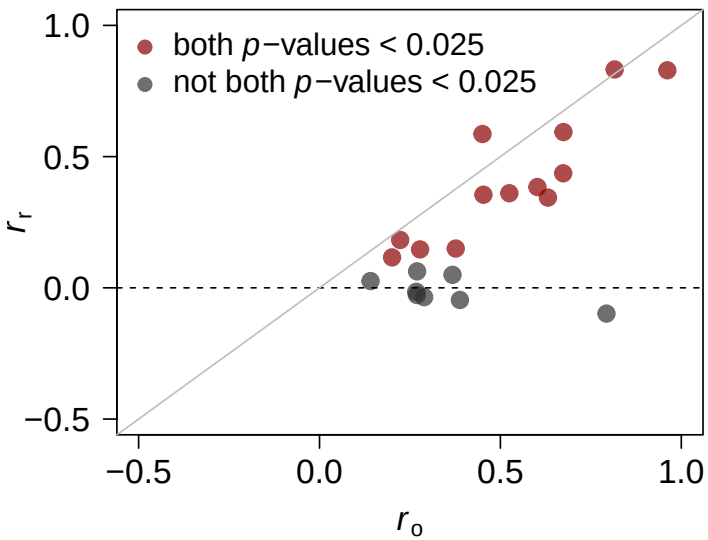
Psychology: 29% (21/73)



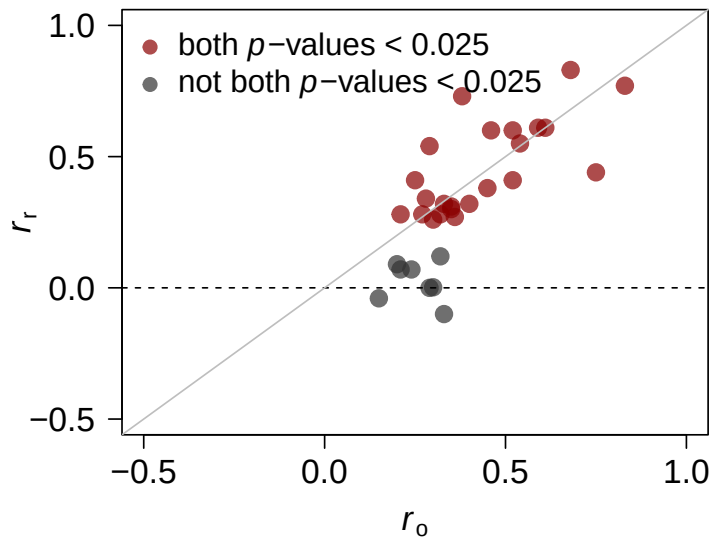
Experimental Economics: 56% (10/18)



Social Sciences: 62% (13/21)



Experimental Philosophy: 74% (23/31)

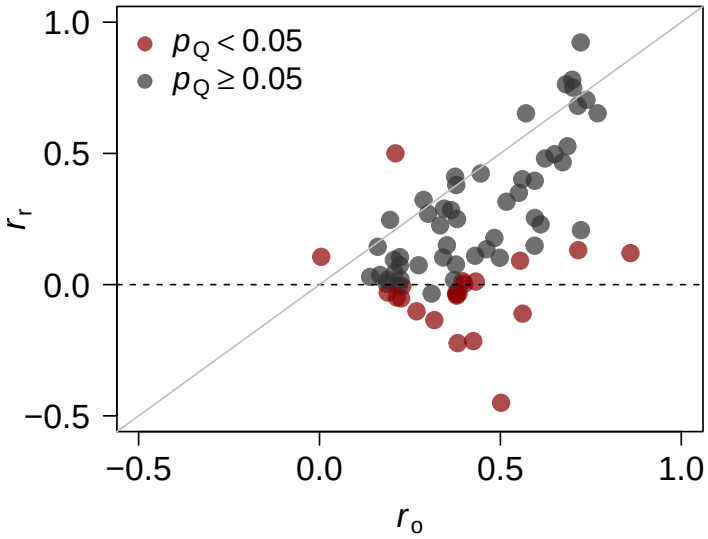


Solution: Exercise 1.2

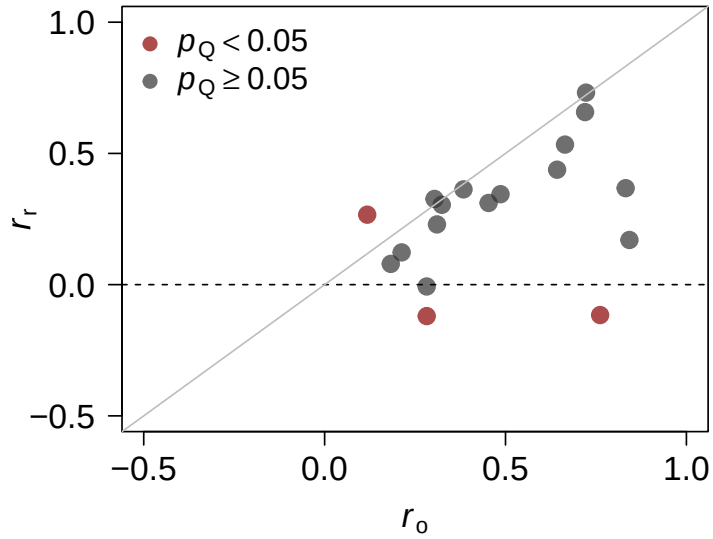
Project	Incompatible estimates ($p_Q < 0.05$)
Psychology	30% (22/73)
Experimental Economics	17% (3/18)
Social Sciences	33% (7/21)
Experimental Philosophy	16% (5/31)
all	26% (37/143)

Solution: Exercise 1.2

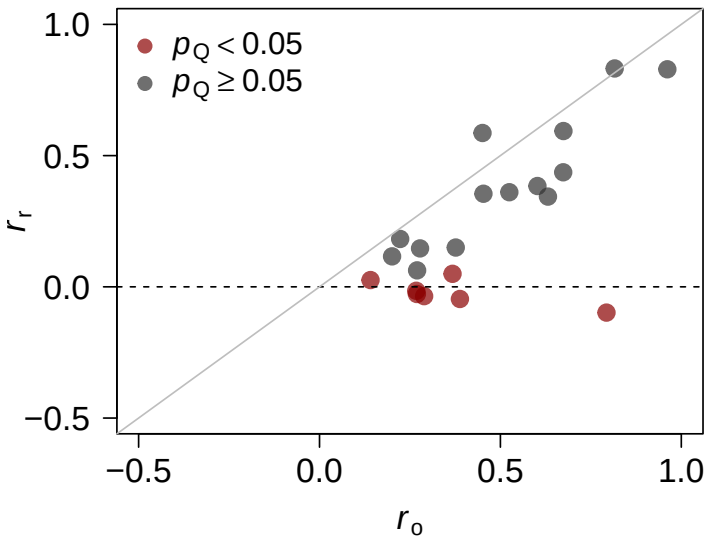
Psychology: 30% incompatible



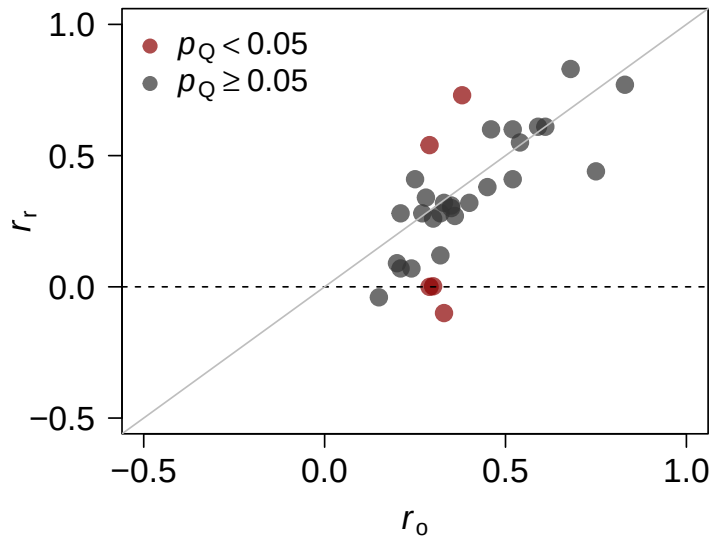
Experimental Economics: 17% incompatible



Social Sciences: 33% incompatible



Experimental Philosophy: 16% incompatible

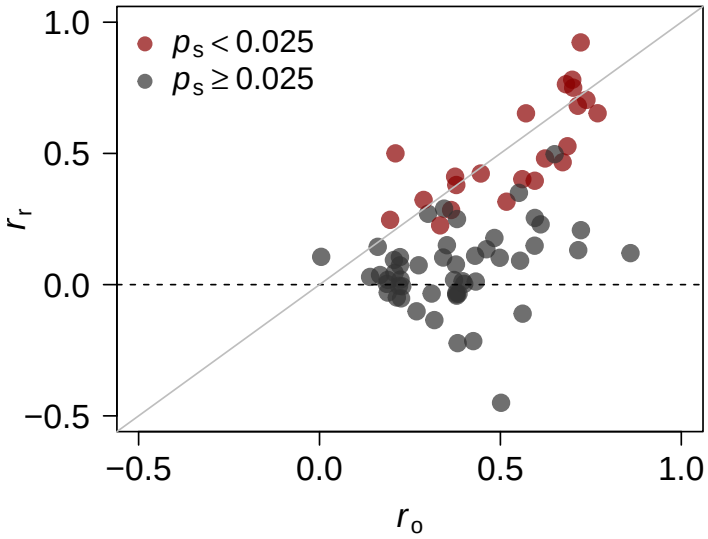


Solution: Exercise 1.3

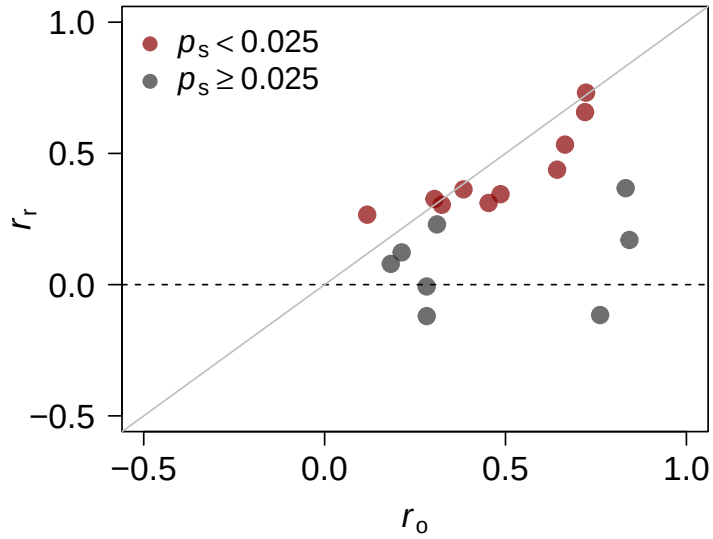
Project	sceptical p-value < 0.025
Psychology	30% (22/73)
Experimental Economics	56% (10/18)
Social Sciences	52% (11/21)
Experimental Philosophy	71% (22/31)
all	45% (65/143)

Solution: Exercise 1.3

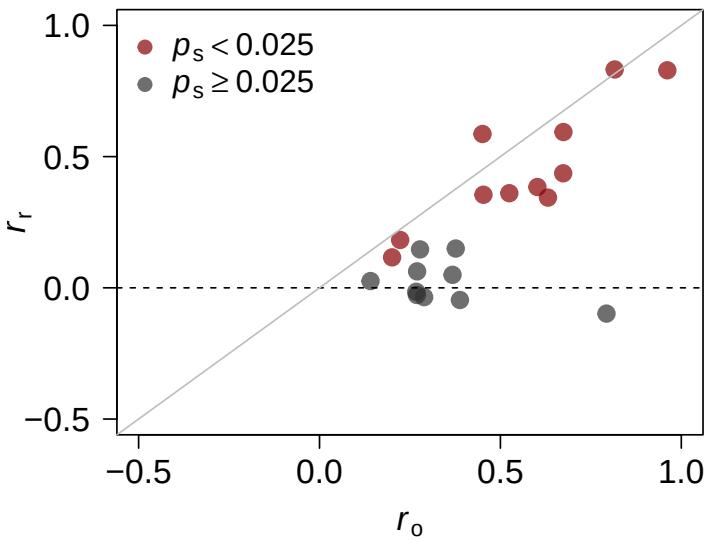
Psychology: 30% (22/73)



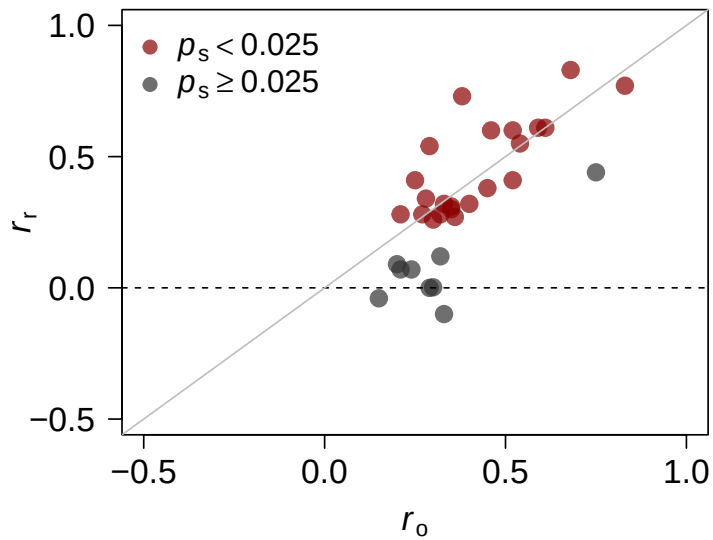
Experimental Economics: 56% (10/18)



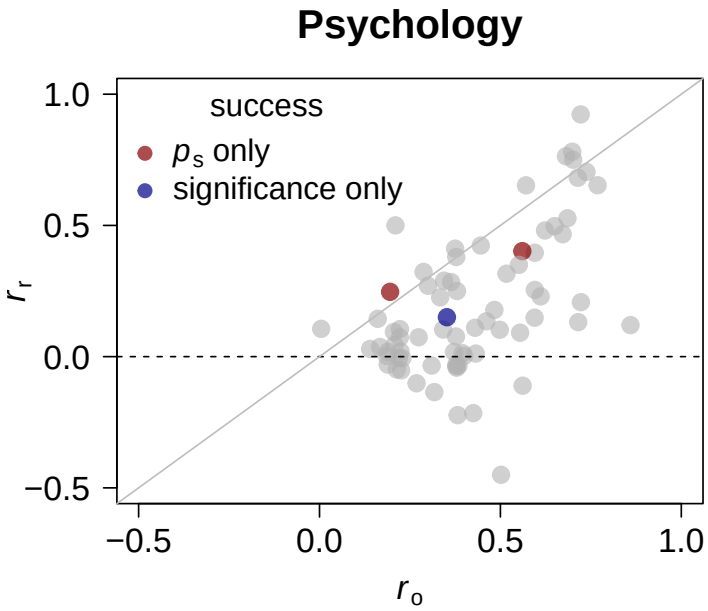
Social Sciences: 52% (11/21)

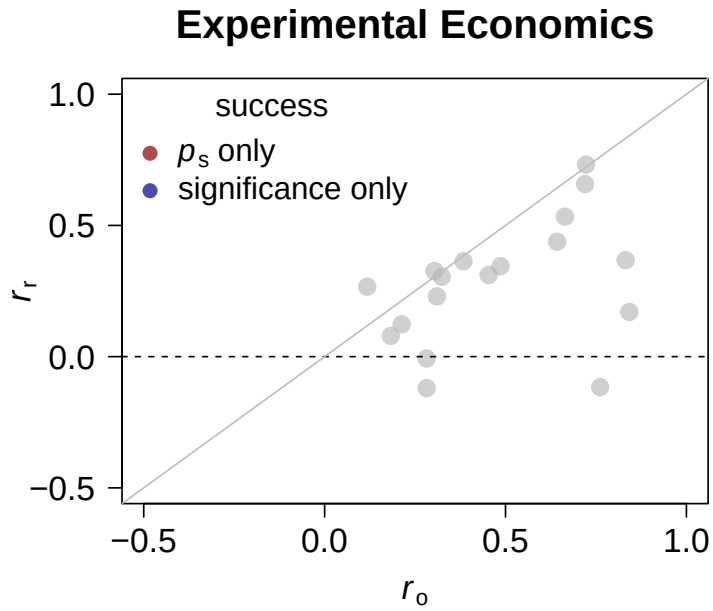


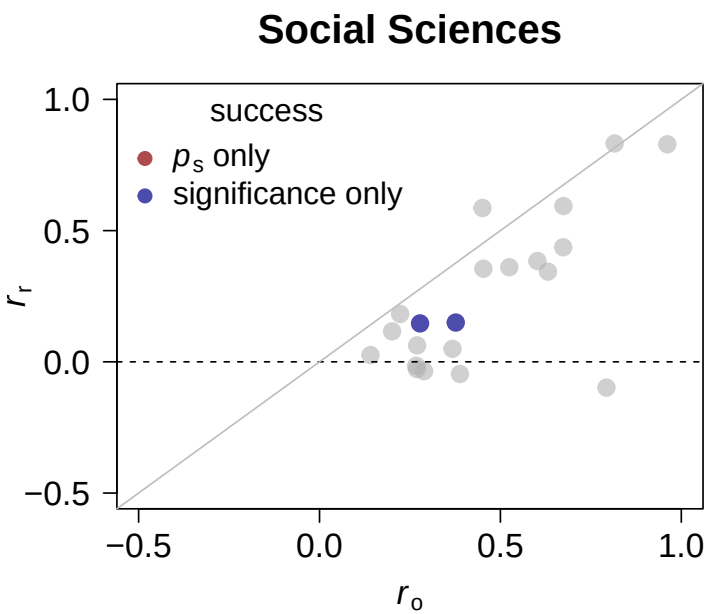
Experimental Philosophy: 71% (22/31)

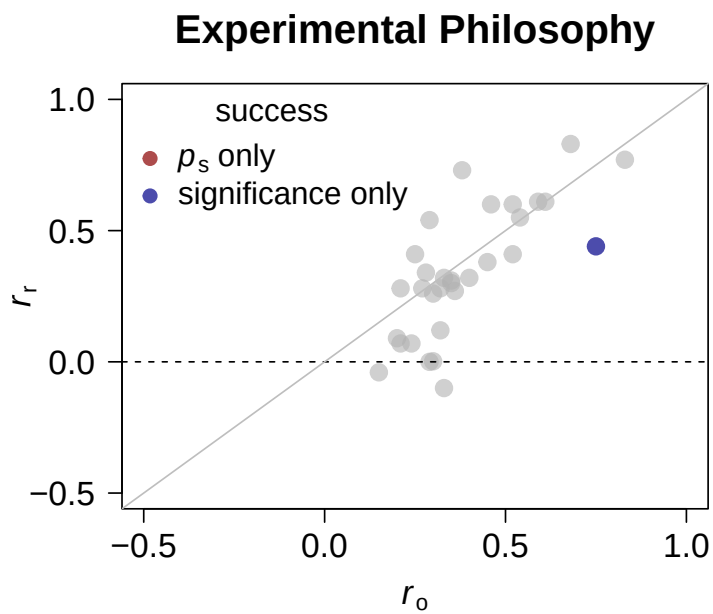


Solution: Exercise 1.4









Solution: Exercise 1.4

Study	n_r/n_o	r_o	r_r	p_o	p_r	p_s
Schmidt and Besner (2008)	2.6	0.2	0.25	0.028	< 0.0001	0.024
Oberauer (2008)	0.6	0.56	0.4	0.0003	0.035	0.017
Payne, Burkley, and Stokes (2008)	2.7	0.35	0.15	0.001	0.023	0.031
Balafoutas and Sutter (2012)	3.5	0.28	0.15	0.009	0.011	0.04
Pyc and Rawson (2010)	9.2	0.38	0.15	0.011	0.004	0.061
Nichols (2006)	9.4	0.75	0.44	0.015	0.0006	0.049

Exercise Session 2

Design based on significance

Design of replication studies

Sample size of replication study

- Direct replication $\rightarrow$ procedures of replication study as closely matched as possible to original study
- But same sample size as in original study can lead to a very low power (Goodman, 1992)

  $\rightarrow$ proper sample size calculation is essential

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

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

STEVEN N. GOODMAN

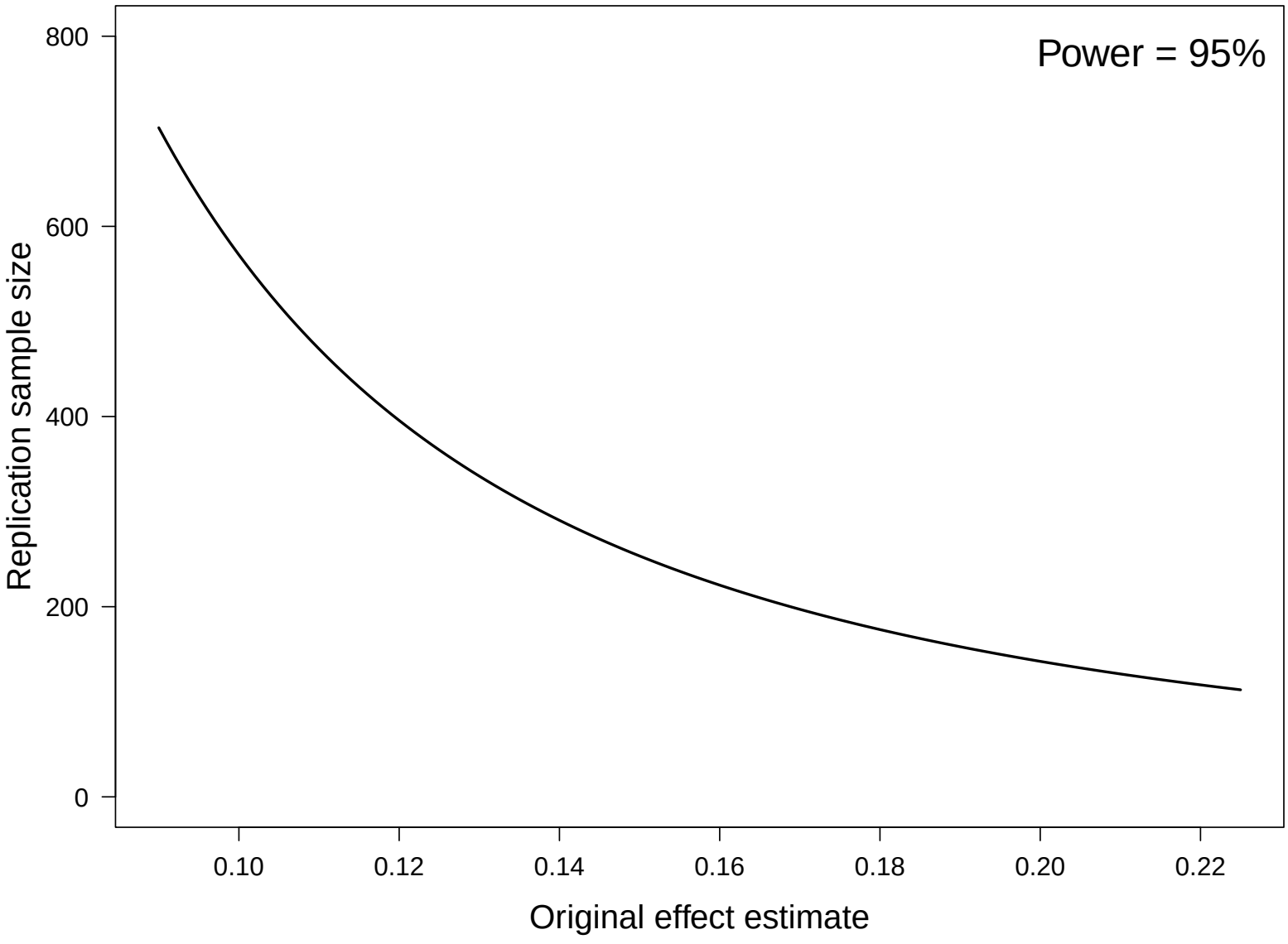
Johns Hopkins University School of Medicine, Department of Oncology, Division of Biostatistics, 550 N. Broadway, Suite 1103, Baltimore MD 21205, U.S.A.

What is used in practice

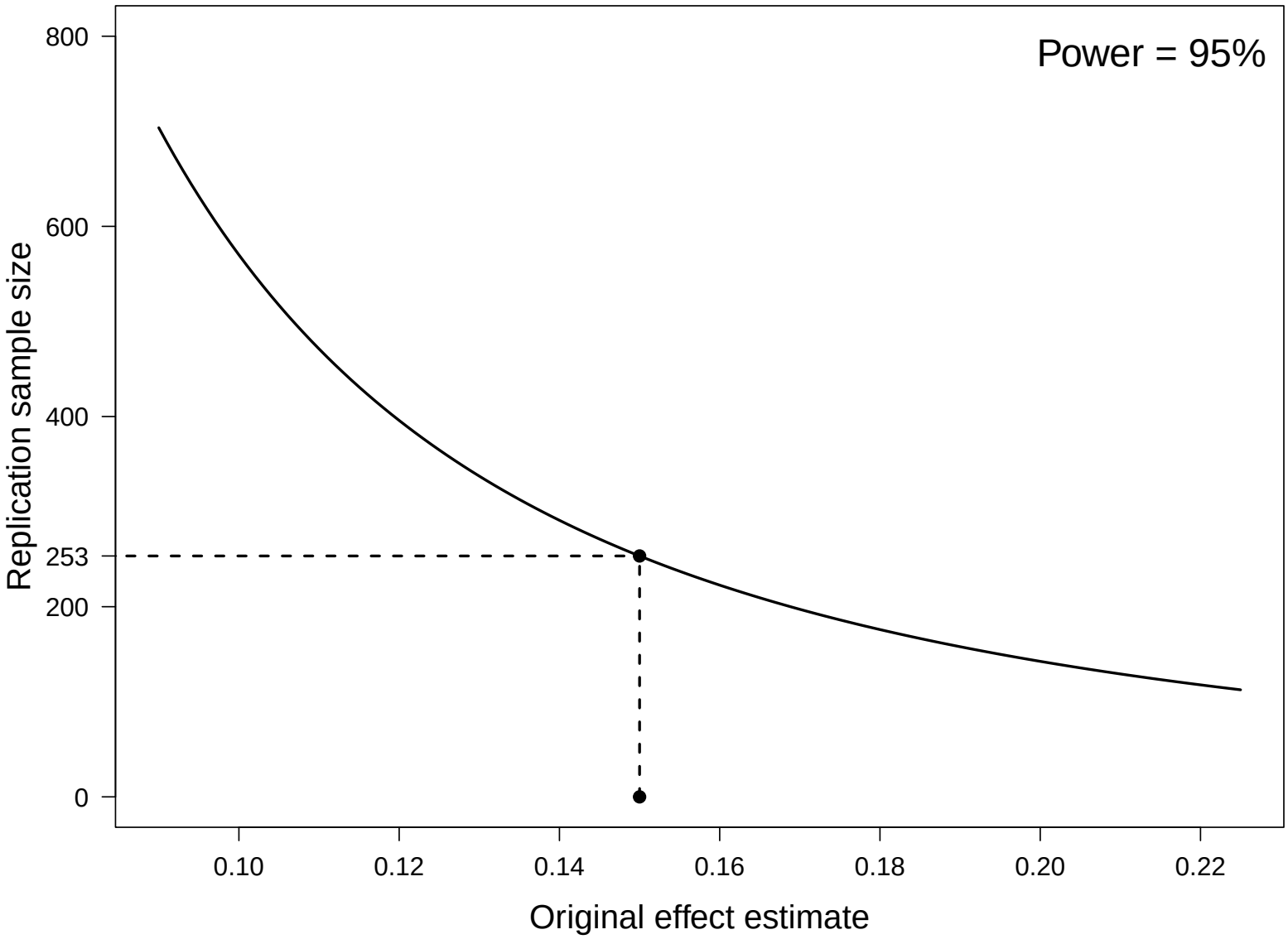
Standard sample size calculation

- Goal is to have between 80% and 95% power in the replication study to detect the effect estimate from the original study.
- Original effect estimate is sometimes shrunken by a factor of 50%.
- Uncertainty of original effect estimate is ignored

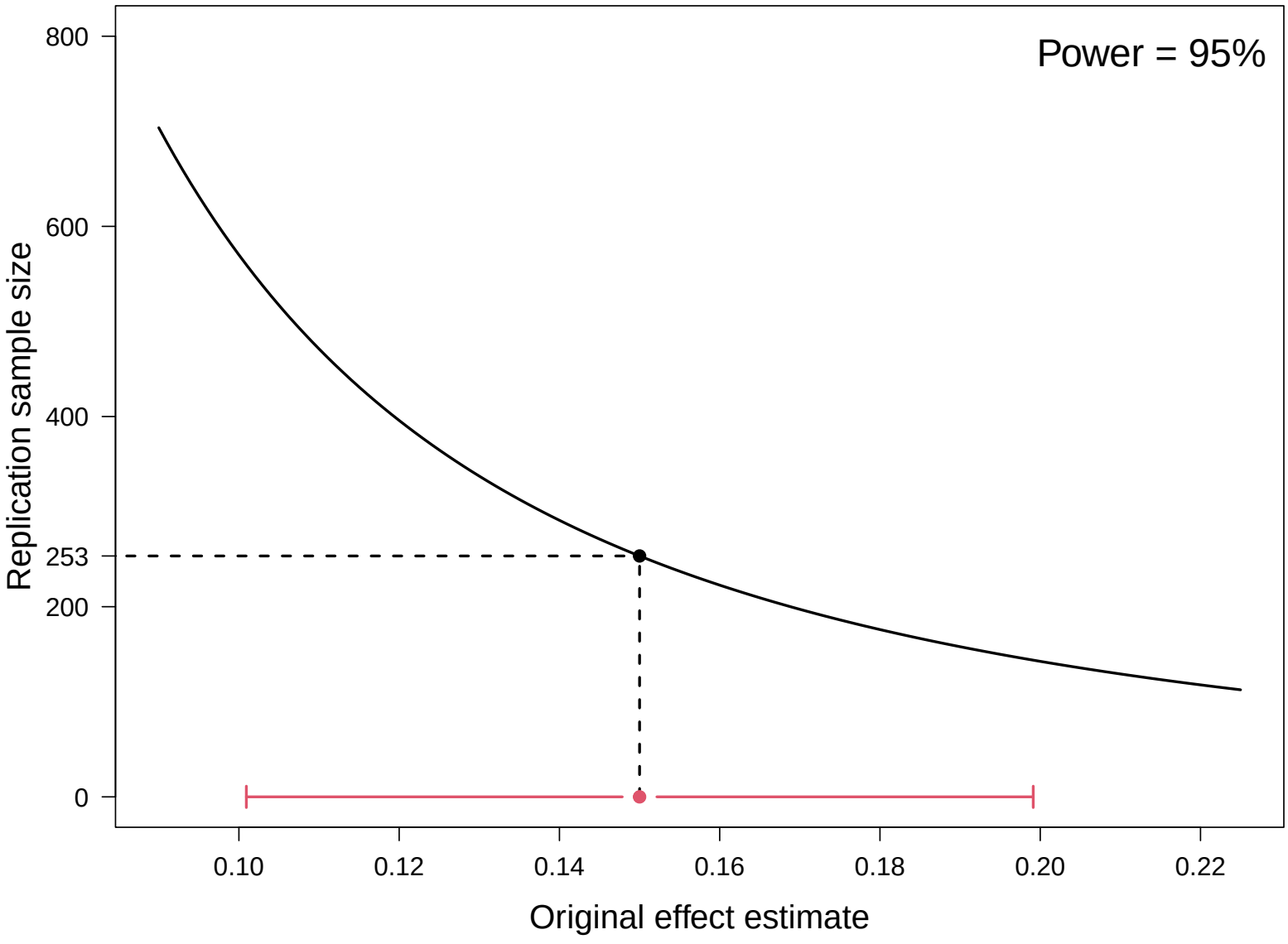
Standard sample size calculation



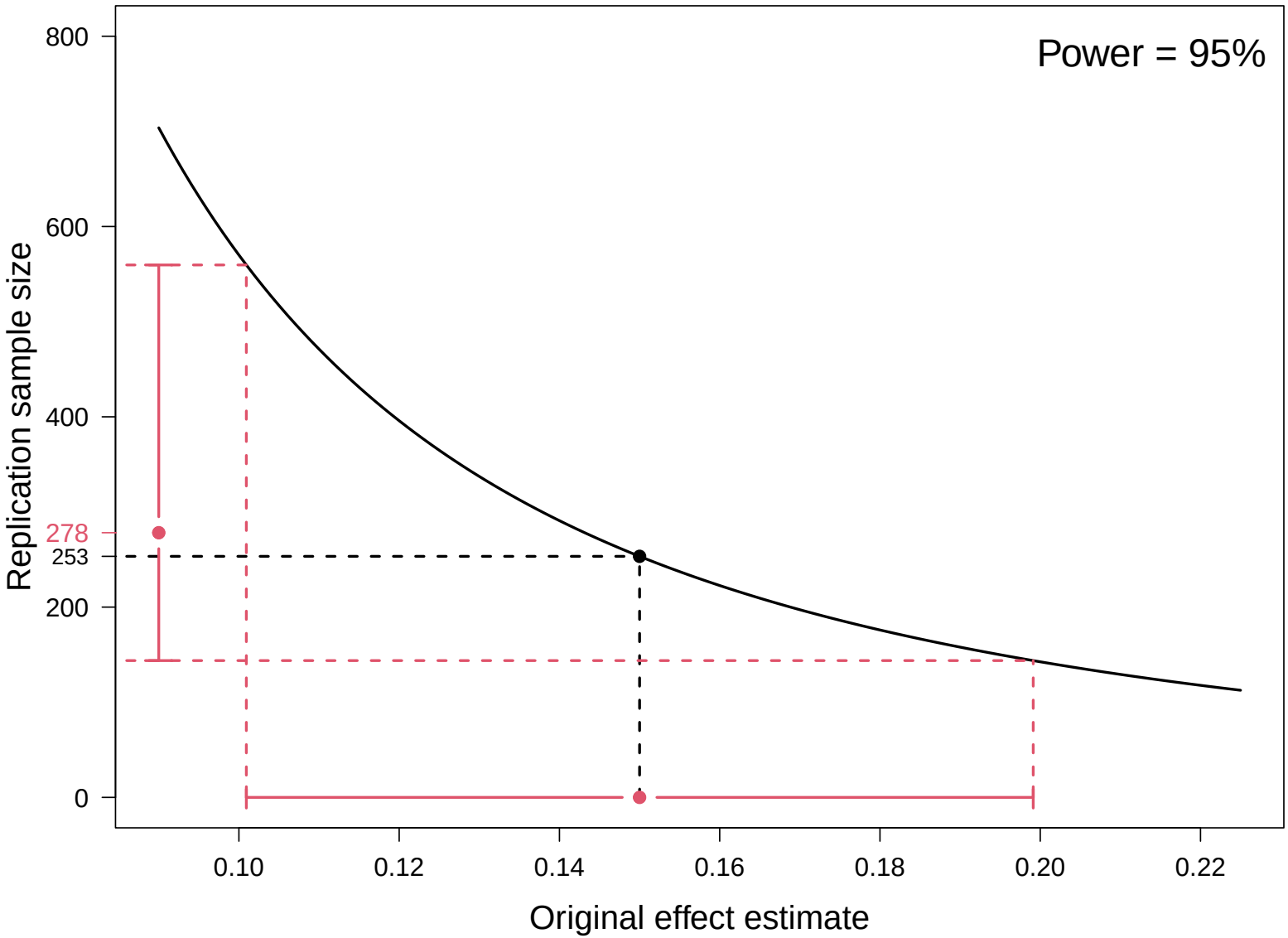
Standard sample size calculation



Incorporation of uncertainty



Incorporation of uncertainty



Incorporation of uncertainty

Design prior

- *Conditional*: ignores uncertainty of original study
- *Predictive*: reflects that there is uncertainty about the true effect after the original experiment

Design based on significance

Two functions:

- `powerSignificance()` and `sampleSizeSignificance()`

Design based on significance

Two functions:

- `powerSignificance()` and `sampleSizeSignificance()`

Main arguments (`default`):

- `zo`
- `c` (1)
- `power`
- `designPrior` (`"conditional"`)
- `shrinkage` (0)
- `level` (0.025)
- `alternative` (`"one.sided"`)

Example from Pyc and Rawson (2010)

- p-value $p_o = 0.011$
- relative sample size $c = 9.2$

```
# power calculation
powerSignificance(zo = p2z(0.011, alternative = "one.sided"),
                  c = 9.2,
                  designPrior = "conditional")
```

```
## [1] 0.9999997
```

Exercises

(Solutions available at https://osf.io/fcrj6/)

Exercise 2.1

We have five original studies that we want to replicate. The one-sided p-values are 0.0001, 0.001, 0.005, 0.01, and 0.025, respectively. We decide to use the same sample size as in the original study ($c = 1$).

- Compute and plot the conditional and predictive power of the five replication studies. Use the function `powerSignificance()`
- Shrink the original effect estimate by a factor of 25% and use a conditional design prior. How does the power compare to the conditional power without shrinkage?

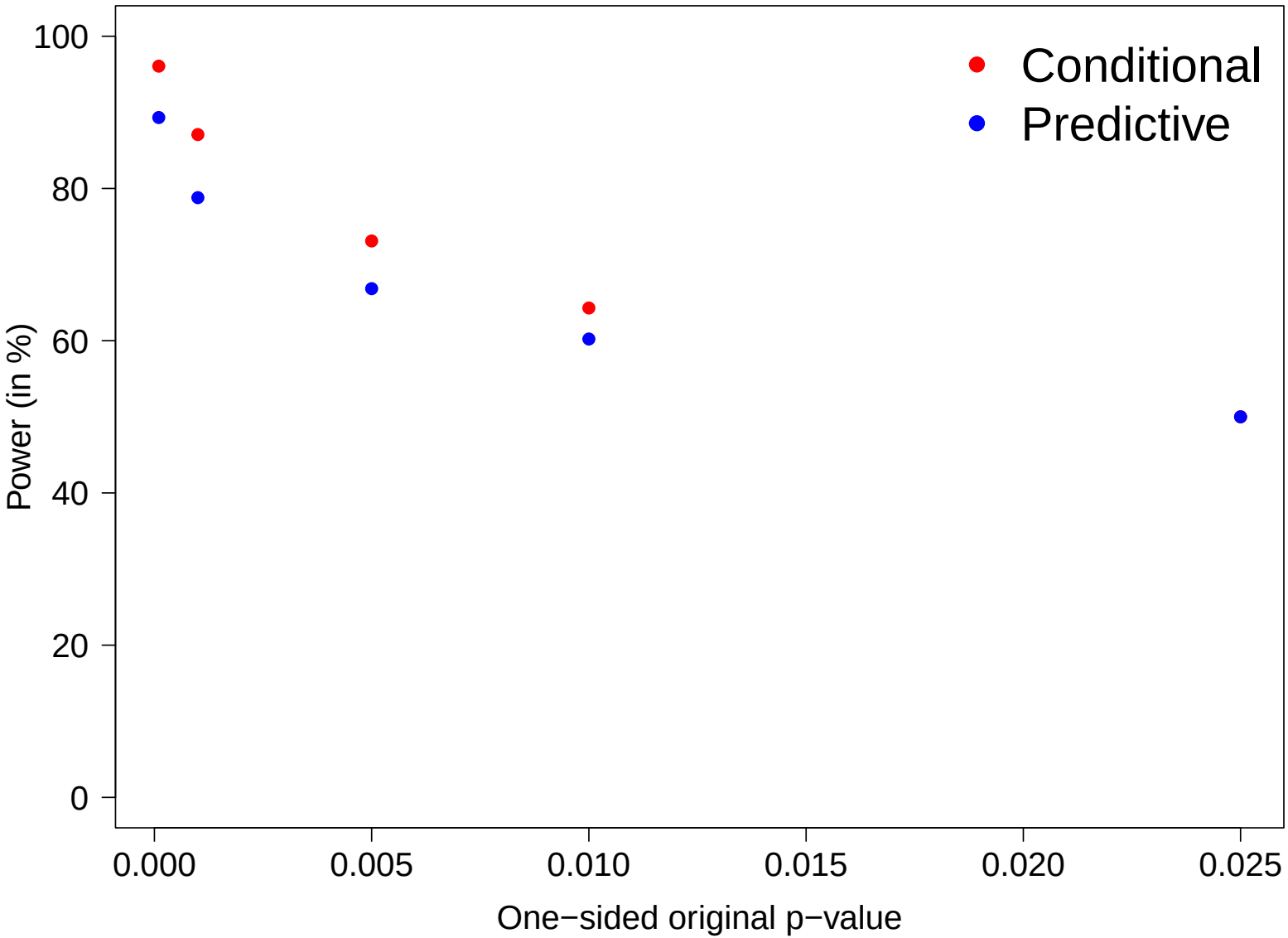
Exercises

(Solutions available at https://osf.io/fcrj6/)

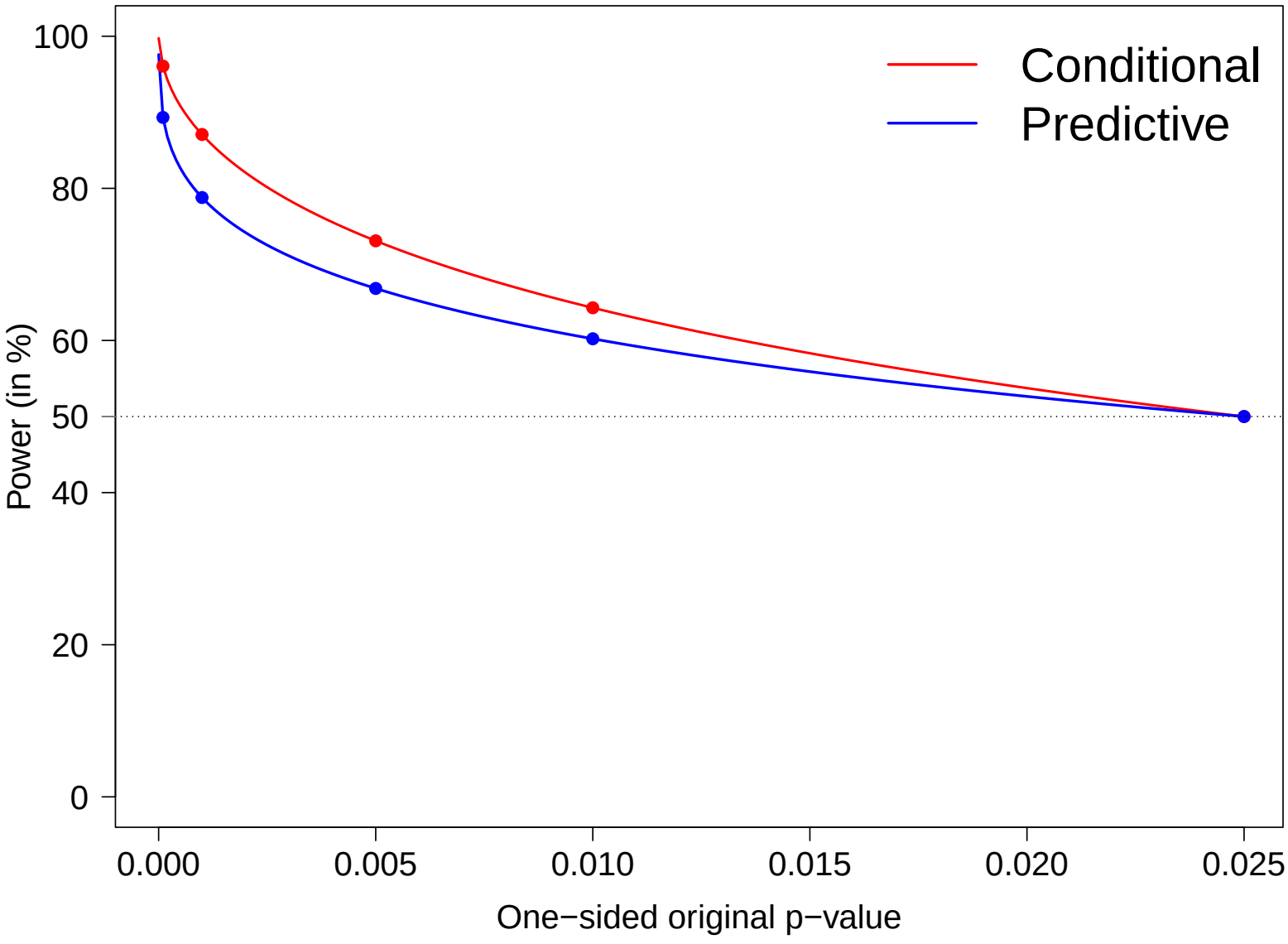
Exercise 2.2

- Compute and plot the relative sample sizes of the five studies to achieve a power of 80% with the conditional and the predictive design prior. Use the function `sampleSizeSignificance()`.
- Shrink the original effect estimate by a factor of 25% and use a conditional design prior. How does the required relative sample size change compared to not shrinking the estimate?

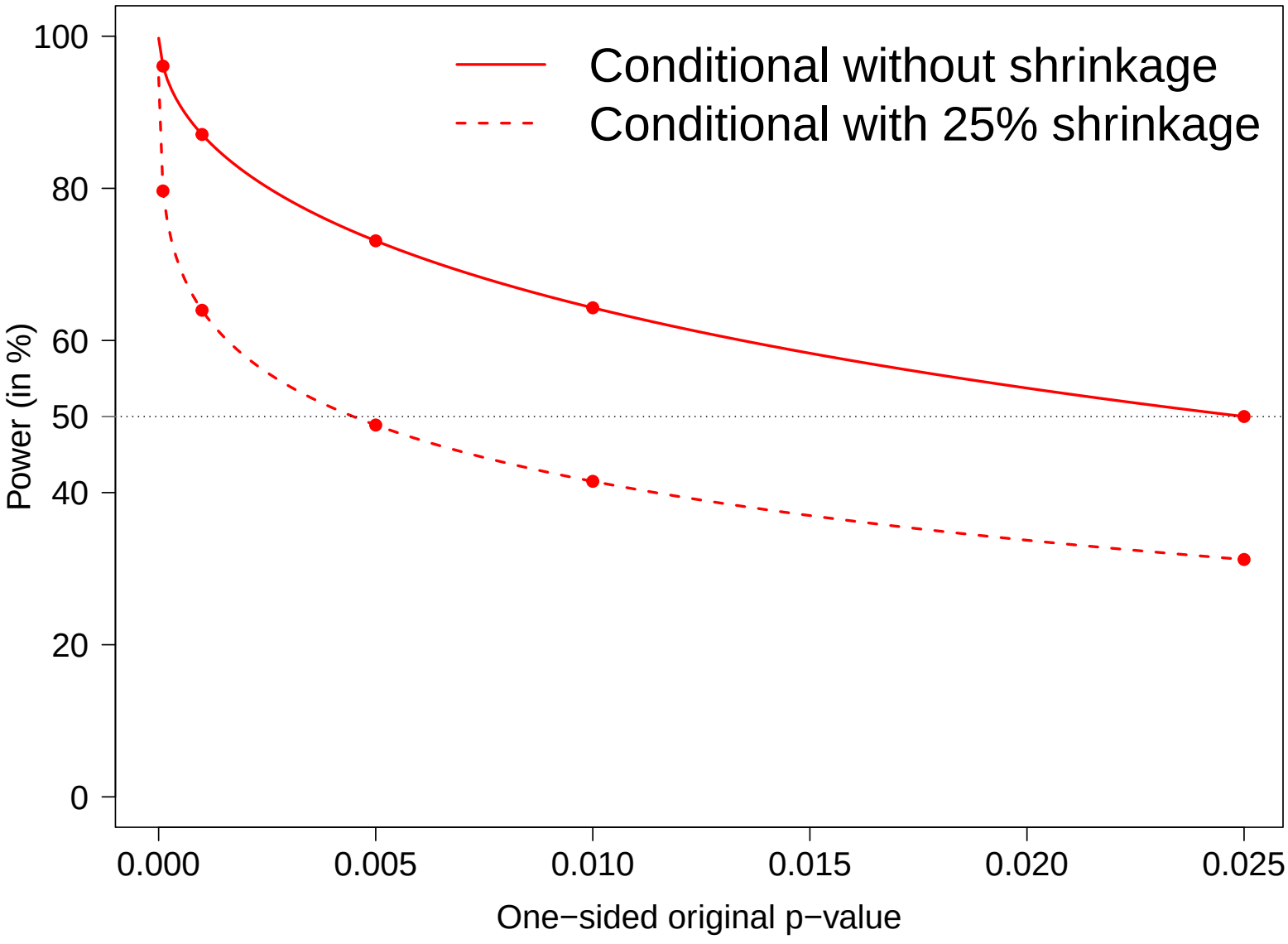
Solution: Exercise 2.1



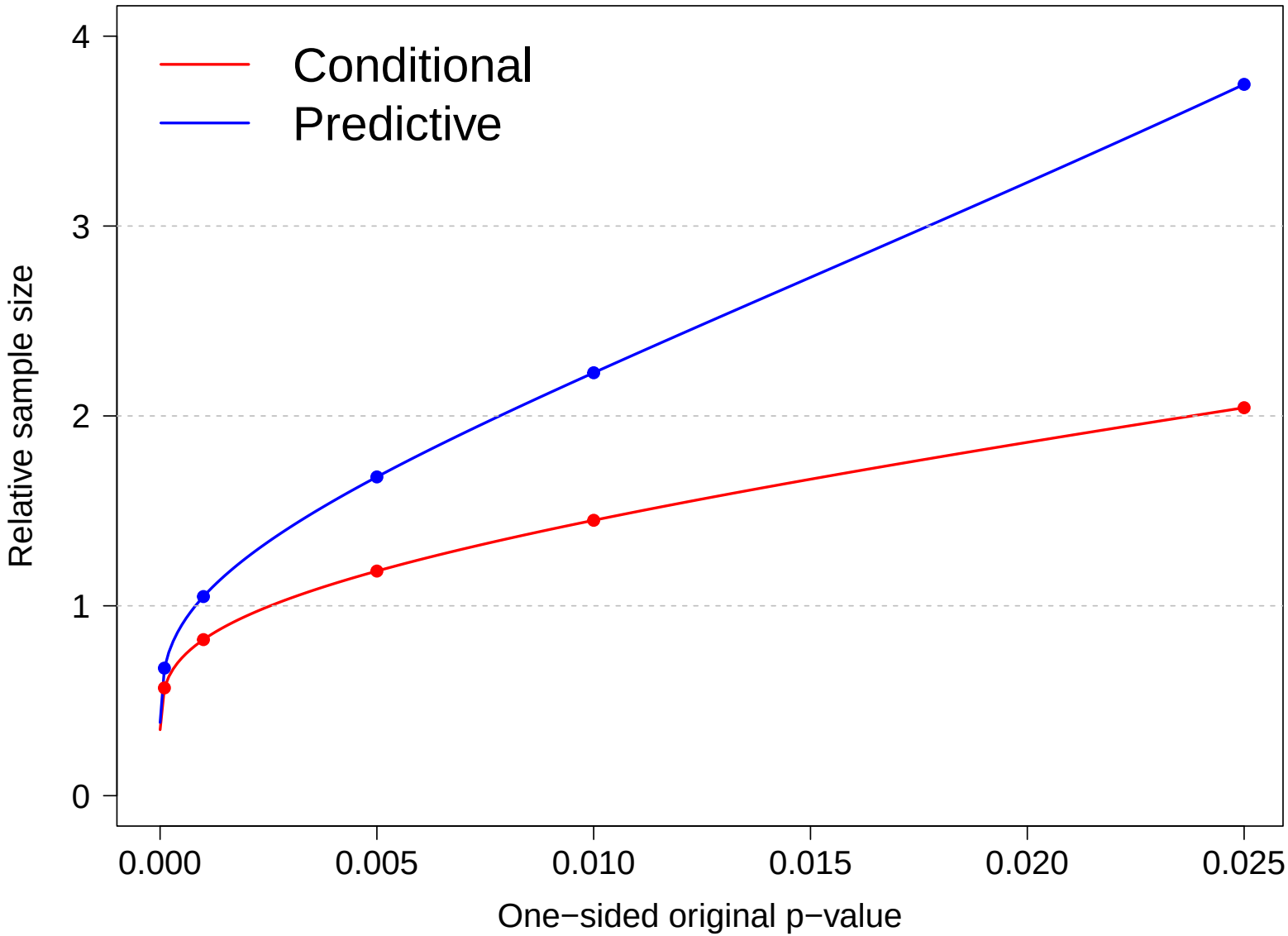
Solution: Exercise 2.1



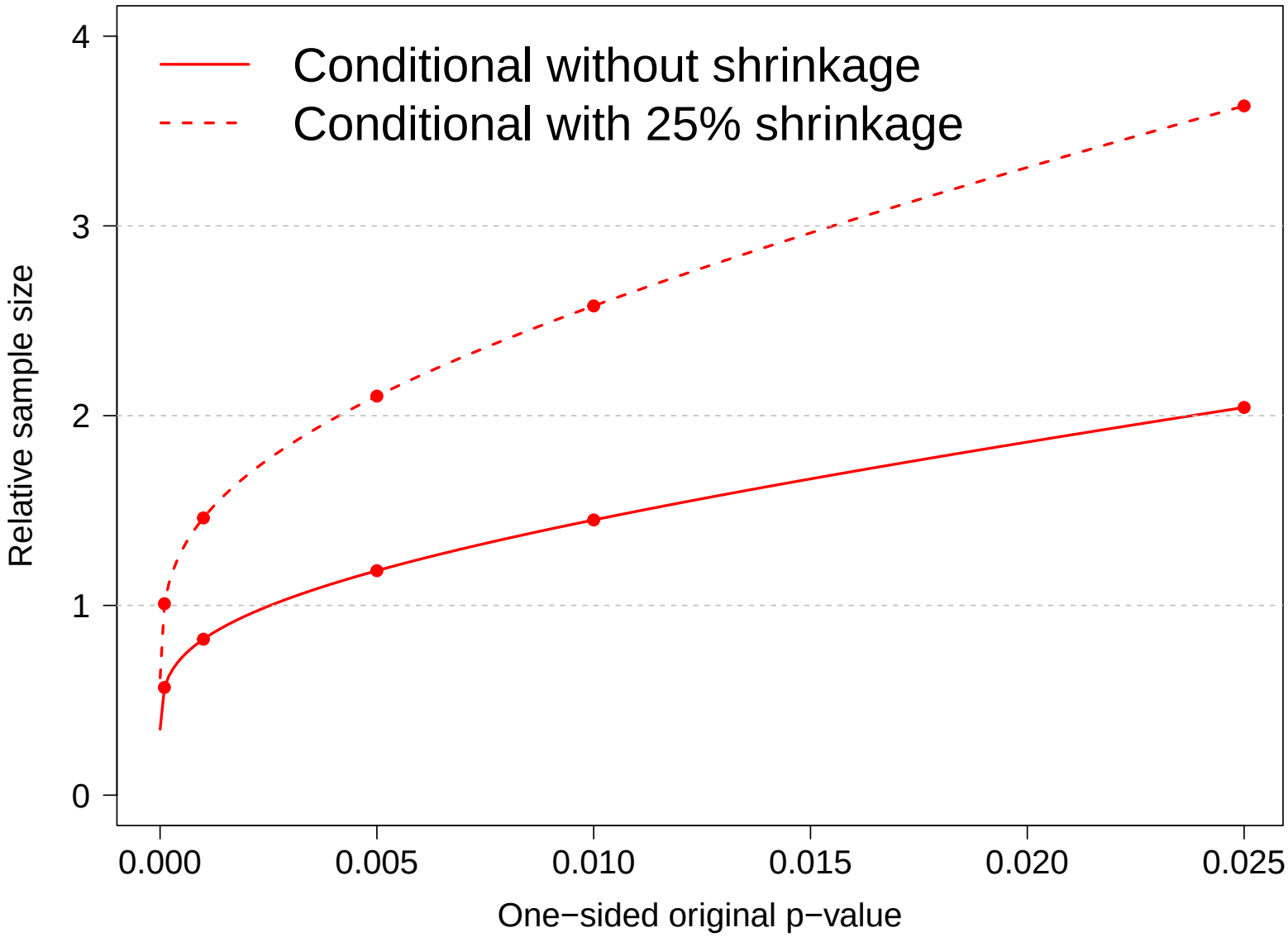
Solution: Exercise 2.1



Solution: Exercise 2.2



Solution: Exercise 2.2



Exercise Session 3

Design based on replication success (sceptical *p*-value)

Design based on replication success

Two functions:

- `powerReplicationSuccess()` and `sampleSizeReplicationSuccess()`

Design based on replication success

Two functions:

- `powerReplicationSuccess()` and `sampleSizeReplicationSuccess()`

Main arguments (`default`):

- `zo`
- `c` (`1`)
- `power`
- `designPrior` (`"conditional"`)
- `level` (`0.025`)
- `alternative` (`"one.sided"`)
- `type` (`"golden"`)

Example from Pyc and Rawson (2010)

- p-value $p_o = 0.011$
- relative sample size $c = 9.2$

```
# power calculation
powerReplicationSuccess(zo = p2z(0.011, alternative = "one.sided"),
                        c = 9.2,
                        designPrior = "conditional")
```

```
## [1] 0.9923838
```

Exercises

(Solutions available at https://osf.io/fcrj6/)

Exercise 3.1

- Compute and plot the conditional and predictive power for replication success. Use the function `powerReplicationSuccess()` with $c = 1$ and $p_o = 0.0001, 0.001, 0.005, 0.01$ and 0.025.
- Compare conditional power for replication success with conditional power for significance (exercise 2.1).

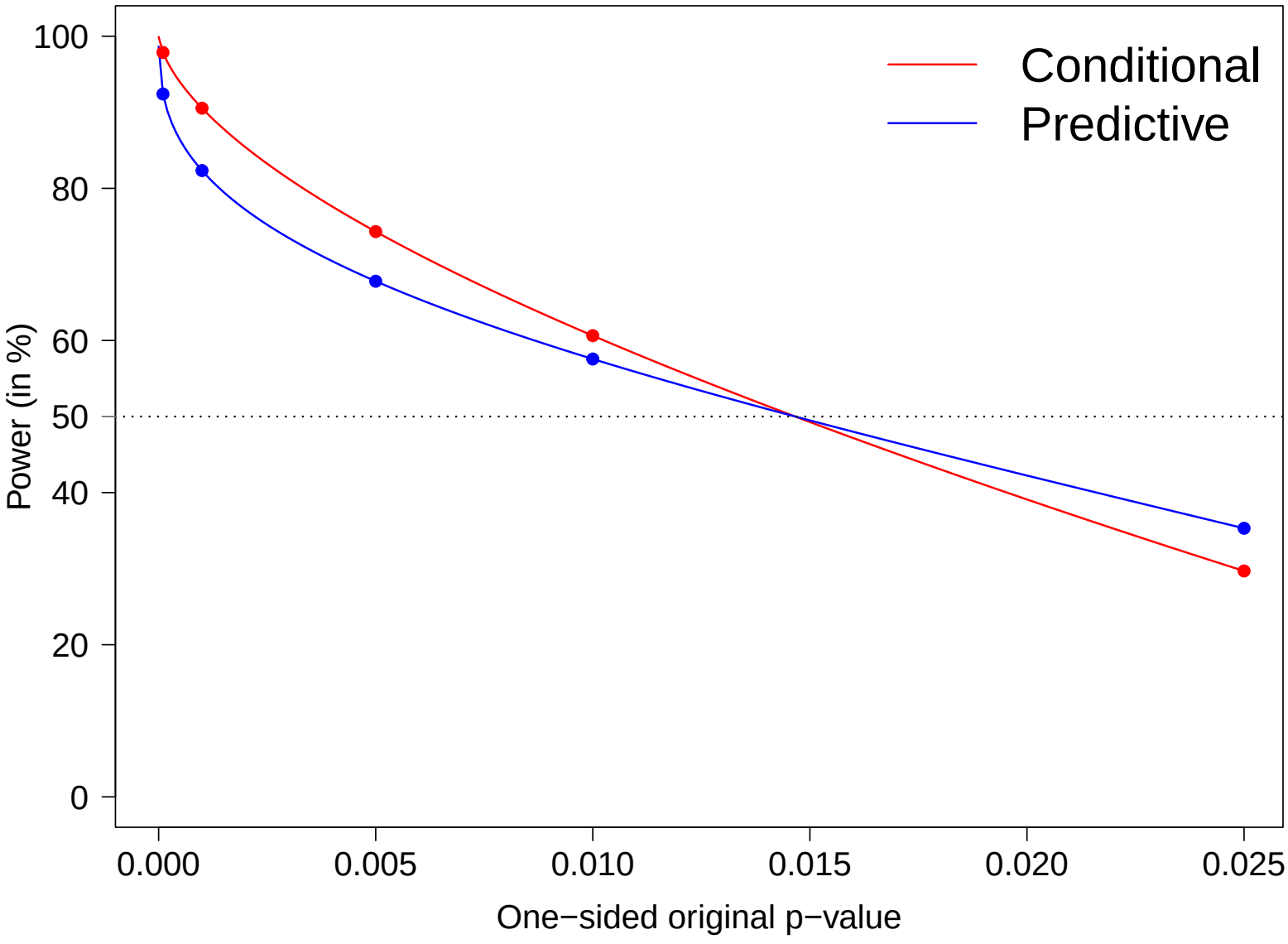
Exercises

(Solutions available at https://osf.io/fcrj6/)

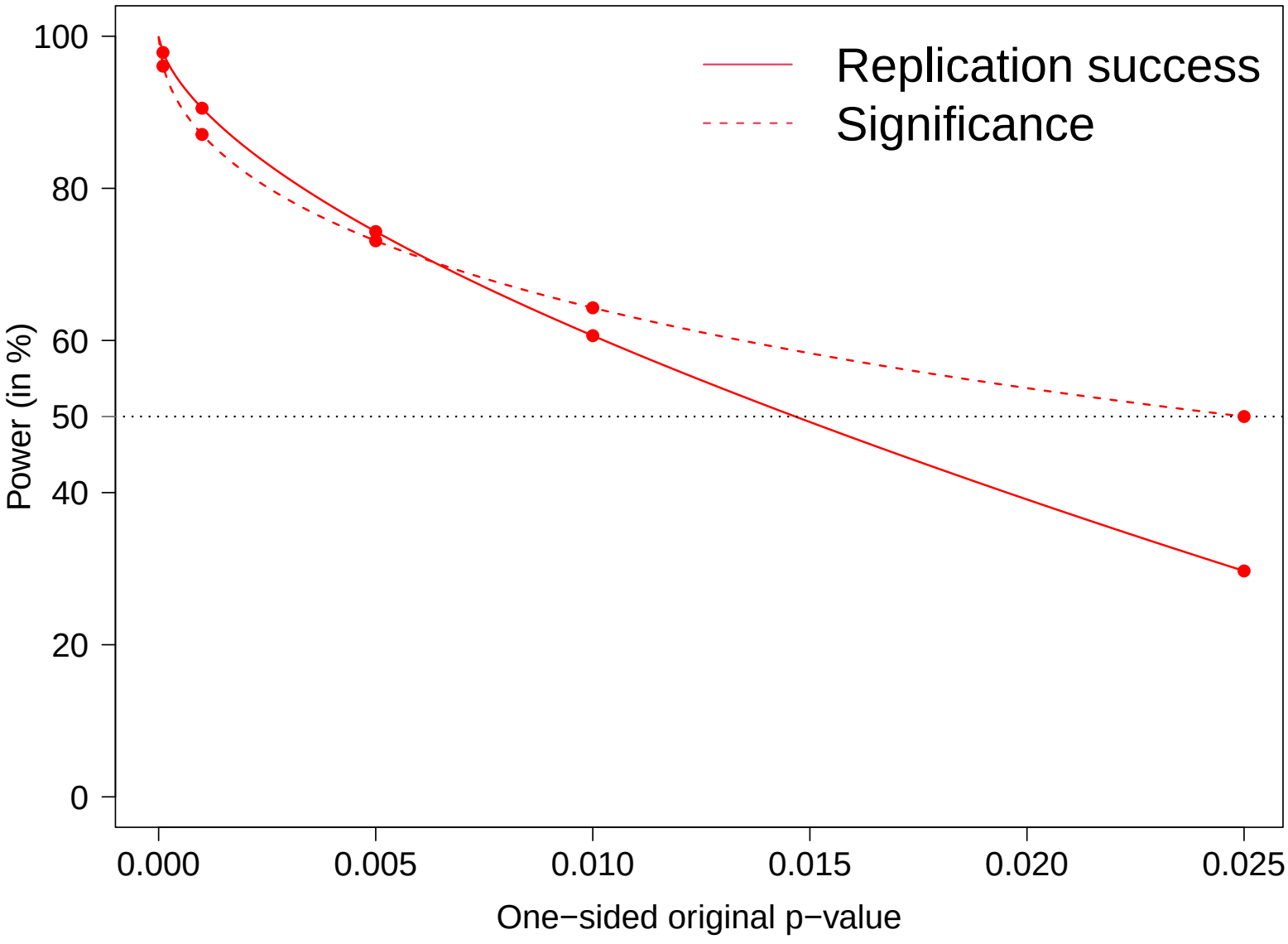
Exercise 3.2

- Compute and plot the relative sample sizes of the five studies to achieve a power of 80% with the conditional and the predictive design prior. Use the function `sampleSizeReplicationSuccess()`.
- Compare the relative sample sizes with the ones obtained in exercise 2.2 (only for the conditional design prior).

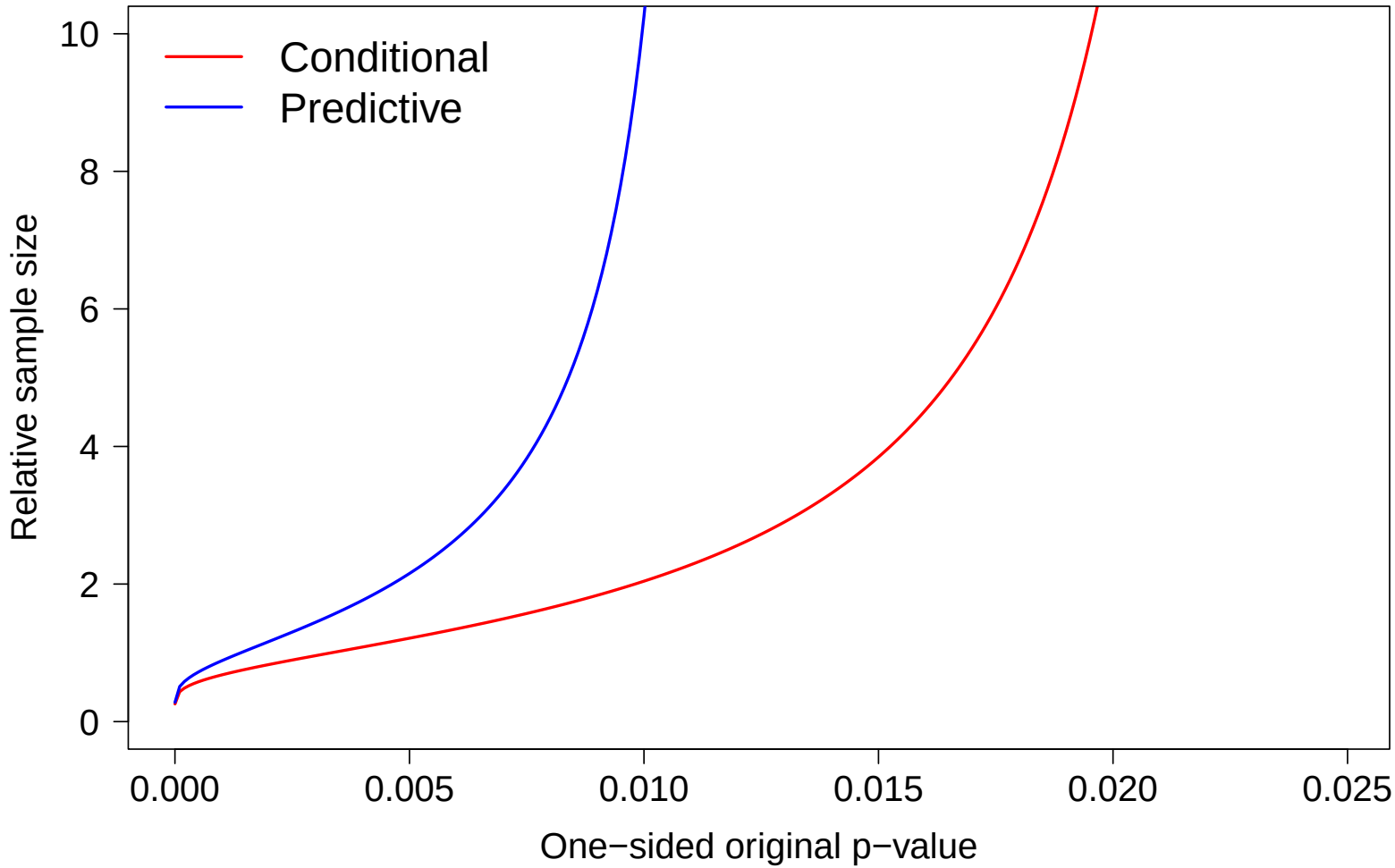
Solution: Exercise 3.1



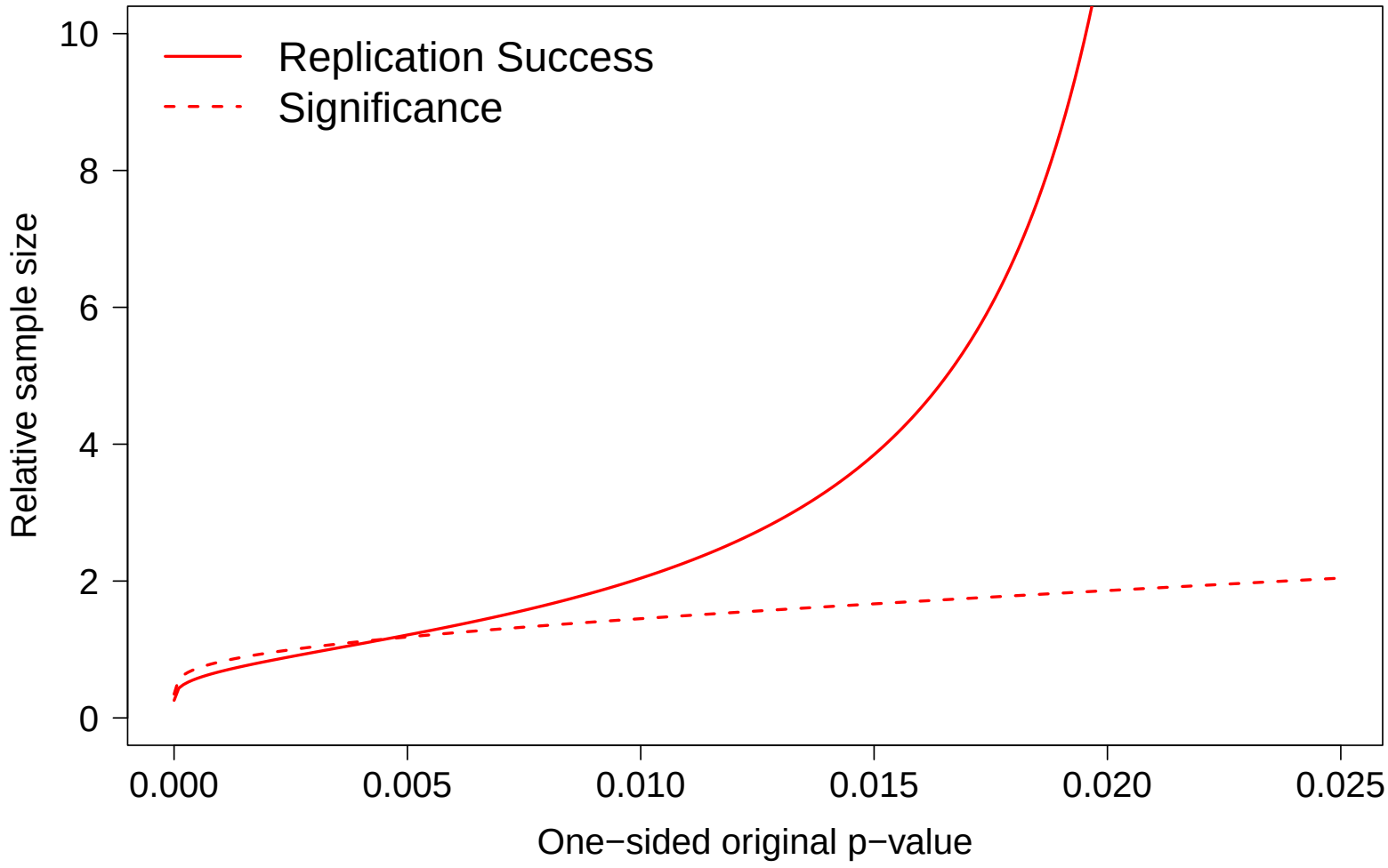
Solution: Exercise 3.1



Solution: Exercise 3.2



Solution: Exercise 3.2



Outlook

- Between-study heterogeneity
  $\rightarrow$ relative heterogeneity `h` can be specified in some functions
- Data-driven shrinkage with empirical Bayes
  $\rightarrow$ `designPrior = "EB"`
- Interim analysis
  $\rightarrow$ `powerSignificanceInterim()`
- Sample size based on relative effect size

References

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.

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.

Cova, F., Strickland, B., Abatista, A., Allard, A., Andow, J., Attie, M., Beebe, J., Berniūnas, R., Boudesseul, J., Colombo, M., Cushman, F., Diaz, R., N'Djaye Nikolai van Dongen, 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., Liao, S.-y., 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*.

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

Held, L. (2020). 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.

Held, L., Micheloud, C., and Pawel, S. (2020). The assessment of replication success based on relative effect size. Technical report.

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

Pawel, S. and Held, L. (2020). Probabilistic forecasting of replication studies. *PLOS ONE*, 15(4):e0231416.

Pyc, M. A. and Rawson, K. A. (2010). Why testing improves memory: Mediator effectiveness hypothesis. *Science*, 330(6002):335–335.