Tutorial on the R package ReplicationSuccess

Leonhard Held, Charlotte Micheloud, Samuel Pawel

Department of Biostatistics, Center for Reproducible Science

CRS@UZH

www.crs.uzh.ch

22.01.2020

Background

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

$\rightarrow$ Tool to assess credibility of scientific discoveries

$\rightarrow$ Regulatory requirement

Replication crisis

– Low replicability of many scientific discoveries

$\rightarrow$ 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

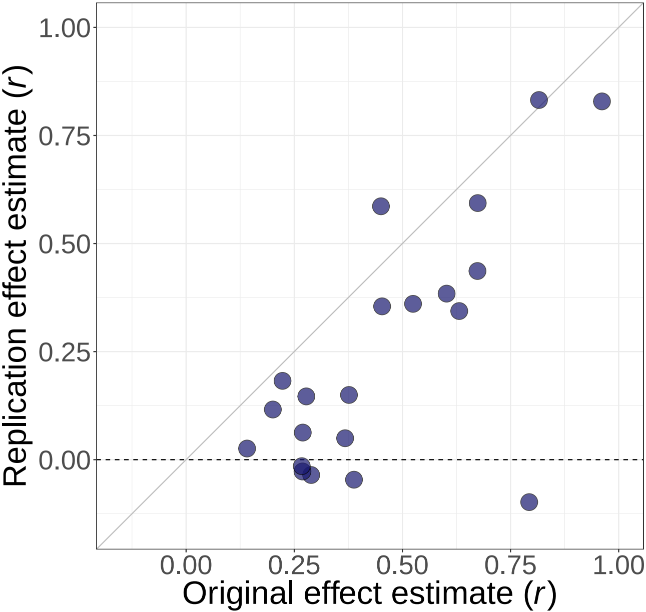
Evaluating the replicability of social science experiments in *Nature* and *Science* between 2010 and 2015

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

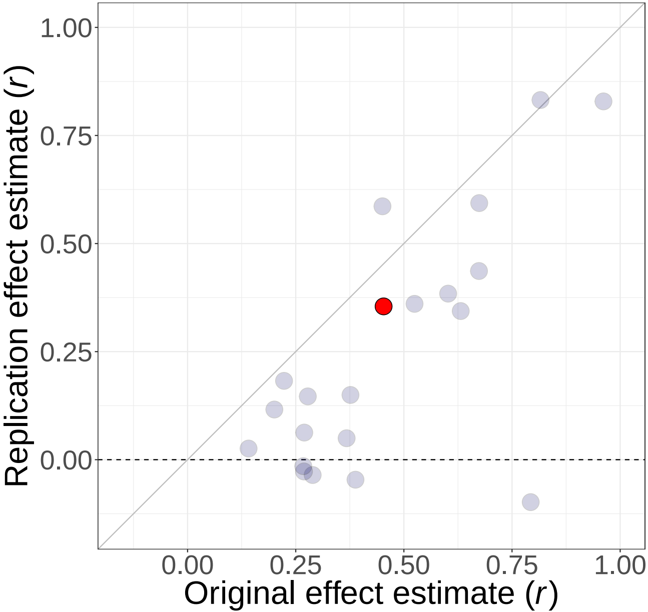
Social sciences replication project

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

Social sciences replication project



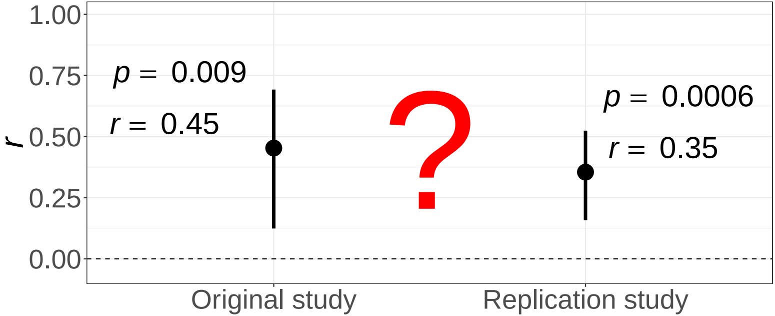
Social sciences replication project



Morewedge et al. (2010). Science

Original discovery

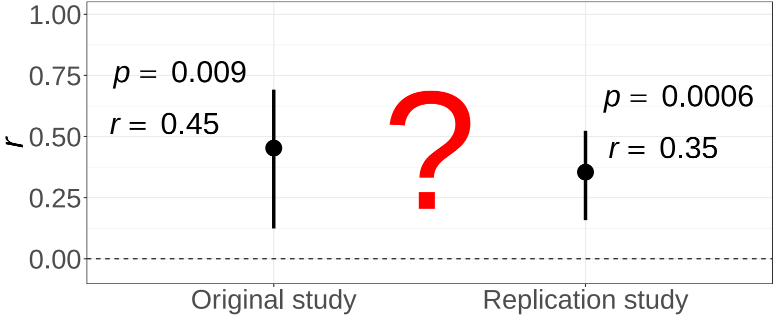
"Repeatedly imagining eating a food subsequently reduces the actual consumption of that food"



When is a replication successful?

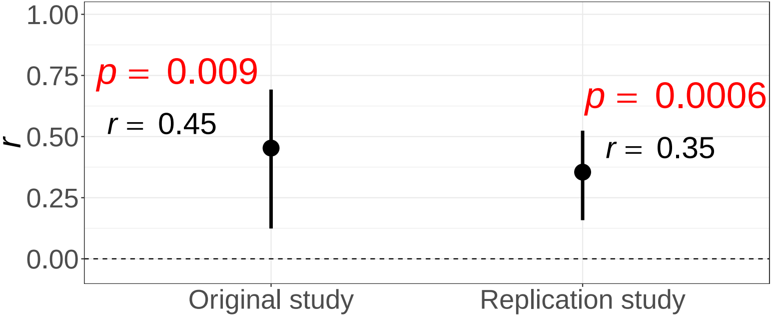
Some proposed criteria

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



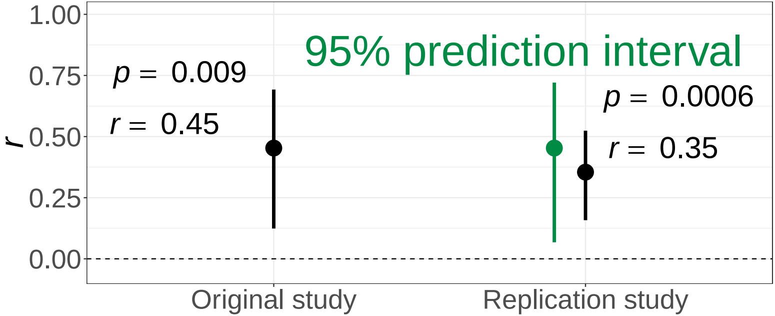
1. Statistical significance

Are original and replication estimates statistically significant?



2. Compatibility of effect estimates

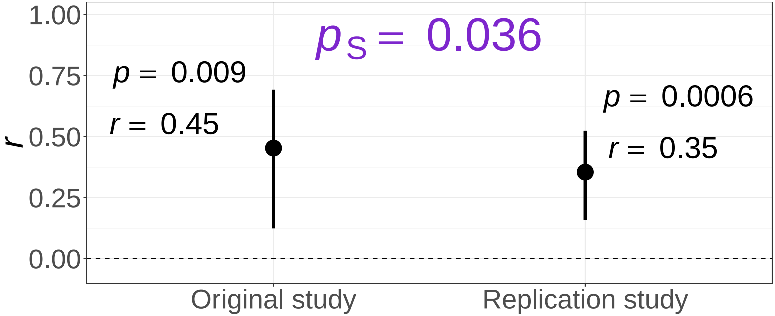
Is the replication estimate contained in its prediction interval?
$\rightarrow$ function: `predictionInterval()`



3. Sceptical *p*-value

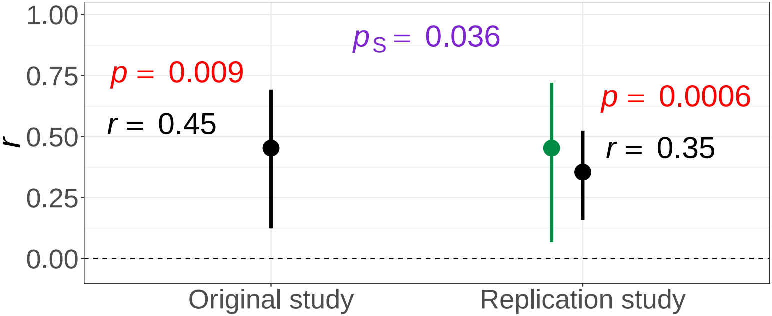
Can we convince a sceptic whose priof beliefs make the original study not significant?

→ function: `pSceptical()`



Drawbacks of classical approaches

- Significance can always be achieved by increasing sample size
- Estimates can be compatible but provide no information about true effect



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 proper sample size calculation is essential and depends on analysis strategy

Design of replication studies

What is used in practice

– Standard sample size calculation

```
sampleSizeZtest = function(delta, sd, sig.level = 0.05, power){
  u <- qnorm(p = power)
  v <- qnorm(p = 1 - sig.level/2)
  n <- 2*(u + v)^2*sd^2/delta^2
  return(n)
}
sampleSizeZtest(delta = 0.25, sd = 0.4, sig.level = 0.01, power = 0.95)

## [1] 91.20852
```

Design of replication studies

What is used in practice

- Standard sample size calculation

```
sampleSizeZtest = function(delta, sd, sig.level = 0.05, power){
  u <- qnorm(p = power)
  v <- qnorm(p = 1 - sig.level/2)
  n <- 2*(u + v)^2*sd^2/delta^2
  return(n)
}
sampleSizeZtest(delta = 0.25, sd = 0.4, sig.level = 0.01, power = 0.95)

## [1] 91.20852
```

- Goal is to have between 80% and 95% power in the replication study to detect the effect estimate from the original study
- Shrinkage of the original effect estimate is sometimes used

Design of replication studies

Issues with standard sample size calculation

- Uncertainty of original effect estimate is ignored
- Heterogeneity between original and replication study is not taken into account
- Arbitrary shrinkage methods

Package

- Functionalities for design and analysis of replication studies
  $\rightarrow$ Traditional methods
  $\rightarrow$ Sceptical *p*-value (Held, 2019)

J. R. Statist. Soc. A (2020)

A new standard for the analysis and design of replication studies

Leonhard Held

University of Zurich, Switzerland

Package

- Functionalities for design and analysis of replication studies
  → Traditional methods
  → Sceptical *p*-value (Held, 2019)

J. R. Statist. Soc. A (2020)

A new standard for the analysis and design of replication studies

Leonhard Held

University of Zurich, Switzerland

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

Statistical framework of package

– Effect estimates are assumed to be normally distributed
  - $\rightarrow$ usually fulfilled after suitable transformation
  - $\rightarrow$ Fisher's z-transformation for correlation coefficients r

Statistical framework of package

- Effect estimates are assumed to be normally distributed

  $\rightarrow$ usually fulfilled after suitable transformation

  $\rightarrow$ Fisher's z-transformation for correlation coefficients r

- Design prior

  $\rightarrow$ Conditional: ignores uncertainty of original study

  $\rightarrow$ Predictive: reflects that there is still uncertainty about the true effect after the original experiment

Statistical framework of package

Key quantities

– relative sample size $c = n_r / n_o$

```
ReplicationProjects$c <- with(ReplicationProjects, z_se_O^2/z_se_R^2)
```

Statistical framework of package

Key quantities

- relative sample size $c = n_r / n_o$

```r
ReplicationProjects$c <- with(ReplicationProjects, z_se_O^2/z_se_R^2)
```

- *p*-value or test statistic of original study

```r
ReplicationProjects$to <- with(ReplicationProjects, z_O/z_se_O)
ReplicationProjects$po <- t2p(ReplicationProjects$to)
ReplicationProjects$to <- p2t(ReplicationProjects$po)
```

Statistical framework of package

Key quantities

- relative sample size $c = n_r / n_o$

```
ReplicationProjects$c <- with(ReplicationProjects, z_se_O^2/z_se_R^2)
```

- p-value or test statistic of original study

```
ReplicationProjects$to <- with(ReplicationProjects, z_O/z_se_O)
ReplicationProjects$po <- t2p(ReplicationProjects$to)
ReplicationProjects$to <- p2t(ReplicationProjects$po)
```

- p-value or test statistic of replication study

```
ReplicationProjects$tr <- with(ReplicationProjects, z_R/z_se_R)
ReplicationProjects$pr <- t2p(ReplicationProjects$tr)
```

Application

Installation

– Linux / Windows

```r
install.packages(pkgs = "ReplicationSuccess",
                 repos ="http://R-Forge.R-project.org")
```

– Mac

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

Application

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

1. Statistical significance

Two functions:

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

1. Statistical significance

Two functions:

– `powerSignificance()` and `sampleSizeSignificance()`

Main arguments:

– `po` or `to`
– `c`
– `power`
– `designPrior`
– `shrinkage`
– `level`
– `alternative`

1. Statistical significance

Example from Morewedge et al. (2010)

- $t_o = 2.63$
- $p_o = 0.009$
- $c = n_r / n_o = 3$

```
# power calculation
powerSignificance(po = 0.009, c = 3, designPrior = "conditional")

## [1] 0.99483

# sample size calculation
sampleSizeSignificance(to = 2.63, power = 0.9, designPrior = "predictive")

## [1] 2.927087
```

1. Statistical significance

Exercise 1.1

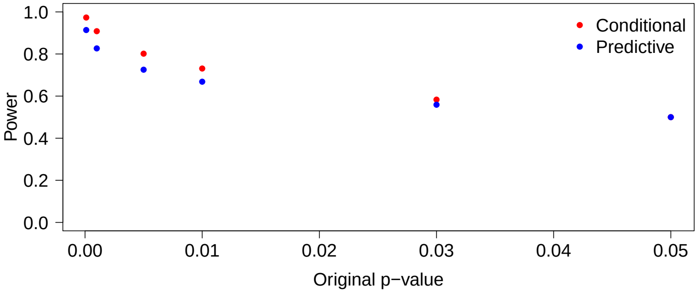
We have six original studies that we want to replicate. Their p-values are 0.0001, 0.001, 0.005, 0.01, 0.03 and 0.05, respectively. We decide to simply use the same sample size as in the original study.

- Compute the conditional and predictive power of the six replication studies and plot it.
- What do you notice?
- What happens if we decide to take less subjects in the replication study as compared to the original study?

1. Statistical significance

Exercise 1.1 - Solutions

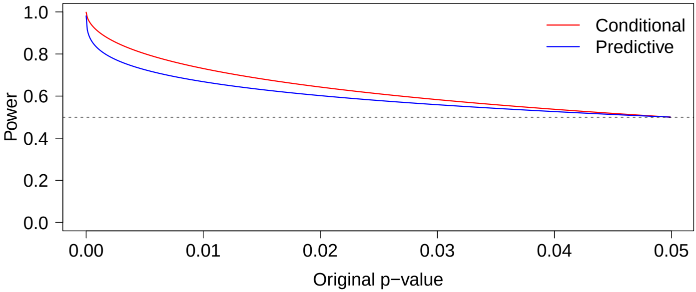
- Compute the conditional and predictive power of the six replication studies and plot it.
- What do you notice?



1. Statistical significance

Exercise 1.1 - Solutions

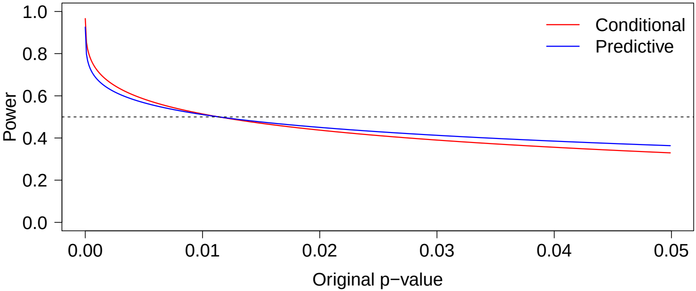
- Compute the conditional and predictive power of the six replication studies and plot it.
- What do you notice?



1. Statistical significance

Exercise 1.1 - Solutions

- What happens if we decide to take less subjects in the replication study as compared to the original study?
  c = 0.6



1. Statistical significance

Exercise 1.2

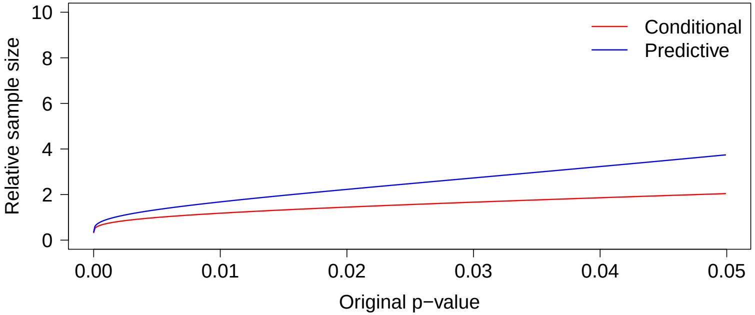
We now know that taking the same sample size as in the original study is not optimal and want to perform a proper sample size calculation.

- Compute and plot the relative replication sample sizes of the six studies to achieve a power of 80% with the conditional and the predictive design prior.
- What happens if the desired power is now 90%?

1. Statistical significance

Exercise 1.2- Solutions

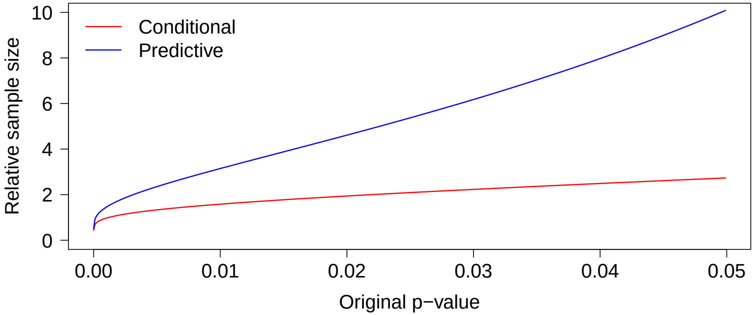
– Compute and plot the relative replication sample sizes of the six studies to achieve a power of 80% with the conditional and the predictive design prior.



1. Statistical significance

Exercise 1.2- Solutions

- What happens if the desired power is now 90%?



1. Statistical significance

Exercise 1.3

We are now interested in the Experimental economics projects.

- Compute the required replication sample size to reach a power of 90% for each study of the project and with the conditional and the predictive design prior.
- What do you notice?

```
data("ReplicationProjects")
eco <- subset(ReplicationProjects, project == "Experimental Economics")
```

1. Statistical significance

Exercise 1.3 - Solutions

- Compute the required replication sample size to reach a power of 90% for each study of the project and with the conditional and the predictive design prior.
- What do you notice?

$\rightarrow$ Most of the required replication sample size are above one with conditional design prior
$\rightarrow$ Predictive design prior gives larger sample sizes than conditional design prior

1. Statistical significance

Exercise 1.4

Some original studies belonging to the psychology data set were not statistically significant at the two-sided 5%-level. This is the case for the study from Reynolds and Besner (2008), for example.

- Compute the required replication sample size to reach a power of 95% for this study with the conditional and the predictive design prior.

```
reynolds <- subset(ReplicationProjects, study == "M Reynolds, D Besner")
```

Exercise 1.4 - Solutions

- Compute the required replication sample size to reach a power of 95% for this study with the conditional and the predictive design prior.
- $p_o = 0.12$

```
sampleSizeSignificance(po = reynolds$pval_O,
                       power = 0.95,
                       designPrior = "conditional")

## [1] 5.300107

sampleSizeSignificance(po = reynolds$pval_O,
                       power = 0.95,
                       designPrior = "predictive")

## Error in sampleSizeSignificance(po = reynolds$pval_O, power
= 0.95, designPrior = "predictive"):  power too large, power
should not exceed 0.941
```

$\rightarrow$ predictive power is bounded by (1- one-sided *p*-value of original study)

2. Compatibility of effect estimates

Two functions:

– `sampleSizePI()` and `sampleSizePIwidth()`

2. Compatibility of effect estimates

Two functions:

- `sampleSizePI()` and `sampleSizePIwidth()`

Main arguments

- `to` or `po`
- `w`
- `conf.level`
- `designPrior`

2. Compatibility of effect estimates

Example from Morewedge et al. (2010)

- $t_o = 2.63$
- $p_o = 0.009$

```
# fix prediction interval limit to 0
sampleSizePI(to = 2.63, designPrior = "predictive")

## [1] 1.249076

# fix relative width of prediction interval
sampleSizePIwidth(w = 1.25, designPrior = "predictive")

## [1] 1.777778
```

2. Compatibility of effect estimates

Exercise 2.1

a) You have five original studies for which you want to conduct replication studies. The test statistics are 2, 2.5, and 3. How much do you need to change the sample size such that a 95% prediction interval of the replication estimate does not include 0?

b) How much do you need to change the sample size such that a 95% prediction interval of the replication estimate is only 25% wider than the confidence interval from the original estimate?

2. Compatibility of effect estimates

Exercise 2.1

a)
```r
to <- c(1.5, 2, 2.5, 3)
sampleSizePI(to = to, designPrior = "predictive")

## [1]         NA 24.2300381  1.5949318  0.7446793
```

b)
```r
w <- 1.25
sampleSizePIwidth(w = w, designPrior = "predictive")

## [1] 1.777778
```

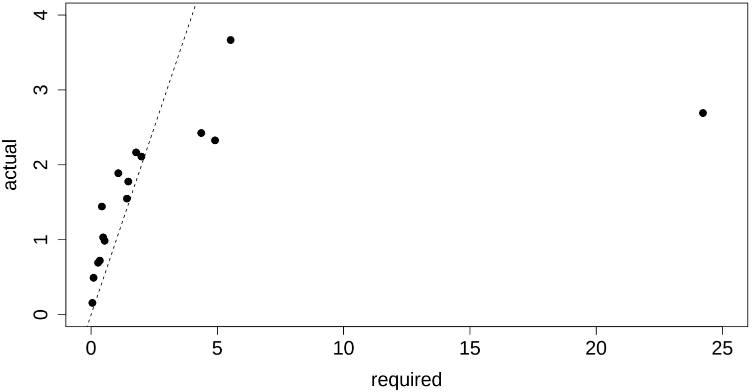
2. Compatibility of effect estimates

Exercise 2.2

– For the replications from experimental economics project compute the required relative sample size for the 95% prediction intervals of the replication estimates not to contain zero. Compare them to the actually used relative sample sizes.

2. Compatibility of effect estimates

```
required_c <- sampleSizePI(to = eco$to, designPrior = "predictive")
```



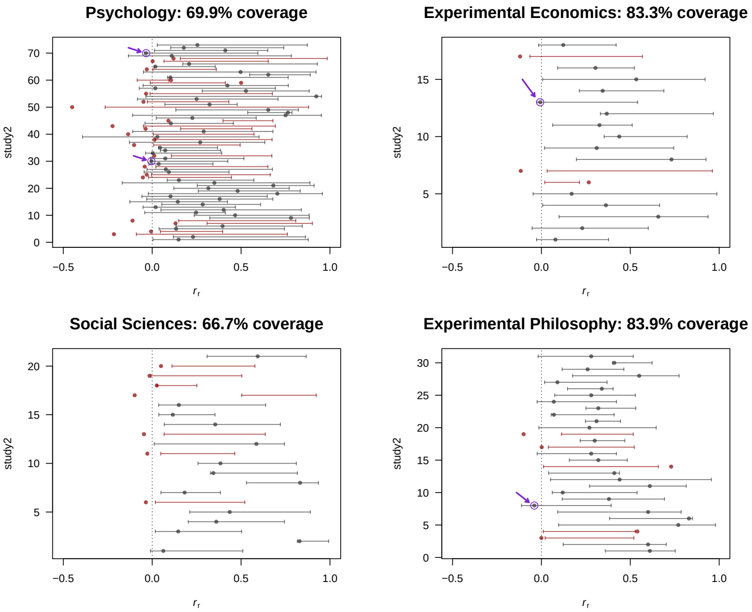
2. Compatibility of effect estimates

Exercise 2.3

a) Look at the documentation of the function `predictionInteval()` with `?predictionInteval`. Run the example code at the bottom to compute and plot the 95% prediction intervals for all four replication projects. Interpret the results.

b) Which situations could have been avoided by more careful design of the replication studies?

2. Compatibility of effect estimates



3. Sceptical *p*-value

Two functions:

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

3. Sceptical *p*-value

Two functions:

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

Main arguments:

- `po` or `to`
- `c`
- `power`
- `designPrior`
- `level`
- `alternative`

3. Sceptical p-value

Example from Morewedge et al. (2010)

- $t_o = 2.63$
- $p_o = 0.009$
- $c = n_r/n_o = 3$

```
# sample size calculation
sampleSizeReplicationSuccess(to = 2.63, power = 0.9,
                             designPrior = "predictive",
                             alternative = "one.sided",
                             level = 0.065)

## [1] 5.673493
```

3. Sceptical *p*-value

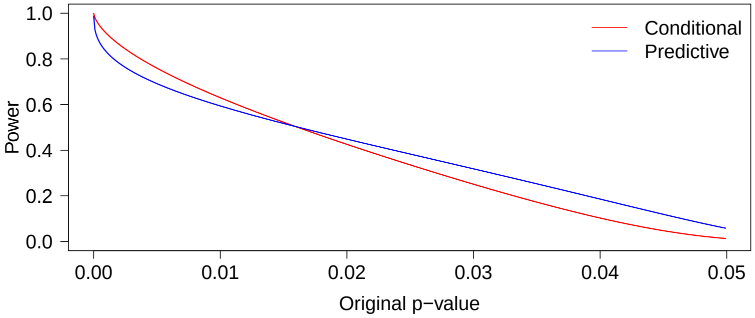
Exercise 3.1

- Compute and plot the conditional and predictive power for Replication Success of the 6 studies from exercise 1.1, using an alpha level of 0.065 and a one-sided alternative
- How does the plot compare with the one from exercise 1.1?

3. Sceptical *p*-value

Exercise 3.1 - Solutions

- Compute and plot the conditional and predictive power for Replication Success of the 6 studies from exercise 1.1, using an alpha level of 0.065 and a one-sided alternative



3. Sceptical *p*-value

Exercise 3.2

– For the replications from experimental economics project compute the required relative sample size to reach a power for Replication Success of 90%. Use the conditional and the predictive design prior, a level of 0.065 and a one-sided alternative.

– Compare them to the actually used relative sample sizes.

3. Sceptical *p*-value

Exercise 3.2 - Solutions

```
par(mfrow = c(1,2), las = 1)
sampleSizeReplicationSuccess(po = eco$pval_O, power = 0.90, level = 0.065,
                             alternative = "one.sided",
                             designPrior = "conditional")

##  [1]        Inf        Inf 2.00793709 3.05422230        Inf 0.92256435
##  [7] 4.84305551 0.60795643 3.26348978 0.18600082 0.82465494 0.74237959
## [13]        Inf 0.09905365 6.46577335 0.49696693        Inf        Inf

sampleSizeReplicationSuccess(po = eco$pval_O, power = 0.90, level = 0.065,
                             alternative = "one.sided",
                             designPrior = "predictive")

##  [1]        Inf        Inf 40.5344149        Inf        Inf  1.5503100
##  [7]        Inf  0.8278017        Inf  0.2029201  1.2900901  1.0986378
## [13]        Inf  0.1037199        Inf  0.6352722        Inf        Inf
```

Outlook

- Between-study heterogeneity
  $\rightarrow$ argument in most functions `d`
- Data-driven shrinkage with empirical Bayes
  $\rightarrow$ `designPrior = "EB"`
- Interim analysis
  $\rightarrow$ `powerSignificanceInterim()`

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*.

Held, L. (2019). A new standard for the analysis and design of replication studies (with discussion). *Journal of the Royal Statistical Society: Series A (Statistics in Society)*.

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

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

Pawel, S. and Held, L. (2019). Probabilistic forecasting of replication studies. Preprint.

Reynolds, M. and Besner, D. (2008). Contextual effects on reading aloud: Evidence for pathway control. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34(1):50 – 64.