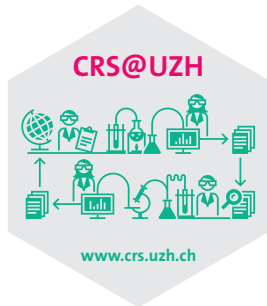


Tutorial on the R package ReplicationSuccess

Leonhard Held, Charlotte Micheloud, Samuel Pawel

Department of Biostatistics, Center for Reproducible Science



Theory

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")
```

Replication studies

Direct replication

- Repeating original study using the same methodology
- Tool to assess credibility of scientific discoveries
- 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
- Increased interest in meta-science
- Large-scale replication projects

Large-scale replication projects

- 2015: Reproducibility project psychology

Large-scale replication projects

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

Large-scale replication projects

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

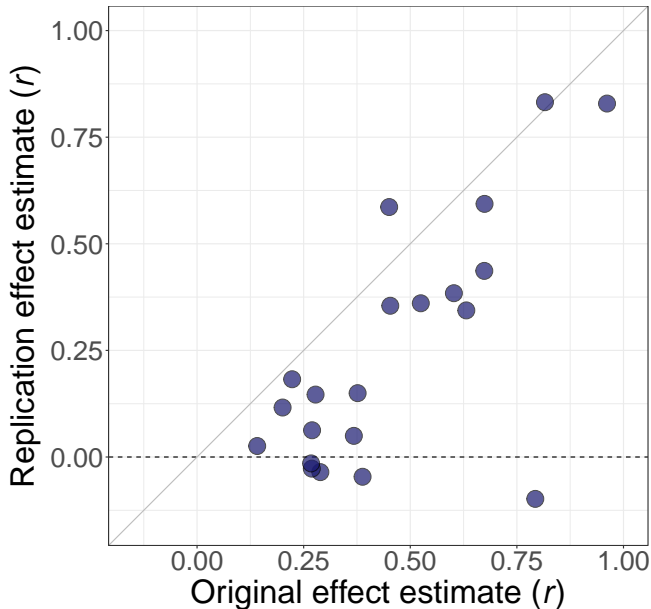
Large-scale replication projects

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

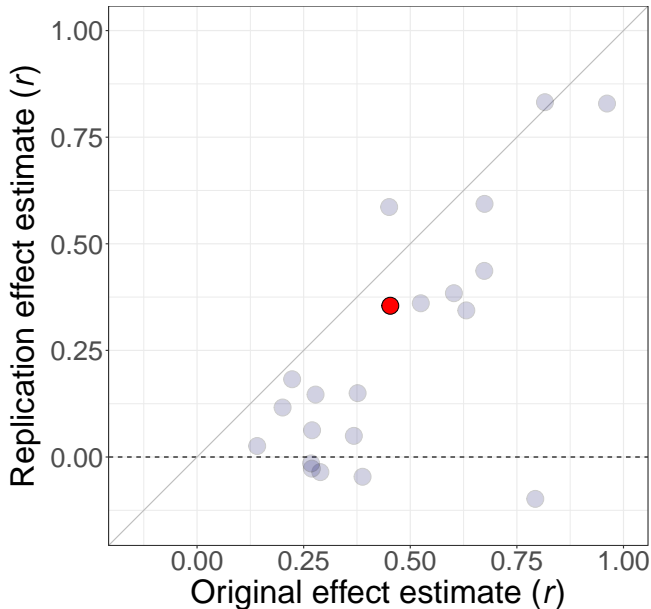
Large-scale replication projects

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

Social sciences replication project



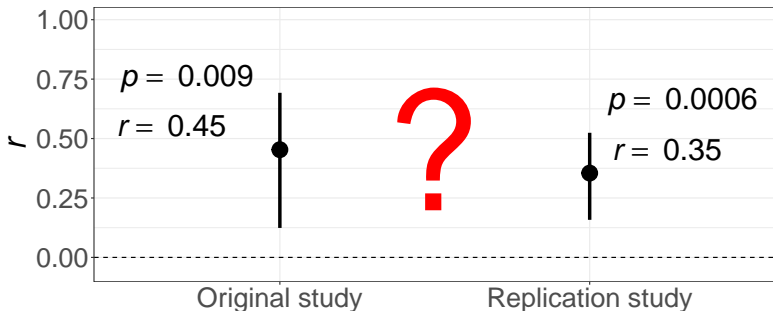
Social sciences replication project



Morewedge et al. (2010). Science

Original discovery

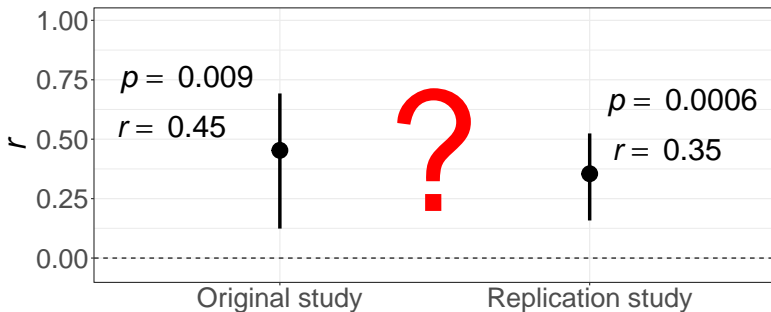
“Repeatedly imagining eating a food subsequently reduces the actual consumption of that food”



When is a replication successful?

Some proposed criteria

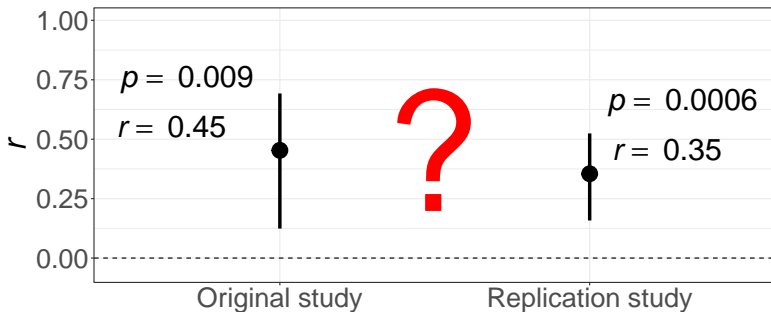
1. Statistical significance



When is a replication successful?

Some proposed criteria

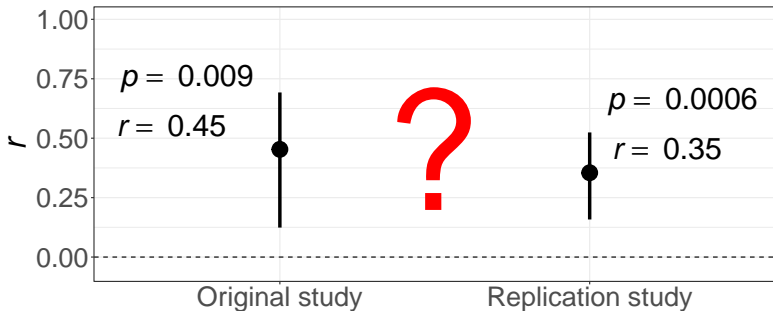
1. Statistical significance
2. Compatibility of effect estimates



When is a replication successful?

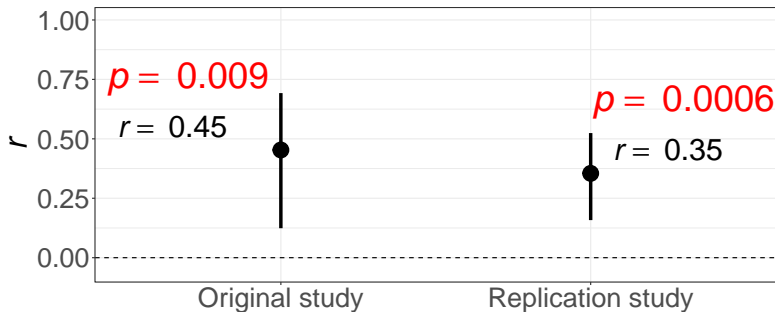
Some proposed criteria

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



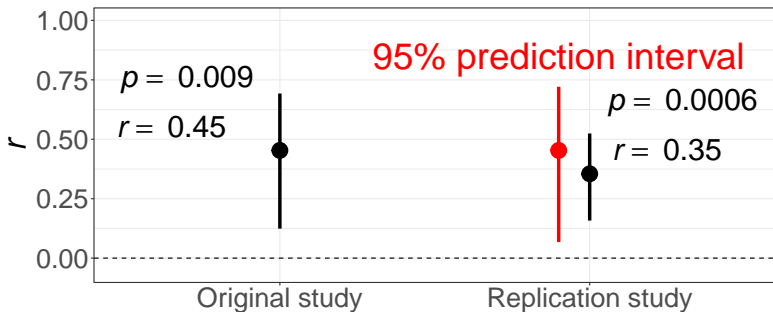
1. Statistical significance

Are original and replication estimates statistically significant?



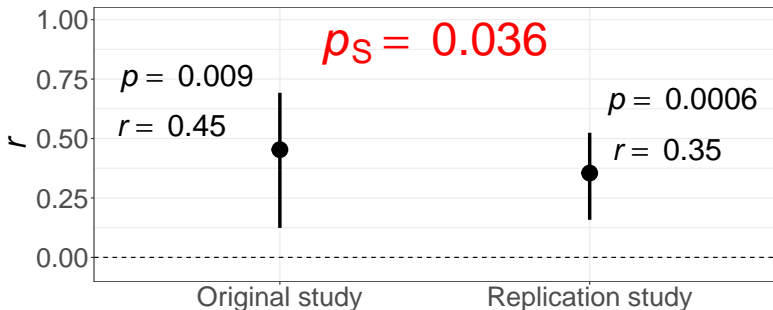
2. Compatibility of effect estimates

Is the replication estimate contained in its prediction interval?



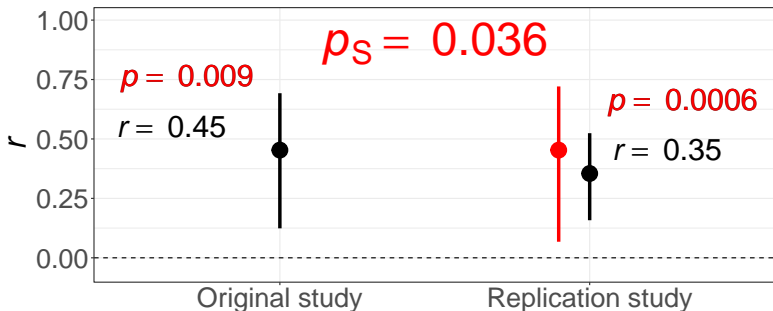
3. Sceptical p -value

?At which level can we convince a sceptic who argues that the original study is no longer significant at that level?



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 → 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 power calculation
- Depending on the projects, 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 (e.g. in Camerer et al. (2018))

Design of replication studies

Issues with this method

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

Package

To add: small intro to package (goal, structure etc)

Statistical framework of package

- Effect estimates are assumed to be normally distributed
 - usually fulfilled after suitable transformation
 - Fisher's z-transformation for correlation coefficients r
- Design prior
 - Conditional: ignores uncertainty of original study
 - Predictive: reflects that there is still uncertainty about the true effect after the original experiment

Statistical framework of package

- Relative quantities (as opposed to absolute quantities)
 - p -value or test statistic of original study
 - Relative sample size n_r/n_o
- Example for Morewedge et al. (2010):
 - $p_o = 0.009$
 - $p_r = 0.0006$
 - $c = 3$

Application

Application

1. Statistical significance
2. Comparison of effect estimates
3. Reverse Bayes methods

Statistical significance

Two functions:

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

Statistical significance

Two functions:

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

Main arguments

- `po` or `to`
- `c`
- `power`
- `designPrior`
- `shrinkage`

Statistical significance

Exercise 1

We have five 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.

- Please compute the conditional and predictive power of the five 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?

Statistical significance

Exercise 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.

- Please 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%?

Statistical significance

Exercise 3

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.

- Please 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%?

Comparison of effect size

– predictionInterval – sampleSizePI – sampleSizePIwidth

Reverse Bayes

– pSceptical – powerReplicationSuccess –
sampleSizeReplicationSuccess

Outlook

– Interim – Heterogeneity – EB shrinkage

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., Viciano, 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)*.
- 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.