Assessment of Replication Success

Sceptical Confidence Interval

Master Thesis in Biostatistics (STA495)

by

Xijin Chen
xijin.chen@uzh.ch

supervised by

Prof. Dr. Leonhard Held
leonhard.held@uzh.ch

Zurich, May 2020

Replication Success: Sceptical Confidence Interval

Xijin Chen

Version January 15, 2020

Contents

Preface

Howdy!

Max Muster
June 2018

Chapter 1

Introduction

Reproducible research is particularly important because it forms the foundation on which future studies are built and it indicates the establishement of credibility of a hypothesis. In the case of biomedical research, it is preclinical research that provides the exciting, new ideas that will eventually find their way into clinical studies and new drugs that provide benefit to humankind (**?**).

The past few year have witnessed impressive efforts on replication studies. Among these reproducibility projects, Reproducibility Project: Psychology (RP:P) is the largest and most well known one (**?**). It involved 270 crowd sourced researchers in 64 different institutions in 11 different countries. Researchers attempted direct replications of 100 studies published in three leading psychology journals in the year 2008. Each study was replicated only once. Replications attempted to follow original protocols as closely as possible. In almost all cases, replication studies used larger sample sizes than the original studies and therefore had greater statistical power, which means a greater probability of correctly rejecting the null hypothesis (i.e., that no relationship exists).

However, a larger number of reports illustrated that the results of large scale reproducibility projects could not be replicated, namely, the problem of 'replication crisis'. And this crisis is not unique to the field of psychology and emerged in many scientific studies. In 2016, a poll conducted by the journal Nature reported that more than half (52%) of scientists surveyed believed science was facing a "replication crisis" (**?**).

In the case of Open Science Collaboration (OSC), depending on the criterion used, only 36% to 47% of the original studies were successfully replicated, which led many to conclude that there is a "replication crisis" in psychological science (**?**).

As indicated by many authors, the RP:P dataset is always regarded as an invaluable resources since there are disagreements on qualification or assessment of the reproducibility (**??**). In this case, The simply descriptive statistics taken at face values in the RP:P project are referred to as underestimates of reproducibility.

There are a great number of reasons for the lack of reproducibility of scientific studies, like multiple testing, P-hacking, publication bias and under-powered studies (**?**). The leading one is over-Reliance on null hypothesis significance testing (**?**). Under the Neyman-Pearson (NP) hypothesis testing framework of hypothesis testing, P-value is connected with the notion of replication. Statistically significant findings and claiming of new discoveries solely based on the criterion of $p<0.05$ would result in a high rate of false positives. Therefore, replication probability, the probability of repeating a statistically significant result is substantially lower than expected.

The inferential inadequacies of statistical significance testing, which result in the problem of nob-reproducibility are now widely recognized. Therefore, in an attempt to assess the Reproducibility of researches, reliable approaches are required. Up untill now, there is no agreement on an unified statistical criterion for assessment of replication success. I summarized these approaches into several categories.

1. 'Significant result' of replication study based on P-value < 0.5.

The use of single P-value is essentially controvertial as indicated in many literatures (**??**).

2. Comparison of the original and replication effect size

To test the whether the original effect size was within the 95% confidence interval of replication study () and vice verse , checking the effect was within the confidence interval of original effect size or not(**??**).

3. Combination of original and replication studies by meta-analytic estimate of the effect size

Combine evidence from both studies to improve results over the consideration of only single studies (**?**).

4. Computation of a prediction interval of the effect estimate of the replication study

Judge whether the effect size of replication sutdy is within the 95% prediction interval (**?**).

8. prediction market: In the prediction market for a particular target study, peers who were likely to be familiar with experimental methods in economics could buy or sell shares whose monetary value depended on whether the target study was replicated (fig. S4 and tables S1 and S2). The prediction markets produce a collective market probability of replication (27) that can be interpreted as a replicability indicator (26).

Chapter 2

Data

Maybe it is the methods section. Here however, we give a couple hints. Note that you can wisely use *preamble*-chunks. Minimal, is likely:

```
library(knitr)
opts_chunk$set(
    fig.path='figure/ch02_fig',
    self.contained=FALSE,
    cache=TRUE
)
```

Defining figure options is very helpful:

```
library(knitr)
opts_chunk$set(fig.path='figure/ch02_fig',
               echo=TRUE, message=FALSE,
               fig.width=8, fig.height=2.5,
               out.width='\\textwidth-3cm',
               message=FALSE, fig.align='center',
               background="gray98", tidy=FALSE, #tidy.opts=list(width.cutoff=60),
               cache=TRUE
)
options(width=74)
```

Notice how in Figure 2.1 everything is properly scaled.

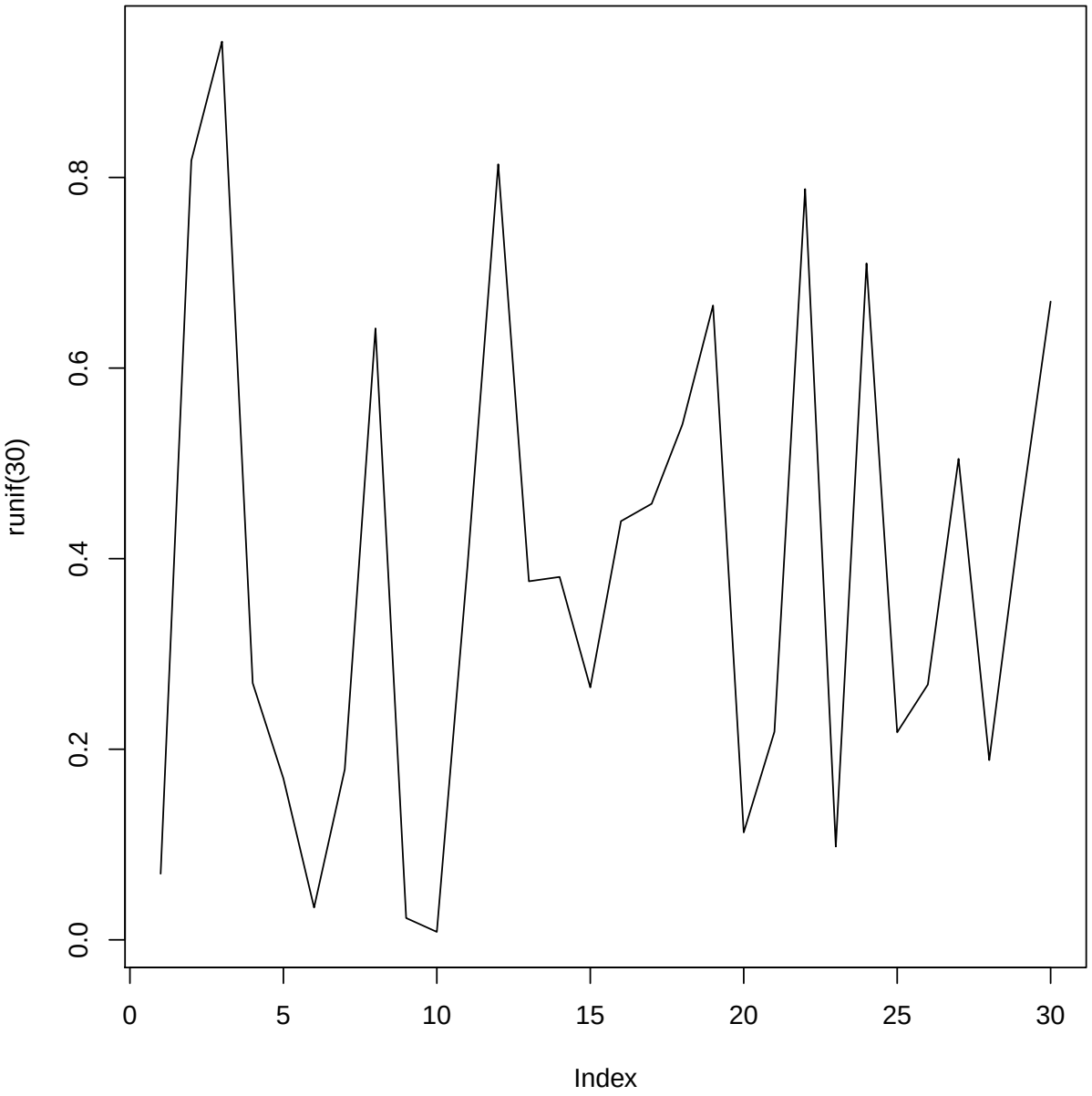

Figure 2.1: Test figure to illustrate figure options used by knitr.

Chapter 3

Results

Chapter 4

Discussion and Outlook

Chapter 5

Conclusions

Appendix A

Appendix

Maybe some R code here, probably a `sessionInfo()`

biblio