„Observed power“ und Umgang mit Zufallsfehler

Michael Höfler,
OSIP-Treffen am 2.12.2020

Einen Zusammenhang testen/schätzen

- Im einfachsten Fall soll unbekannte Mittelwertsdifferenz, **MD**, mittels Mittelwertsdifferenz aus Daten, **md**, untersucht werden
- Teststatistik **t** = **md** / **se**
- **se** = Standardfehler der Schätzung
  in Stichprobe = Zufallsfehler in **md**

Power bestimmen

Vor Durchführung einer Studie

- **Power** = Wahrscheinlichkeit, dass ein Niveau-α-Test einen wahren Zusammenhang von angenommener Größe **md*** bei angenommener Standardabweichung, **sd*,** erkennt
- Vorgegebenes β = 1 – Power, üblicherweise β = 0.1, 0.2
- Berechnung hängt tatsächlich nur von Verhältnis **md*/ sd*** (**Cohen‘s d**) ab
- Standardfehler = **se*** = **sd***/$\sqrt{n}$, n zu bestimmende Stichprobengröße

Aus den blauen Größen dann

- Hypothetische Teststatistik **t* = md* / se***
- n schließlich aus *t*-Verteilung: 1 - Wahrscheinlichkeit für Teststatistik > |t*|, falls wahre **MD** = 0

Beispiel, Berechnung in Stata

für Cohen‘s d = 0.8, α =. 05, β =.20, zweiseitiger Test

Mittelwert in Gruppe 1 bzw. 2

```
. power twomeans 0.8 0 , alpha(0.05) power(0.8)

Performing iteration ...

Estimated sample sizes for a two-sample means test
t test assuming sd1 = sd2 = sd
Ho: m2 = m1  versus  Ha: m2 != m1

Study parameters:

        alpha =    0.0500
        power =    0.8000
        delta =   -0.8000
           m1 =    0.8000
           m2 =    0.0000
           sd =    1.0000

Estimated sample sizes:

            N =        52
  N per group =        26
```

Gibt man kein SD* an, wird angenommen, dass SD* = 1 in beiden Gruppen

Äquivalent zu „Effektstärke“ = 0.8:
mean2 = 0.8
mean1 = 0
sd1* = sd2 = sd = 1

„Effektstärke“ = $(mean_2\text{-}mean_1) / SD^*$

Statistische Power hängt hier nur von Effektstärke ab! (gleiches Ergebnis z.B. für m2 = 1.6, m1 = 0, SD* = 2)

n/2 = 26 pro Gruppe benötigt, also n = 52

„Observed Power“

- Berechnung wie eben, aber mit aus den Daten geschätzten **md** und **sd**
- In SPSS automatisch mit dem Test ausgegeben
- Kann man also auch benutzen, nicht wahr?

Tests of Between-Subjects Effects

Dependent Variable: Bigband Music

Source	Type II Sum of Squares	df	Mean Square	F	Sig.	Noncent. Parameter	Observed Power[a]
Corrected Model	155.171[b]	7	22.167	20.661	.000	144.629	1.000
Intercept	16839.959	1	16839.959	15695.893	.000	15695.893	1.000
AGECAT4	140.810	3	46.937	43.748	.000	131.244	1.000
SEX	.761	1	.761	.709	.400	.709	.134
AGECAT4 * SEX	12.250	3	4.083	3.806	.010	11.418	.819
Error	1425.870	1329	1.073				
Total	18421.000	1337					
Corrected Total	1581.041	1336					

a. Computed using alpha = .05

b. R Squared = .098 (Adjusted R Squared = .093)

Wo liegt jetzt das Problem??

- Man kann die Daten nutzen, um Mittelwertsunterschied, **MD**, zu testen
- Genauso legitim: mittels der Daten die Power schätzen
- Problem aber: Nimmt man **dieselben Daten für beides**, liegt beiden Analysen derselbe Zufallsfehler zugrunde
- Denn man setzt in beide Formeln das Beobachtete (**md** und **sd**) ein

- Zufällig große **md** (**md/sd**) = zufällige Evidenz für große Power

Man kann sogar beides exakt ineinander umrechnen

The Abuse of Power: The Pervasive Fallacy of Power Calculations for Data Analysis

John M. HOENIG and Dennis M. HEISEY

The American Statistician, February 2001, Vol. 55, No. 1

Observed power can never fulfill the goals of its advocates because the observed significance level of a test ("*p* value") also determines the observed power; for any test the observed power is a 1:1 function of the *p* value. A

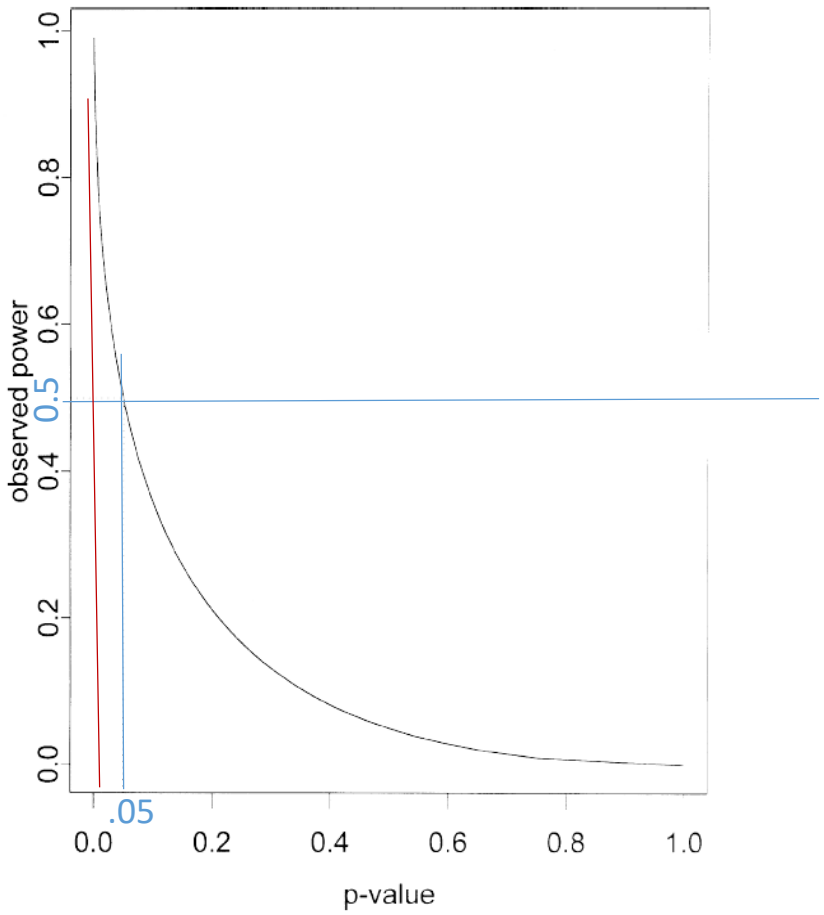


Figure 1. "Observed" Power as a Function of the p Value for a One-Tailed Z Test in Which α is Set to .05. When a test is marginally significant (P = .05) the estimated power is 50%.

Subtilere Falle: nichtsignifikant + „observed power“

entsteht, wenn man nichtsig. Studien mit unterschiedlichem p-Wert vermengt

Nonsignificance Plus High Power Does Not Imply Support for the Null Over the Alternative

SANDER GREENLAND, MA, MS, DrPH

Ann Epidemiol 2012;22:364–368.

type I error (false positive) rate. Among the problems with power computed from completed studies are these:

1. Irrelevance: Power refers only to future studies done on populations that look exactly like our sample with respect to the estimates from the sample used in the power calculation; for a study as completed (observed), it is analogous to giving odds on a horse race after seeing the outcome.
2. Arbitrariness: There is no convention governing the free parameters (parameters that must be specified by the analyst) in power calculations beyond the α-level.
3. Opacity: Power is more counterintuitive to interpret correctly than *P* values and confidence limits. In particular, high power plus "nonsignificance" does not imply that the data or evidence favors the null (6).

Even more startling is the "power approach paradox" detailed by Hoenig and Heisey (6): Among nonsignificant results, those with higher observed power are commonly interpreted as stronger evidence for the null, when in fact just the opposite is the case. Observed power is merely

p =.06, **kleine Stichprobe**, großer Mittelwertsunterschied:
high observed power, but **little** evidence for H_0

p =.50, **große Stichprobe**, kein Mittelwertsunterschied:
low observed power, but **high** evidence for H_0

APA-Experten-Empfehlung schon von 1998:

Statistical Methods in Psychology Journals

Guidelines and Explanations

Leland Wilkinson and the Task Force on Statistical Inference
APA Board of Scientific Affairs

in power calculations. Because power computations are most meaningful when done before data are collected and examined, it is important to show how effect-size estimates have been derived from previous research and theory in order to dispel suspicions that they might have been taken from data used in the study or, even worse, constructed to justify a particular sample size. Once the study is analyzed, confidence intervals replace calculated power in describ-
ing results

Konfidenzintervalle → unten

Kritikwelle an Tests, p-Werten … und observed power im Nachgang der Replikationskrise

Methodology (2011), 7, pp. 81-87. https://doi.org/10.1027/1614-2241/a000025. © 2011 Hogrefe Publishing.

Rethinking Observed Power
Concept, Practice, and Implications

Shuyan Sun ⓘ , Wei Pan ⓘ , Lihshing Leigh Wang ⓘ

Onlineveröffentlichung: Juni 27, 2011

https://doi.org/10.1027/1614-2241/a000025

Abstract | Volltext (HTML) | Literaturnachweis | Zitiert von | PDF

Abstract

Observed power analysis is recommended by many scholarly journal editors and reviewers, especially for studies with statistically nonsignificant test results. However, researchers may not fully realize that blind observance of this recommendation could lead to an unfruitful effort, despite the repeated warnings from methodologists. Through both a review of 14 published empirical studies and a Monte Carlo simulation study, the present study demonstrates that observed power is usually not as informative or helpful as we think because (a) observed power for a nonsignificant test is generally low and, therefore, does not provide additional information to the test; and (b) a low observed power does not always indicate that the test is underpowered. Implications and suggestions of statistical power analysis for quantitative researchers are discussed.

Safeguard Power as a Protection Against Imprecise Power Estimates

Marco Perugini, Marcello Gallucci, Giulio Costantini

First Published May 6, 2014 | Research Article | Find in PubMed | Check for updates
https://doi.org/10.1177/1745691614528519

Article information

Altmetric 6

A correction has been published:Erratum

Abstract

An essential first step in planning a confirmatory or a replication study is to determine the sample size necessary to draw statistically reliable inferences using power analysis. A key problem, however, is that what is available is the sample-size estimate of the effect size, and its use can lead to severely underpowered studies when the effect size is overestimated. As a potential remedy, we introduce *safeguard power analysis*, which uses the uncertainty in the estimate of the effect size to achieve a better likelihood of correctly identifying the population effect size. Using a lower-bound estimate of the effect size, in turn, allows researchers to calculate a sample size for a replication study that helps protect it from being underpowered. We show that in most common instances, compared with nominal power, safeguard power is higher whereas standard power is lower. We additionally recommend the use of safeguard power analysis to evaluate the strength of the evidence provided by the original study.

COMMENTARY

Misplaced Confidence in Observed Power

Zad Rafi[1]

[1]Department of Population Health, NYU Langone, New York, NY

Correspondence
Zad Rafi
Department of Population Health, NYU Langone, New York, NY

Contact
[1]Email: zad@lesslikely.com
[1]Twitter: @DailyZad

A recently published randomized controlled trial in *JAMA* investigated the impact of the selective serotonin reuptake inhibitor, escitalopram, on the risk of major adverse events (MACE). The authors estimated a hazard ratio (HR) of 0.69 (95% CI: 0.49, 0.96; $p = 0.03$) and then attempted to calculate how much statistical power their study (test) had attained, and used this measure to assess how reliable their results were. Here, we discuss why this approach, along with other post-hoc power analyses, are highly misleading.

KEYWORDS
Statistical Power · Sample Size · Confidence Intervals · Data Interpretation · Randomized Trials

JSR Journal of Surgical Research

LETTER TO THE EDITOR | ARTICLES IN PRESS

Post Hoc Power: Not Empowering, Just Misleading

Andrew D. Althouse, PhD

Published: August 16, 2020 • DOI: https://doi.org/10.1016/j.jss.2019.10.049

LETTERS TO THE EDITOR

Post Hoc Power Calculation: Observing the Expected

Plate, Joost D. J. MD; Borggreve, Alicia S. MD; van Hillegersberg, Richard MD, PhD; Peelen, Linda M. PhD
Author Information

Annals of Surgery: January 2019 - Volume 269 - Issue 1 - p e11
doi: 10.1097/SLA.0000000000002910

The 20% Statistician

A blog on statistics, methods, philosophy of science, and open science.
Understanding 20% of statistics will improve 80% of your inferences.

Friday, December 19, 2014

Observed power, and what to do if your editor asks for post-hoc power analyses

Observed power (or post-hoc power) is the statistical power of the test you have performed, based on the effect size estimate from your data. Statistical power is the probability of finding a statistical difference from 0 in your test (aka a 'significant effect'), if there is a true difference to be found. Observed power differs from the true power of your test, because the true power depends on the true effect size you are examining. However, the true effect size is typically unknown, and therefore it is tempting to treat post-hoc power as if it is similar to the true power of your study. In this blog, I will explain why you should never calculate the observed power (except for blogs about why you should not use observed power). Observed power is a useless statistical concept, and at the end of the post, I'll give a suggestion how to respond to editors who ask for post-hoc power analyses.

About Me

Blog by Daniel Lakens, experimental psychologist at the Human-Technology Interaction group at Eindhoven University of Technology, The Netherlands.

Search This Blog

Search

Subscribe To

PM&R WILEY

Statistically Speaking | Full Access

Post hoc Power

Regina L. Nuzzo PhD

First published: 25 August 2020 | https://doi.org/10.1002/pmrj.12476

Disclosure: none

SECTIONS PDF TOOLS SHARE

Introduction

Statistical power, or the "true positive" rate in significance testing, is a valuable and important concept in study design. "Post hoc power" or "observed power," however, is a different concept, and in practice it generally provides little useful information. In this article, I explain the difference and discuss why authors and reviewers should avoid reporting or requesting post hoc power based on study results.

Lösung

- **Unterschiedliche Daten** nehmen für beides
- **„external cross-validation“:** andere Daten für die Schätzung der Power verwenden (üblicherweise Literaturstudium)

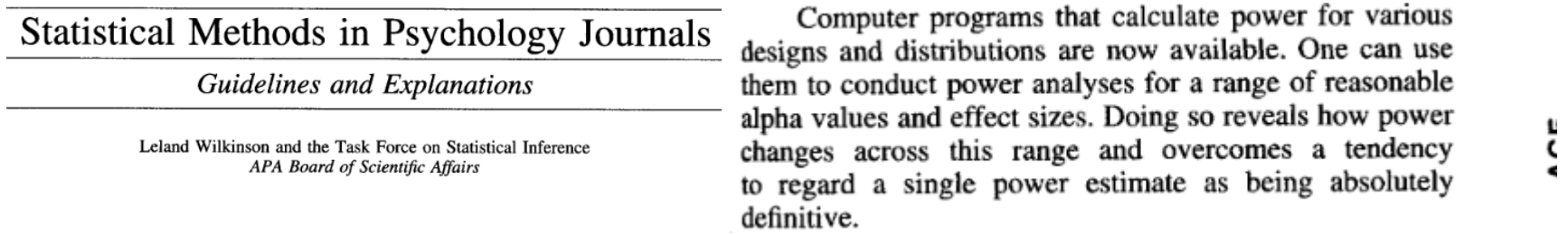
Daten können sich zufällig + systematisch unterscheiden

- **„internal cross-validation“:** Daten zufällig einteilen, im einen Teil auf Unterschiede testen, im anderen Power schätzen

Daten können sich nur zufällig unterscheiden

Grundsätzliche Kritik am Konzept Power

- Bei der Berechnung tut man so, als kenne man die wahre **MD** (MD/SD) mit 100%-iger Sicherheit (bayesianische Betrachtung)
- Dabei basiert der in die Formel eingesetzte Wert oft auf kleinen Stichproben (großer Zufallsfehler) und auf sehr selektiven Stichproben (systematischer Fehler, Bias)
- Besser wäre es, statt **MD** eine **Verteilung** zu nehmen, die die Unsicherheit über **MD** ausdrückt (ergäbe eine Wahrscheinlichkeitsverteilung für das benötigte **n**)



Statistical Methods in Psychology Journals

Guidelines and Explanations

Leland Wilkinson and the Task Force on Statistical Inference
APA Board of Scientific Affairs

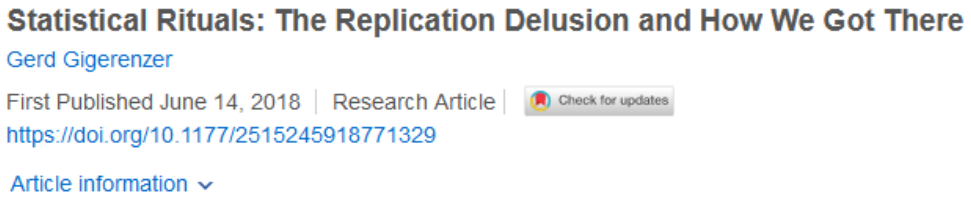
Computer programs that calculate power for various designs and distributions are now available. One can use them to conduct power analyses for a range of reasonable alpha values and effect sizes. Doing so reveals how power changes across this range and overcomes a tendency to regard a single power estimate as being absolutely definitive.

.. und seiner Umsetzung

- Warum α fast immer = .05 gewählt, aber β fast immer =.20 oder .10?
- Warum sollten **falsch positive Schlüsse** immer 2* bzw. 4* so schwer wiegen wie **falsch negative**? (Gegenbeispiel: Neue Behandlungsart bei Schwerstkranken mit Therapieresistenz)

- Eigene Beobachtung: Psychologinnen tun sich schon äußerst schwer darin, überhaupt einen Wert für **md** zu nennen
- Angst vor Fehlern? Aber an einer Annahme über **MD** führt bei der Stichprobenplanung kein Weg vorbei (keine Stichprobenplanung wäre unethisch)
- Vermeidung ist der noch größere Fehler und generell nicht gut (Klinische Psychologie)
- Manchmal kann man ihnen bestenfalls die Zustimmung zu etwas Halbgarem wie Cohen‘s d = 0,5 („mittlere Effektstärke“) abringen

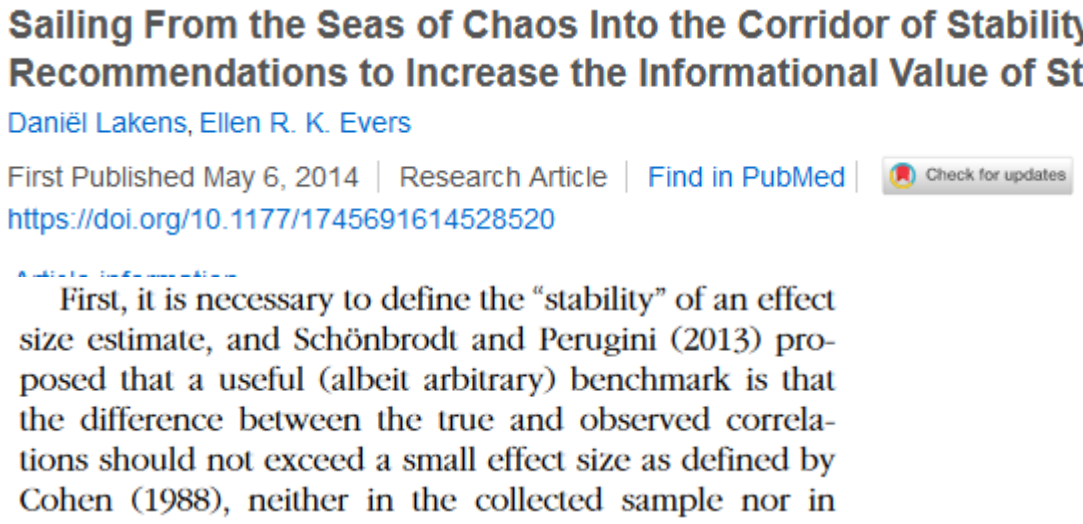
- These: hängt mit Ritual zusammen, Ergebnisse auf „signifikant“ und „nicht signifikant zu reduzieren“ (→ unten „dichotomia“)
- Wenig quantitatives Verständnis der verwendeten Skalen
- Mängel in der Lehre, z.B. bei der Interpretation von Regressionskoeffizienten (z.B. Koeffizient = **MD** in linearer Regression, wenn binäre Einflussvariable dummy-kodiert; = Korrelationskoeffizient, wenn X und Y z-standardisiert)
- Oft auch schlechte Skalen, psychometrische Mängel (Ergebnisse explorativ auf „gute“ Psychometrie“ getrimmt)



Umgang mit der Unsicherheit in Replikationsstudien

- **Einfache Antwort** nur, wenn man der ursprünglichen Studie gar nicht traut.
- Dann nur Ergebnis der Replikation trauen: per Design α und β und Bias klein (soll mittels besserer Methoden geringer sein als in Ausgangs, eher „conceptual replication“).
- Oder falls man hinter dem Ergebnis der ersten Studie p-hacking, harking, publication bias vernmutet (dann „direct replication“).

- Ansonsten die Ergebnisse der ersten Studie einbeziehen
- Die Frage nach dem Wie ist aber kompliziert und besser Gegenstand eines **eigenen Themas** (Format Journal Club?)



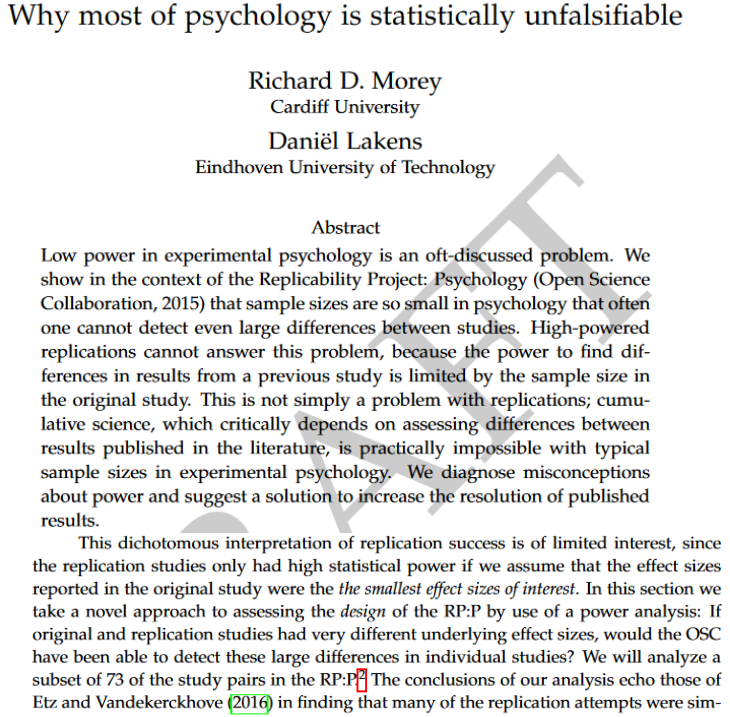
Sailing From the Seas of Chaos Into the Corridor of Stability: Practical Recommendations to Increase the Informational Value of Studies

Daniël Lakens, Ellen R. K. Evers

First Published May 6, 2014 | Research Article | Find in PubMed | Check for updates

https://doi.org/10.1177/1745691614528520

First, it is necessary to define the "stability" of an effect size estimate, and Schönbrodt and Perugini (2013) proposed that a useful (albeit arbitrary) benchmark is that the difference between the true and observed correlations should not exceed a small effect size as defined by Cohen (1988), neither in the collected sample nor in



Why most of psychology is statistically unfalsifiable

Richard D. Morey
Cardiff University

Daniël Lakens
Eindhoven University of Technology

Abstract

Low power in experimental psychology is an oft-discussed problem. We show in the context of the Replicability Project: Psychology (Open Science Collaboration, 2015) that sample sizes are so small in psychology that often one cannot detect even large differences between studies. High-powered replications cannot answer this problem, because the power to find differences in results from a previous study is limited by the sample size in the original study. This is not simply a problem with replications; cumulative science, which critically depends on assessing differences between results published in the literature, is practically impossible with typical sample sizes in experimental psychology. We diagnose misconceptions about power and suggest a solution to increase the resolution of published results.

This dichotomous interpretation of replication success is of limited interest, since the replication studies only had high statistical power if we assume that the effect sizes reported in the original study were the *the smallest effect sizes of interest*. In this section we take a novel approach to assessing the *design* of the RP:P by use of a power analysis: If original and replication studies had very different underlying effect sizes, would the OSC have been able to detect these large differences in individual studies? We will analyze a subset of 73 of the study pairs in the RP:P[2] The conclusions of our analysis echo those of Etz and Vandekerckhove (2016) in finding that many of the replication attempts were sim-

Original Article | Open Access

A new standard for the analysis and design of replication studies

Leonhard Held

First published: 26 December 2019 | https://doi.org/10.1111/rssa.12493 | Citations: 6

Summary

A new standard is proposed for the evidential assessment of replication studies. The approach combines a specific reverse Bayes technique with prior-predictive tail probabilities to define replication success. The method gives rise to a quantitative measure for replication success, called the sceptical *p*-value. The sceptical *p*-value integrates traditional significance of both the original and the replication study with a comparison of the respective effect sizes. It incorporates the uncertainty of both the original and the replication effect estimates and reduces to the ordinary *p*-value of the replication study if the uncertainty of the original effect estimate is ignored. The framework proposed can also be used to determine the power or the required replication sample size to achieve replication success. Numerical calculations highlight the difficulty of achieving replication success if the evidence from the original study is only suggestive. An application to data from the Open Science Collaboration project on the replicability of psychological science illustrates the methodology proposed.

Konfidenzintervalle sind besserer
Umgang mit Zufallsfehler als stat. Tests

cognitive bias of ***dichotomania***: the compulsion to replace quantities with dichotomies (“black-and-white thinking”), even when such dichotomization is unnecessary and misleading for inference.

Invited Commentary: The Need for Cognitive Science in Methodology FREE

Sander Greenland ✉

American Journal of Epidemiology, Volume 186, Issue 6, 15 September 2017, Pages 639–645, https://doi.org/10.1093/aje/kwx259

Mehrinformation:

Man kann p-Wert aus KI berechnen, aber nicht umgekehrt

```
. regress sf_mcs6 sexmale

      Source |       SS           df       MS      Number of obs   =    10,115
-------------+----------------------------------   F(1, 10113)     =     90.60
       Model |  90.579455         1   90.579455   Prob > F        =    0.0000
    Residual |  10111.204    10,113  .999822411   R-squared       =    0.0089
-------------+----------------------------------   Adj R-squared   =    0.0088
       Total |  10201.7835    10,114   1.0086794   Root MSE        =    .99991

------------------------------------------------------------------------------
     sf_mcs6 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     sexmale |   .1895941   .0199192     9.52   0.000     .1505485    .2286396
       _cons |  -.1916603   .0136616   -14.03   0.000    -.2184397    -.164881
------------------------------------------------------------------------------

. disp (.2286396-.1505485 )/(2*invnorm(0.975))
.01992157

. disp .1895941/.01992157
9.517026

. disp 1-normal(9.517026)
0
```

Aber verschiedene KI entsprechen p=.000,

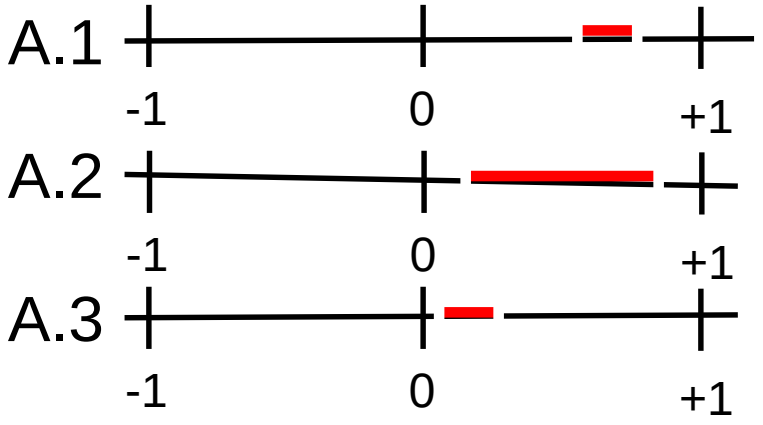
z.B. hier für Cohen‘s d:
0.15 – 0.22, 1.07 – 3.50

Das Problem der Dichotomisierung

Signifikant

Situation A: $p < 0.05$

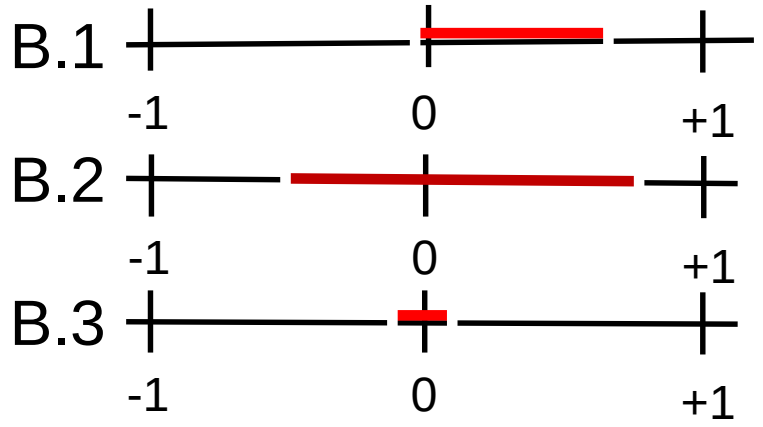
KIs A1 bis A3 gleich interpretieren, ernsthaft?



Nicht signifikant

Situation B: $p \geq 0.05$

KIs B1 bis B3 gleich interpretieren, ernsthaft?



Folie von John Venz

Ganze Bandweite eines KI interpretieren Beispiel: Gruppenunterschied in IQ

Intervall		Interpretation
0.1	12.1	Praktisch gleicher bis um 12.1 höherer IQ in Gruppe B möglich
0.1	40.0	Höherer IQ in Gruppe B, aber praktisch fast jedes Ausmaß hierfür möglich
10.0	20.0	In Gruppe B um mindestens 10, maximal um 20 größer
-20.0	20.0	In beiden Gruppen könnte IQ viel höher sein, Stichprobe nicht informativ!
-0.1	0.1	Mittlerer IQ in beiden Gruppen praktisch gleich groß, Nullhypothese näherungsweise gezeigt

Aber in Medizin/Epidemiologie hat die Pflicht, in Papers KIs statt p-Werte zu berichten, nichts gebracht. Wissenschaftlerinnen interpretieren nur, ob die 0 drin liegt oder nicht.

Research Article

Editors Can Lead Researchers to Confidence Intervals, but Can't Make Them Think

Statistical Reform Lessons From Medicine

Fiona Fidler,[1,2] Neil Thomason,[2] Geoff Cumming,[1] Sue Finch,[1] and Joanna Leeman[1]

[1]La Trobe University, Melbourne, Australia, and [2]The University of Melbourne, Melbourne, Australia

STATISTICAL REFORM IN PSYCHOLOGY: CIS ARE IMPORTANT BUT RARELY USED

For decades, many advocates of statistical reform in psychology have recommended CIs as an alternative (or at least a supplement) to *p* values. The APA (2001) *Publication Manual* now calls them "the best reporting strategy" (p. 22). Although it is too early to assess the impact of the new APA recommendation, previous decades of encouraging researchers to report CIs have had little effect. They remain relatively little used in psychology (Finch, Cumming, & Thomason, 2001; Kieffer, Reese, & Thompson, 2001).

Researchers in psychology rarely see CIs reported, so some may not understand why CIs are important; researchers may also have the misconception that CIs are equivalent to NHST. CIs can indeed be used to perform NHST, by noting whether the null value is within the interval. Unlike *p* values, however, they also provide information on precision: CI width is a guide to this. Effect sizes are important because they are the primary outcome of research and are needed for meta-analysis, and CIs *are* estimates of effect size. Unlike NHST, CIs "provide information on both location and precision" (APA, 2001, p. 22).

Studie planen mittels Breite eines Konfidenzintervalls

Planning Study Size Based on Precision Rather Than Power

Kenneth J. Rothman[a,b] *and Sander Greenland*[c]

Epidemiology • Volume 29, Number 5, September 2018

Power calculations suffer from several drawbacks. The most glaring is the connection with statistical significance testing. A focus on statistical power promulgates the "dichotomania"[1] that is characteristic of significance testing, classifying the results of a quantitative exercise into two ultimate categories, statistically

Sailing From the Seas of Chaos Into the Corridor of Stability: Practical Recommendations to Increase the Informational Value of Studies

Daniël Lakens, Ellen R. K. Evers

First Published May 6, 2014 | Research Article | Find in PubMed | Check for updates
https://doi.org/10.1177/1745691614528520

The corridor of stability

With a small number of observations, effect size estimates have very wide CIs and are relatively unstable. An effect size estimate observed after collecting 20 observations can change dramatically if an additional 20 observations are added. An important question when designing an experiment is how many observations are needed to observe relatively stable effect size estimates, such that the effect size estimate will not change considerably when more participants are collected. On the basis of approaches in statistics that stress accuracy, and not just statistical significance (e.g., Kelley & Maxwell, 2003), Schönbrodt and Perugini (2013) have recently performed simulations that address this question.

Beispiel:

(zweiseitiges) KI für Cohen‘s d soll maximale Breite von 0.5 (= SD/2) haben

Durch Ausprobieren*: 124 pro Gruppe!

```
. ttesti 124 0.5  1 124 0 1

Two-sample t test with equal variances
```

	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf.	Interval]
x	124	.5	.0898027	1	.3222412	.6777588
y	124	0	.0898027	1	-.1777588	.1777588
combined	248	.25	.0653374	1.028934	.1213106	.3786894
diff		.5	.1270001		.2498537	.7501463

```
    diff = mean(x) - mean(y)                                      t =   3.9370
Ho: diff = 0                                     degrees of freedom =      246

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.9999         Pr(|T| > |t|) = 0.0001          Pr(T > t) = 0.0001
```

* Variieren der Gruppengröße, bis KI gewünschte Länge (Annahme: gleiche SD in beiden Gruppen → gleichgroße Gruppen)

Generelles Problem: Als Menschen tun wir uns schwer damit, Unsicherheit einzuräumen

When 90% confidence intervals are 50% certain: On the credibility of credible intervals

May 2005 · Applied Cognitive Psychology 19(4):455 - 475 · Follow journal

DOI: 10.1002/acp.1085

Karl Halvor Teigen · Magne Jørgensen

Overview Stats Comments Citations (84) References (48) Related resea

Abstract and figures

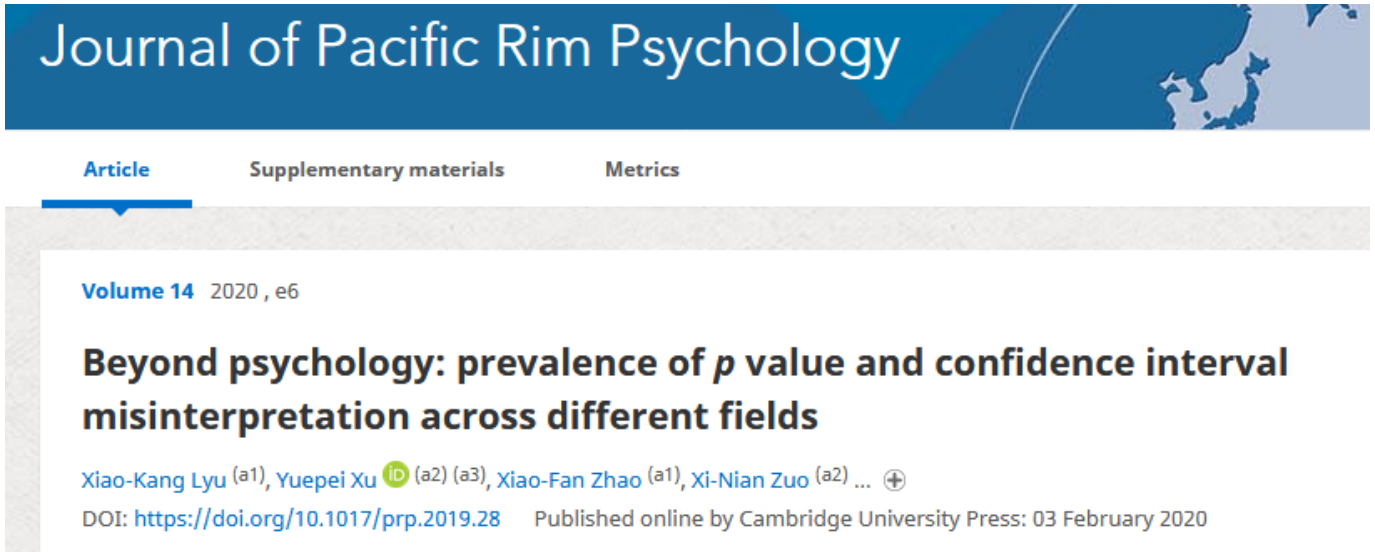
Estimated confidence intervals for general knowledge items are usually too narrow. We report five experiments showing that people have much less confidence in these intervals than dictated by the assigned level of confidence. For instance, 90% intervals can be associated with an estimated confidence of 50% or less (and still lower hit rates). Moreover, interval width appears to remain stable over a wide range of instructions (high and low numeric and verbal confidence levels). This leads to a high degree of overconfidence for 90% intervals, but less for 50% intervals or for free choice intervals (without an assigned degree of confidence). To increase interval width one may have to ask exclusion rather than inclusion questions, for instance by soliciting 'improbable' upper and lower values (Experiment 4), or by asking separate 'more than' and 'less than' questions (Experiment 5). We conclude that interval width and degree of confidence have different determinants, and cannot be regarded as equivalent ways of expressing uncertainty. Copyright © 2005 John Wiley & Sons, Ltd.

Personality and Individual Differences

Volume 98, August 2016, Pages 345-354

Better the devil you know than a world you don't? Intolerance of uncertainty and worldview explanations for belief in conspiracy theories

Richard Moulding [a], Simon Nix-Carnell [a], Alexandra Schnabel [a], Maja Nedeljkovic [b], Emma E. Burnside [a], Aaron F. Lentini [a], Nazia Mehzabin [a]



Statistical Tests, *P*-values, Confidence Intervals, and Power: A Guide to Misinterpretations

Sander GREENLAND, Stephen J. SENN, Kenneth J. ROTHMAN, John B. CARLIN, Charles POOLE, Steven N. GOODMAN, and Douglas G. ALTMAN

Was nun? Was tun?

Lehre

- Mehr quantitatives Verständnis für Effektgrößen schon im Bachelor?
- Wie kann man einmal Gelerntes bei Masterstudentinnen und Wissenschaftlerinnen ersetzen?

Wissenschaftliche Fragen

- Wie lässt sich die mechanische Verwendung von Statistik überwinden? Wie begegnet man der wahrscheinlich dahinter steckenden Hilflosigkeit?
- Ist solche Hilflosigkeit auf Statistik beschränkt oder betrifft sie ganz allgemein das Verständnis von Wissenschaft?

Studienideen:

- Onlineexperimente: Durch welches Lehren richtige Interpretation von Konfidenzintervallen?
- …