awm

angewandte **w**irtschafts- und **m**edienpsychologie

Fallzahlplanung

Thema 09

HOCHSCHULE ANSBACH

Warum große Stichproben gut (schlecht) sind

Meehls Paradox

Paul Meehl, 1920-2003

- Jede Hypothese der Art „der Effekt ist ungleich Null“ kann zu Meehls Paradox führen.
- Je größer die Stichprobe, desto einfacher wird es, die Hypothese zu bestätigen und desto schwieriger, sie zu verwerfen.
- Aber Wissenschaft sollte genau umgekehrt laufen: Große Stichproben erlauben genauere Messungen und sollten daher kleinere Fehler aufdecken bzw. Fehler mit höherer Sicherheit aufdecken. Daher sollten größere Stichproben es einer Hypothese schwieriger machen.
- Beispiel: Die Lichtgeschwindigkeit beträgt (laut der aktuellen Hypothese) 299,792.458 m/s. Genauere Messungen (durch größere Stichproben) sollten es der Hypothese mehr zusetzen als kleinere Stichproben.
- Das Testen von Nullhypothesen (z.B. in den Sozialwissenschaften, kaum in der Physik) lädt das Meehlsche Paradox ein.

Bildquelle

Lösungen zu Meehls Paradox

ROPE

- Eine Theorie sagt einen Wert hervor sowie einen Bereich „praktisch äquivalenter“ Werte darum (ROPE: Region of practical Equivalence).
- Je mehr Daten gesammelt werden, desto schmaler wird ROPE, so wie es sich für die Wissenschaft gehört.

Parameterschätzung

- NullHypothesen zu testen kann zu „Schwarz-Weiß-Denken“ führen: Ja-Nein trifft die Wirklichkeit aber nicht so gut wie Grautöne.
- Außerdem ist die Nullhypothese in vielen Forschungsfeldern meist falsch oder irrelevant.
- Daher bietet sich an, die relevanten theoretischen Größen mit einem Bereich plausibler Werte zu schätzen.
- „Wir schätzen, dass die Lichtgeschwindigkeit etwa zwischen X und Y liegt.“
- Ist ein Wert außerhalb des Schätzbereichs, so ist die zugehörige Hypothese automatisch verworfen. Insofern sind Parameterschätzung Obermengen von Hypothesentests.

ROPE illustriert

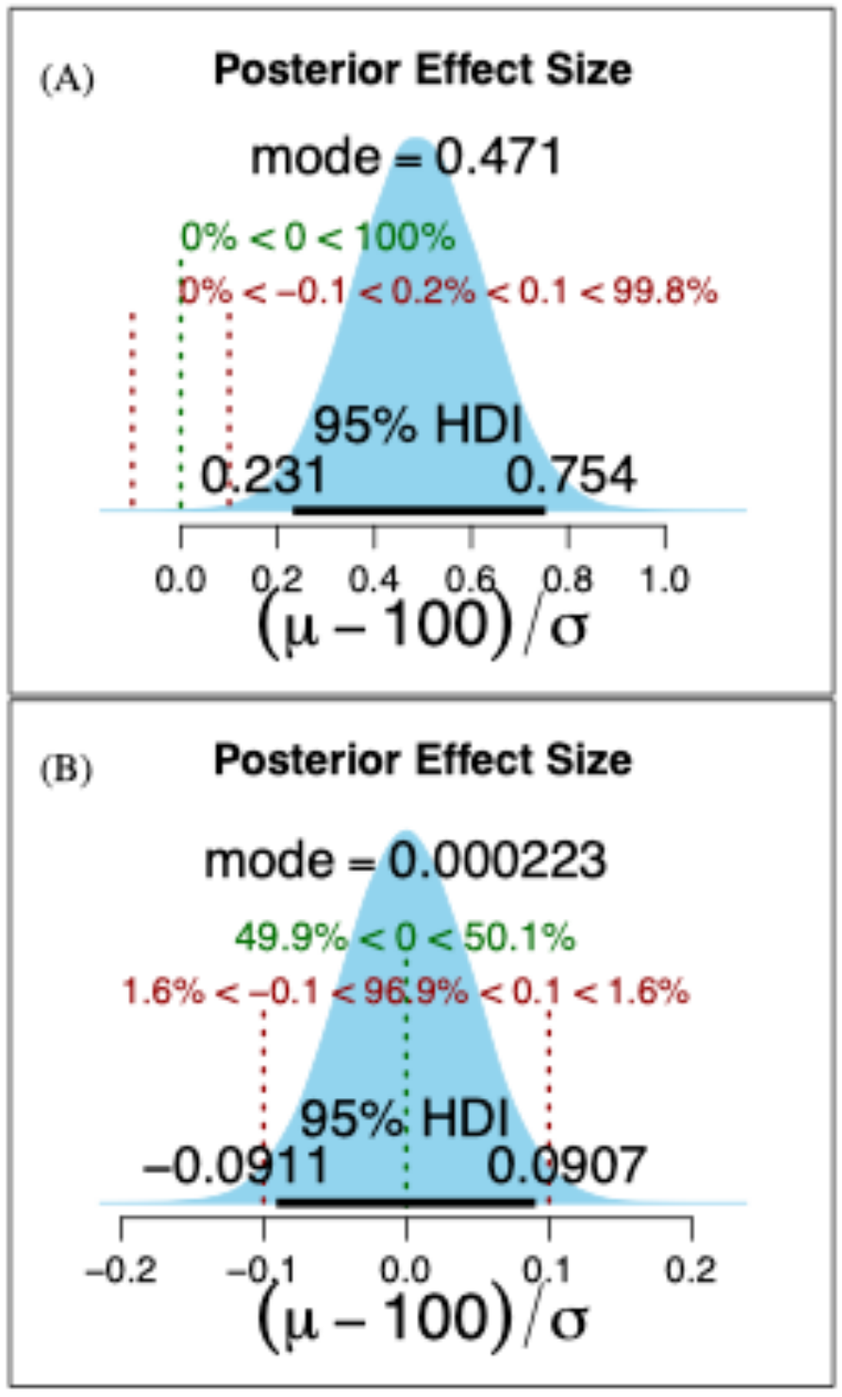


Figure 5. Posterior distributions on effect size for IQ, $(\mu - 100)/\sigma$, marked with 95% HDI and ROPE. The null value of 0 is marked by a vertical dotted line, annotated with the percentage of the posterior distribution that falls below it and above it. The ROPE limits are also marked with vertical dotted lines and the percentage of the posterior distribution that falls below, within, and above the ROPE. (A) Posterior distribution of effect size when $N = 63$ with sample mean of 110 and sample standard deviation of 20. This distribution is just a different perspective on the same posterior distribution shown in Figure 3. Notice that the 95% HDI falls entirely outside the ROPE, and there is only 0.2% probability that the effect size is practically equivalent to zero. (B) Posterior distribution of effect size when $N = 463$ with sample mean of 100 and sample standard deviation of 15. Notice that the 95% HDI falls entirely within the ROPE, and there is 96.9% probability that the effect size is practically equivalent to zero.

Kruschke, J. K., & Liddell, T. (2017). *Bayesian data analysis for newcomers*. PsyArXiv. https://doi.org/10.31234/osf.io/nqfr5

Parameterschätzung illustriert (Bayes)

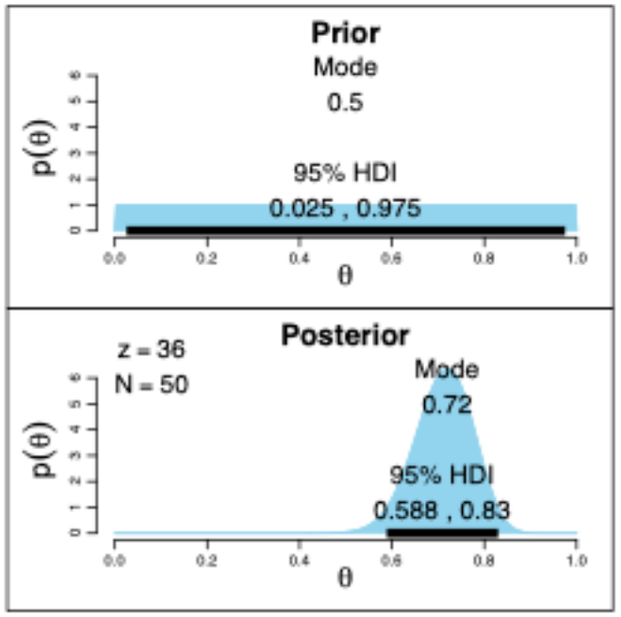


Figure 2. Estimating the probability θ that a patient is cured when given a particular drug. (A) Prior distribution is broad over possible values of the parameter θ. (B) For $N = 50$ patients with $z = 36$ cures, the posterior distribution over the parameter θ is much narrower. HDI limits and modes are displayed to first three digits only. *Compare with Figure 1.*

Kruschke, J. K., & Liddell, T. (2017). *Bayesian data analysis for newcomers*. PsyArXiv. https://doi.org/10.31234/osf.io/nqfr5

Size matters: Großes N erhöht die Präzision

- Sowohl für ROPE als auch für Parameterschätzung gilt: Je größer N, desto besser.
- ROPE: Ein größeres N ficht eine Hypothese stärker an als ein kleineres N
- Paramterschätzung: Je größer N, desto präziser wird ein Parameter geschätzt.
- Die Wissenschaft funktioniert also, wie sie sollte mit beiden Ansätzen.
- Dagegen schlägt leider Meehls Paradox zu beim „normalen“ Testen von Nullhypothesen.
- Vom Testen von Nullhypothesen sollte man daher Abstand nehmen, so einige Statistiker (z.B. Gelman).

Fallzahlplanung

- Plant man vorab (vor der Datenerhebung), wie groß die Stichprobe (das „N“) sein soll, spricht man von Fallzahlplanung.
- Die „Kraft“ (Wahrscheinlichkeit) einer Untersuchung einen Effekt zu finden, bezeichnet man als Power. Daher spricht man auch von Poweranalyse (synonym zu Fallzahlplanung).
- Man kann vorab ausrechnen, wie groß N sein muss, um eine gewissen Power zu erreichen. Üblich sind werte von 80%.
- Anstelle von der Vorab-Berechnung der Power kann man auch ausrechnen, wie groß N sein muss, um einen Effekt mit einer gewissen Präzision zu schätzen. Genauer gesagt legt man fest, wie breit ein Schätzbereich (z.B. 95%-HDI) sein soll und prüft, welches N man wohl braucht, um diese Präzision zu erreichen.
- Gründe für Fallzahlplanung
  - Projektplanung (Planung von Zeit- und Geld-Ressourcen)
  - Man wird gezwungen (Gutachter, Geldgeber, …)
  - Einige statistische Ansätze (Frequentismus) benötigt eine Fallzahlmessung
  - Fallzahlplanung verhindert Betrug, wenn vorab dokumentiert

Fallzahlplanung gehört zur Inferenzstatistik

Stichprobe

- Fallzahlplanung ist ein Ansatz, um von einer Stichprobe auf eine Population zu schließen
- Genauer gesagt um zu prüfen, wie sicher/präzise dieser Schluss gelingt

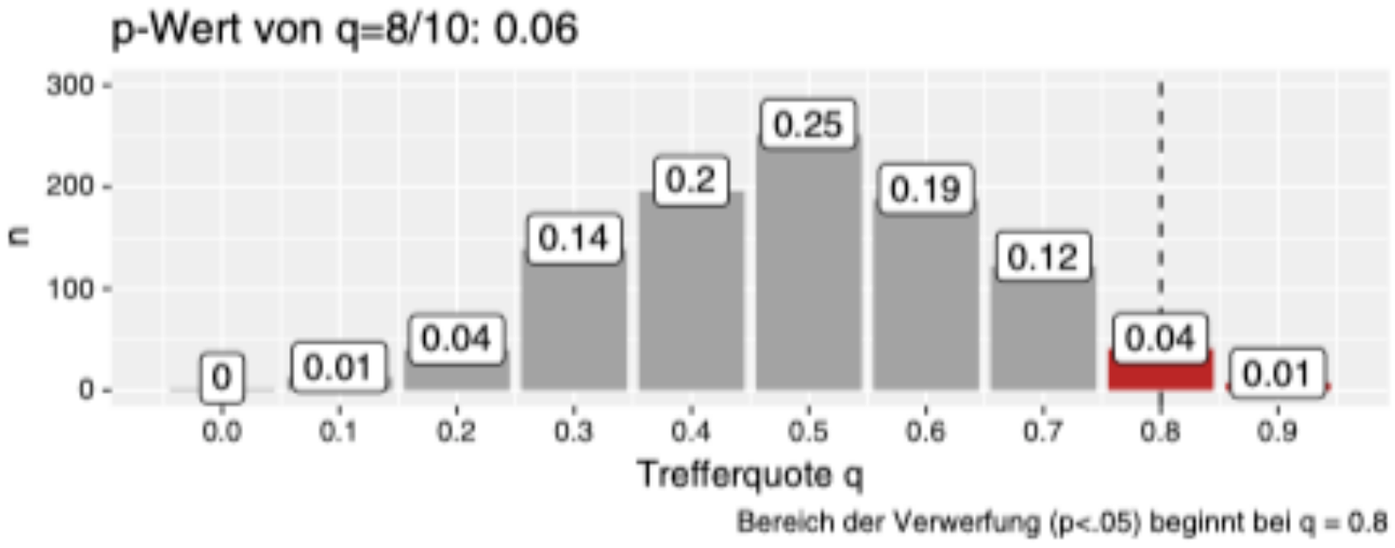
Vollerhebung
(komplette Population)

Sauer, S. (2019). *Moderne Datenanalyse mit R: Daten einlesen, aufbereiten, visualisieren und modellieren* (1. Auflage 2019). Springer. https://www.springer.com/de/book/9783658215866

Fixes N, um einen p-Wert zu berechnen

Berechnung des p-Wert braucht fixes N

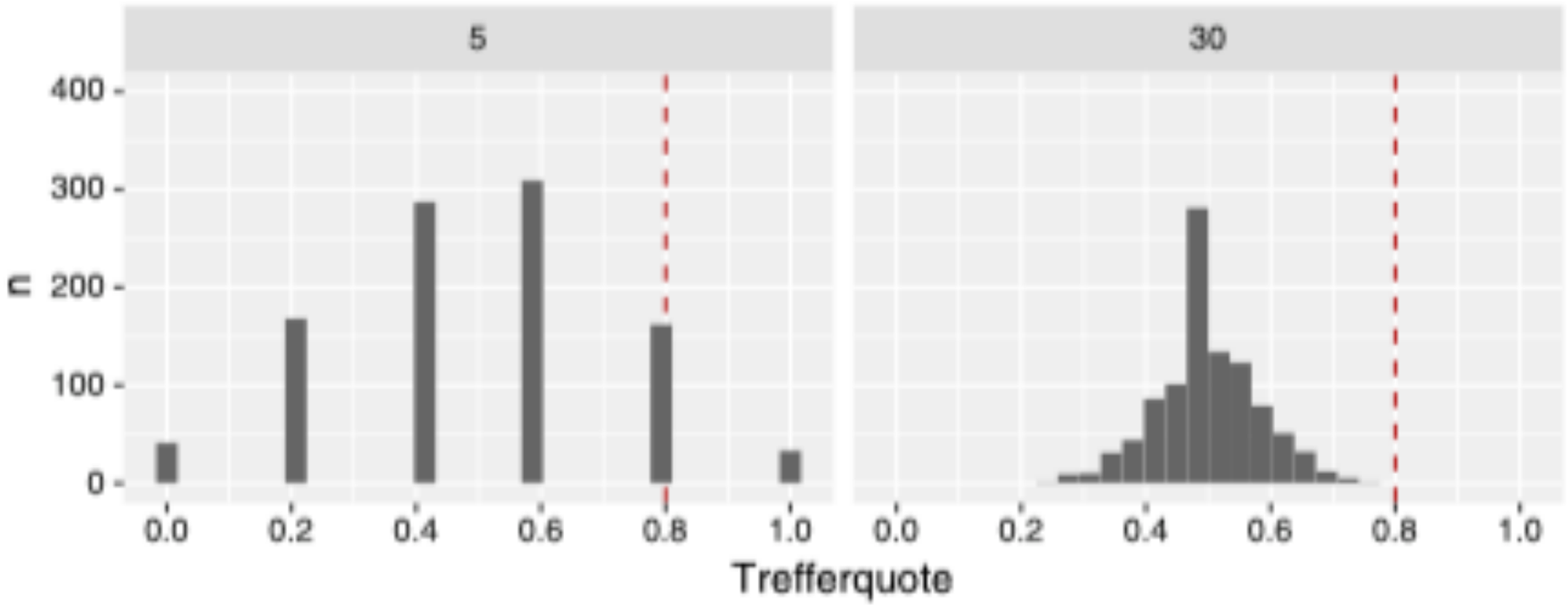
- Der p-Wert ist die zentrale Größe des Frequentismus.
- Um einen p-Wert zu berechnen, braucht es ein vorab definiertes („fixes“) N.
- Alternativ könnte man z.B. vorab die Erhebungsdauer bestimmen, dann wäre N eine Zufallsvariable (hat also eine Verteilung). Ginge auch, aber komplizierter. In der Praxis geht man (bis heute) von einem fixen N aus.
- Man vergleicht dann die Verteilung der Stichproben unter der angenommenen Hypothese mit dem empirischen Kennwert.
- Je extremer der Kennwert in der Verteilung, desto kleiner der p-Wert.
- Bei sehr kleinen p-Werten verwirft man die fragliche Hypothese und verkündet „Signifikanz“.



Sauer, S. (2019). *Moderne Datenanalyse mit R: Daten einlesen, aufbereiten, visualisieren und modellieren* (1. Auflage 2019). Springer. https://www.springer.com/de/book/9783658215866

Je größer N, desto eher kann man die Hypothese verwerfen

- Leider schlägt jetzt Meehl(s Paradox) zu:
- Wie man sieht, wird die Streuung in der Stichprobenverteilung (der sog. Standardfehler) kleiner, wenn N größer wird.
- Der Anteil der simulierten Stichproben, die mindestens so extrem sind wie das echte (empirische) Stichprobenergebnis, summiert sich zum p-Wert.



Sauer, S. (2019). *Moderne Datenanalyse mit R: Daten einlesen, aufbereiten, visualisieren und modellieren* (1. Auflage 2019). Springer. https://www.springer.com/de/book/9783658215866

Optional Stopping (Sequenzielles Testen)

Messen, Testen, Repeat until signifikant (?)

- Intuitiv hört sich Optionale Stopping (Sequenzielles Testen) stimmig an:
  - Sammle ein paar Daten
  - Dann prüfe, ob sich schon ein Effekt findet, wenn ja, stoppe (aus ökonomischen Gründen)
  - ansonsten wiederhole, bis signifikant oder Geld alle
- Ist auch plausibel, aber:
- Wenn das Ziel die Kontrolle von Fehlalarmen ist, dann ist von Optional Stopping abzuraten.
- Das Ziel des Frequentismus ist die Kontrolle von Fehlalarmen: Man möchte die Anzahl von Fehlalarmen minimieren oder zumindest wissen, wie groß die Gefahr (Wahrscheinlichkeit) von Fehlalarmen ist.
- Unter Fehlalarm ist die Wahrscheinlichkeit verstanden, eine (Null-)Hypothese fälschlich zu verwerfen, also einen Effekt zu verkünden, obwohl es keinen gibt.
- Allerdings sind Nullhypothesen meistens falsch, so dass man gar keinen Fehlalarm begehen kann, so einige Statistiker (ist auch meine Meinung).

Beispiel: Prof. Süß forscht

Mich würde mal interessieren, wieviel Lernen für eine Klausur eigentlich nutzt.

- Prof. Süß übt sich im Testen von Nullhypothesen.
- Er will folgende Nullhypothese verwerfen: „Lernen bringt nix“ (vor allem in seinen Fächern).
- Er hofft auf Meehls Paradox: „Sichere Sache, fleißig erheben, irgendwann ist es sicher, dass ich die feindliche Nullhypothese los bin, harhar!“
- Er untersucht
  - 10 Studis und testet auf Signifikanz. Nix.
  - Die nächsten 10. Immer noch nix.
  - Wieder 10. Nicht signifikant.
  - 10 neue Studis. Jetzt signifikant!
- Wie groß ist wohl die Gefahr eines Fehlalarms (wenn man an Nullhypothesen glaubt, wovon wir hier jetzt mal ausgehen, da es viele tun)?

Alphafehler-Inflation im Frequentismus

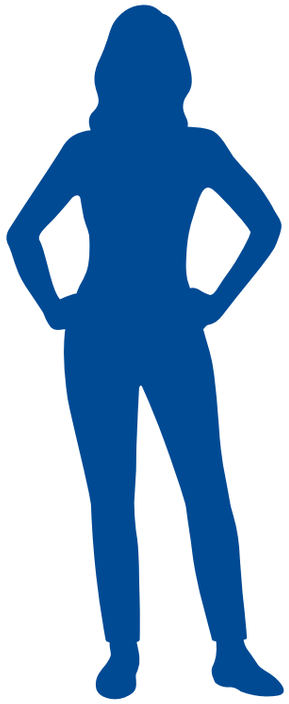
$$R = r \cdot r = .95 \cdot .95 \approx .90$$

$$R = r^k \rightarrow 1 - R = 1 - r^k$$

- Ein statistischer Test wird zumeist auf Fehlalarm-Sicherheit von r = 95% eingestellt: In 5% der Fälle wird der Test fälschlich einen Effekt finden, obwohl die Nullhypothese wahr ist (kein Effekt in Wahrheit).
- Testet man zwei Mal, so ist die Wahrscheinlichkeit, insgesamt richtig (R) zu sein, gleich dem Quadrat von r.
- *Insgesamt Richtig* soll heißen, dass man in keinem der Tests einen Fehlalarm produziert.
- Allgemein ist R bei k Tests gleich r hoch k.
- Die Gefahr, nicht insgesamt richtig zu sein, ist dann 1 minus R.
- Das Aufaddieren der Fehlalarm-Wahrscheinlichkeit bezeichnet man als Alphafehler-Inflation.

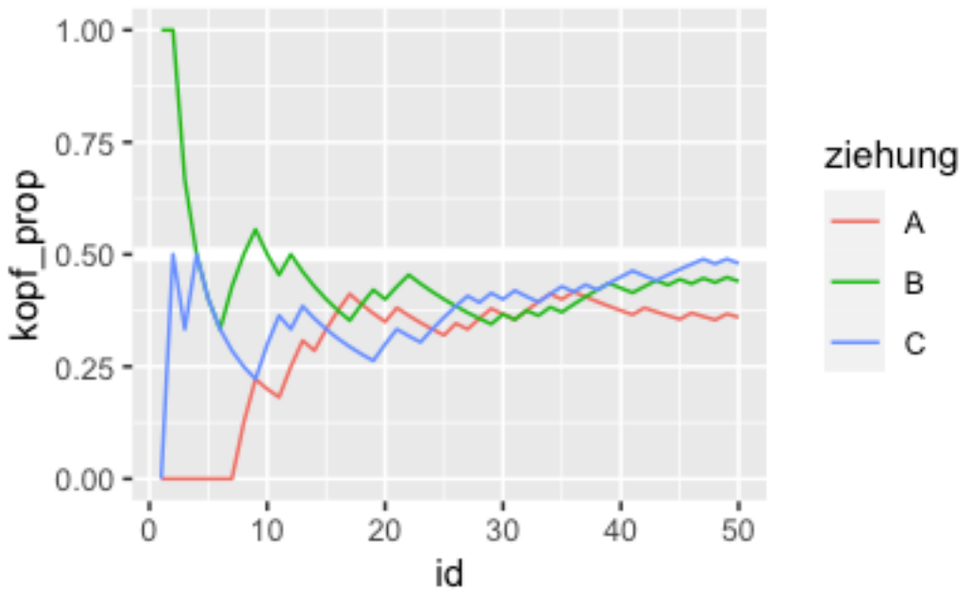
Climber's doom

Das ist doch nur Wahrscheinlichkeit! Wer klettert mit mir mit?!



- Eine Klettererin verwendet ein Seil, dass eine Sicherheit von 99% hat: mit einer Wahrscheinlichkeit von 1% reißt das Seil.
- Jetzt knüpft sie 10 dieser Seile zusammen.
- Wie groß ist die Gefahr, dass das „Gesamtseil" reist?

Optional Stopping ist immer problematisch



50-facher Wurf einer fairen Münze (3 Mal wiederholt). Wie man sieht, ist der kumulierte Anteil von *Kopf* nicht immer 50%, aber nähert sich tendenziell der 50%-Marke an mit der Zeit.

- Für Bayes ist optional stopping weniger problematisch als für den Frequentismus.
- Aber es ist immer (oder kann immer) ein Problem sein, auch für Bayes (wenn auch evtl. in geringerem Maße als für den Frequentismus).
- Bei kleinen Stichproben sind extreme Ergebnisse wahrscheinlicher als bei großen: Große Stichproben spiegeln den wahren Wert gut wieder (normalerweise), kleine Stichproben sind durch viel Variabilität gekennzeichnet.
- Beispiel: Wirft man 1000 Mal eine faire Münze, so wird sich ziemlich exakt 50% Trefferquote für *Kopf* zeigen. Wirft man aber nur 10 Mal eine faire Münze, sind extreme Ereignisse wie 80% Trefferquote weit häufiger (wahrscheinlicher).
- Insofern kann man von Falsch-Positiv-Ergebnissen sprechen, wenn man den Versuch vorzeitig beenden würde, wenn die Trefferquote gerade über 50% liegt.

Methoden der Fallzahlplanung

Ansätze der Fallzahlplanung

- **Methode 1: Mehrebenenmodelle** (Multi-Level-Modellierung)
  - Mit Multi-Level-Modellen können Falsch-Positiv-Ergebnisse vermieden werden.
  - Betrügen kann damit nicht kontrolliert werden.
  - Statistisch ein schöner Weg, in Bayes recht einfach darzustellen.
  - Hier nicht weiter betrachtet, der Einfachheit halber.
- **Methode 2: Simulation**
  - Auf Basis unseres Modells (Prior und Likelihood) simuliert man sich viele Stichproben
  - Dann schaut man, bei welchem N die Mehrzahl dieser Stichproben einen Effekt mit gewünschter Präzision/Power finden
- **Methode 3: Ausrechnen**
  - Je nach Modell kann man die benötigte Stichprobengröße (N) ausrechnen
  - Gute Nachricht: Das haben schon Menschen vor uns gemacht und es uns bereitgestellt.
  - Schlechte Nachricht: Leider zumeist nur für frequentistische Tests, nicht für für Bayes-Verfahren.
  - Im Folgenden werden wir diese Werte als grobe Schätzwerte für Bayes-Verfahren hernehmen (und hoffen das Beste).

Simulation von Daten

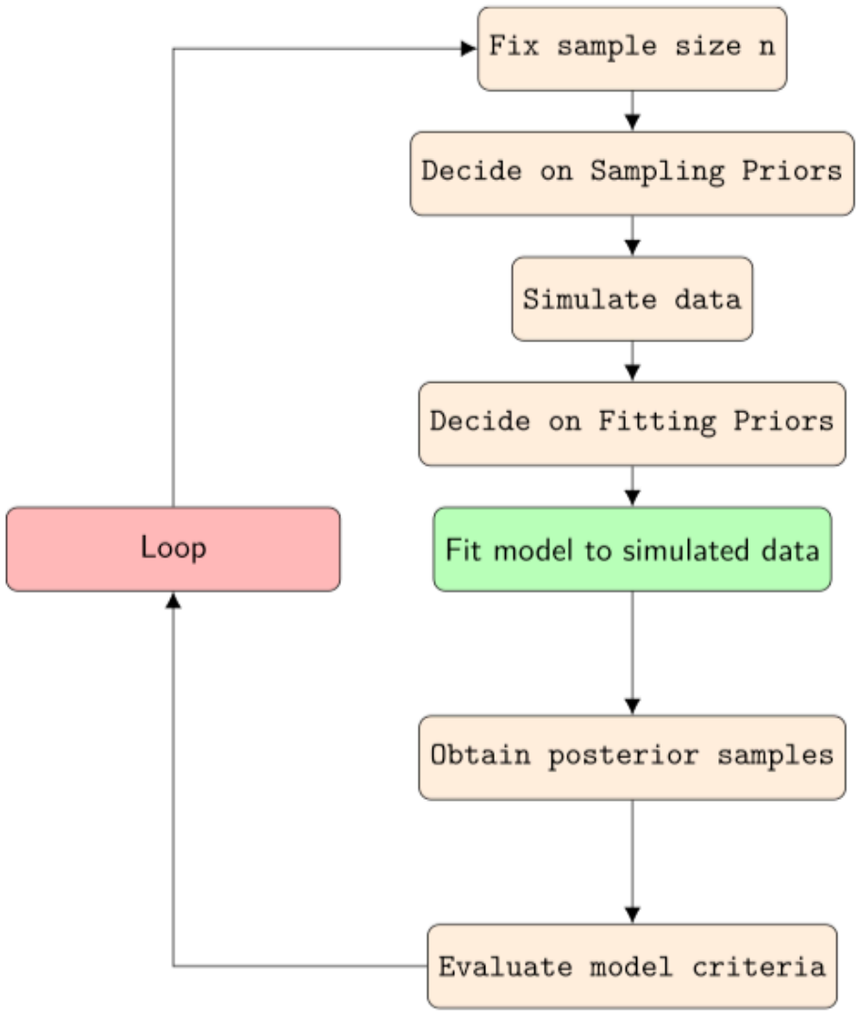


Fig. 1 A modified version of the workflow suggested by Wang and Gelfand (2002). The box colored green (labeled "Fit model to simulated data") can be computationally very intensive. The box labeled "Loop" indicates that the procedure has to be repeated for each sample size chosen; this step will also be computationally intensive

1. Bestimmte eine Verteilung für den gesuchten Parameter, z.B. IQ = N(100, 15)
2. Bestimmte Entscheidungskriterium (z.B. HDI schmäler als 10 IQ-Punkte)
3. Dann wiederhole für steigende Stichproben-Größen (N):
    1. Simuliere eine Prior-Prädiktiv-Verteilung oft (z.B. n_{iter}=100) für eine bestimmte Stichprobengröße n
    2. Berechne dein Modell mit vagen Priors and berechne die Posteriori-Verteilung
    3. Erstelle ein Konfidenzintervall mit den Schätzwerten, ist es schmäler als der Grenzwert, bist du fertig, sonst mache weiter mit größerem n

Vasishth, S., Yadav, H., Schad, D. J., & Nicenboim, B. (2022). Sample Size Determination for Bayesian Hierarchical Models Commonly Used in Psycholinguistics. *Computational Brain & Behavior*. https://doi.org/10.1007/s42113-021-00125-y

Effektstärke in der Population muss bekannt (geschätzt) sein

… damit man die optionale Stichprobengröße berechnen kann.

Table 1
ES Indexes and Their Values for Small, Medium, and Large Effects

Test	ES index	Effect size		
		Small	Medium	Large
1. m_A vs. m_B for independent means	$d = \frac{m_A - m_B}{\sigma}$	.20	.50	.80
2. Significance of product–moment r	r	.10	.30	.50
3. r_A vs. r_B for independent rs	$q = z_A - z_B$ where z = Fisher's z	.10	.30	.50
4. $P = .5$ and the sign test	$g = P - .50$	.05	.15	.25
5. P_A vs. P_B for independent proportions	$h = \phi_A - \phi_B$ where ϕ = arcsine transformation	.20	.50	.80
6. Chi-square for goodness of fit and contingency	$w = \sqrt{\sum_{i=1}^{k} \frac{(P_{1i} - P_{0i})^2}{P_{0i}}}$	.10	.30	.50
7. One-way analysis of variance	$f = \frac{\sigma_m}{\sigma}$	.10	.25	.40
8. Multiple and multiple partial correlation	$f^2 = \frac{R^2}{1 - R^2}$	.02	.15	.35

Note. ES = population effect size.

Für Regressionen kann man f^2 verwenden.

Cohen, J. (1992). A power primer. Psychological Bulletin, 112(1), 155–159.

Cohens Richtlinien zur Stichprobengröße

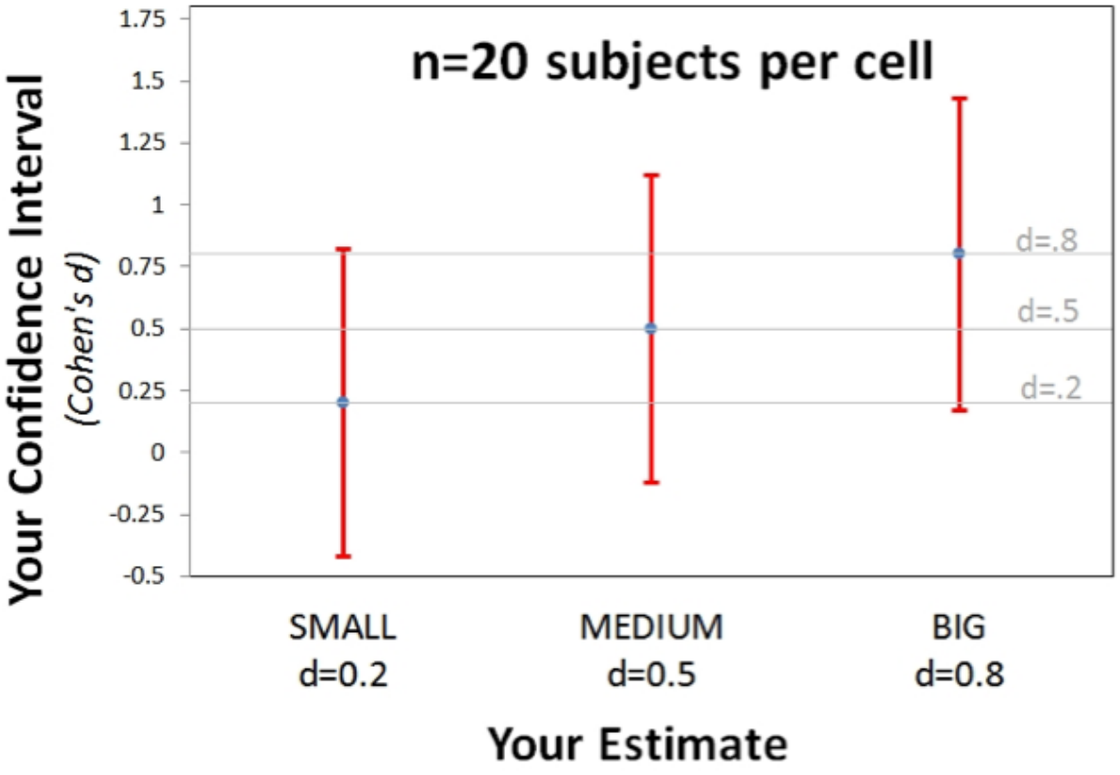
Table 2
N for Small, Medium, and Large ES at Power = .80 for α = .01, .05, and .10

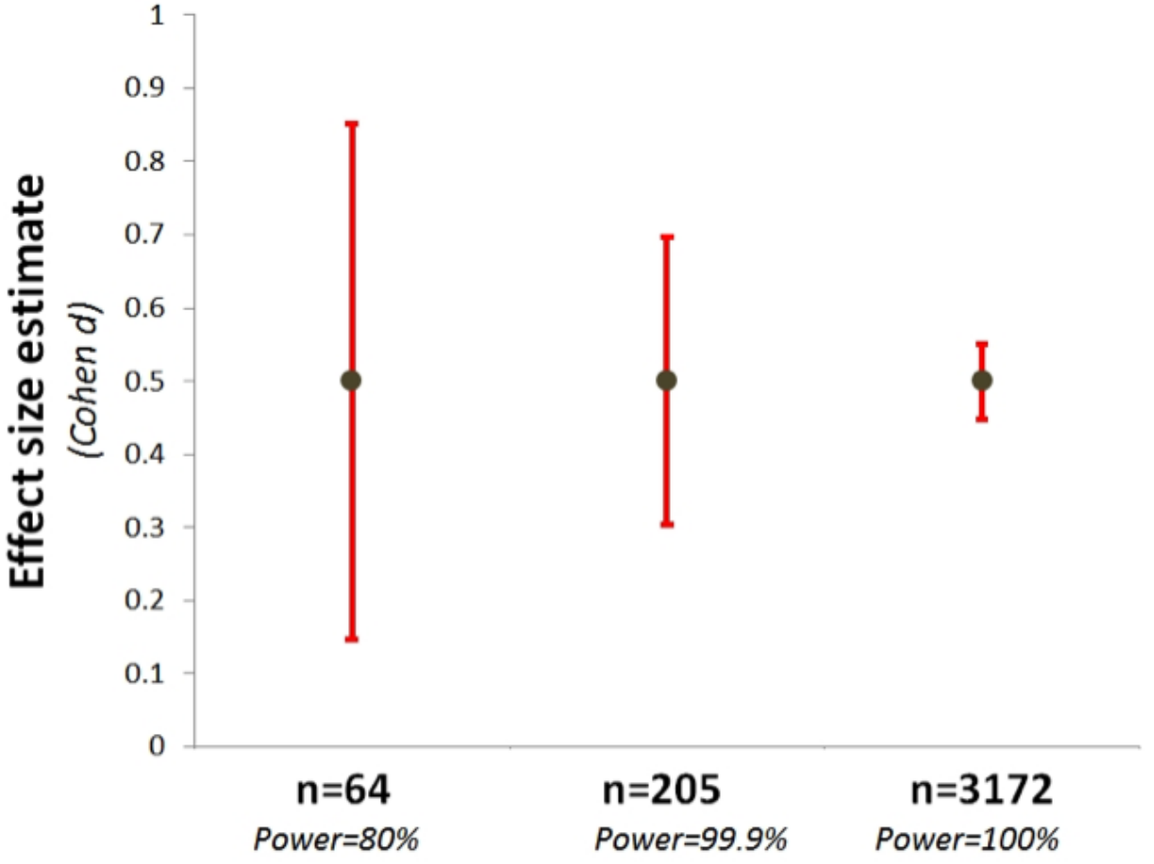
	α								
	.01			.05			.10		
Test	Sm	Med	Lg	Sm	Med	Lg	Sm	Med	Lg
1. Mean dif	586	95	38	393	64	26	310	50	20
2. Sig *r*	1,163	125	41	783	85	28	617	68	22
3. *r* dif	2,339	263	96	1,573	177	66	1,240	140	52
4. *P* = .5	1,165	127	44	783	85	30	616	67	23
5. *P* dif	584	93	36	392	63	25	309	49	19
6. χ^2									
1*df*	1,168	130	38	785	87	26	618	69	25
2*df*	1,388	154	56	964	107	39	771	86	31
3*df*	1,546	172	62	1,090	121	44	880	98	35
4*df*	1,675	186	67	1,194	133	48	968	108	39
5*df*	1,787	199	71	1,293	143	51	1,045	116	42
6*df*	1,887	210	75	1,362	151	54	1,113	124	45
7. ANOVA									
2*g*[a]	586	95	38	393	64	26	310	50	20
3*g*[a]	464	76	30	322	52	21	258	41	17
4*g*[a]	388	63	25	274	45	18	221	36	15
5*g*[a]	336	55	22	240	39	16	193	32	13
6*g*[a]	299	49	20	215	35	14	174	28	12
7*g*[a]	271	44	18	195	32	13	159	26	11
8. Mult *R*									
2*k*[b]	698	97	45	481	67	30			
3*k*[b]	780	108	50	547	76	34			
4*k*[b]	841	118	55	599	84	38			
5*k*[b]	901	126	59	645	91	42			
6*k*[b]	953	134	63	686	97	45			
7*k*[b]	998	141	66	726	102	48			
8*k*[b]	1,039	147	69	757	107	50			

Note. ES = population effect size, Sm = small, Med = medium, Lg = large, diff = difference, ANOVA = analysis of variance. Tests numbered as in Table 1.
[a] Number of groups. [b] Number of independent variables.

Cohen, J. (1992). A power primer. Psychological Bulletin, 112(1), 155–159.

Präzision ist teuer

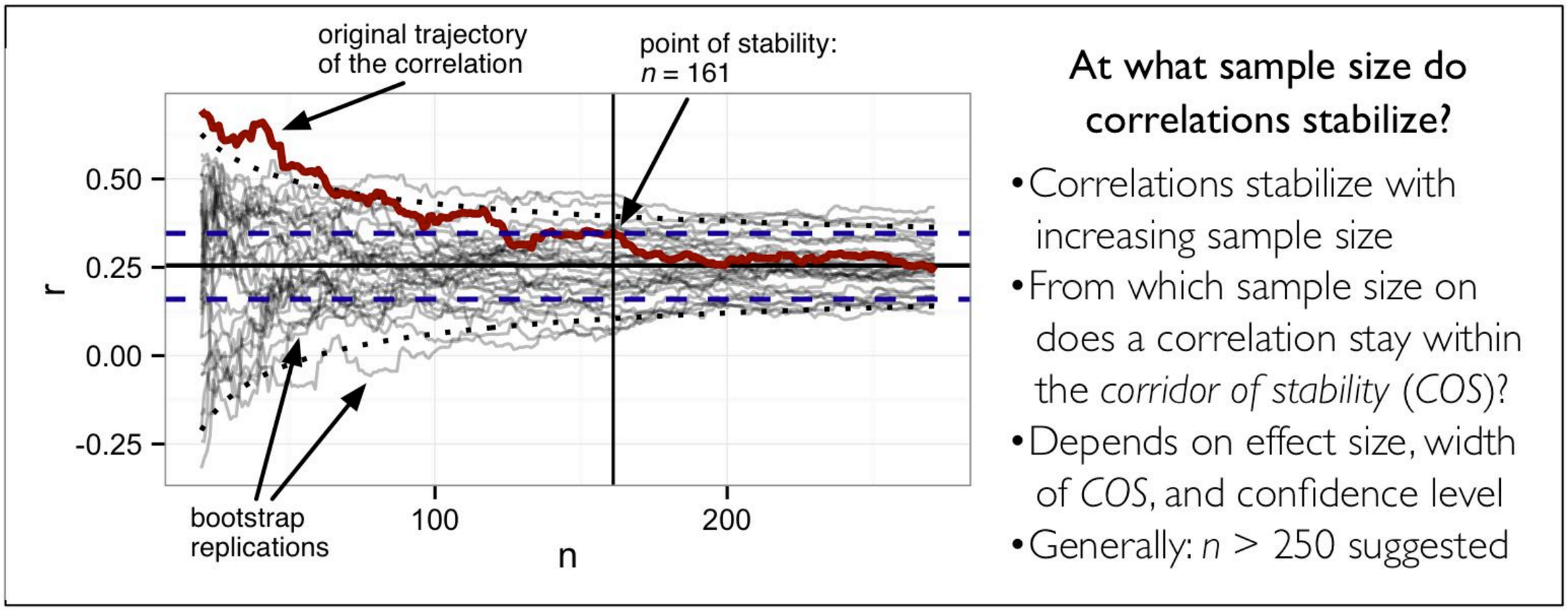




Daher führen kleine Studien auch zu einer Überschätzung der Effektstärke: Aufgrund der großen Spannweite an Effekten treten auch wenn die H0 gilt (d=0) mitunter recht starke Effekte auf (z. B. d=.5), die auch signifikant sein können. Die schwachen Effekte werden dann unterschlagen („file drawer effect“ bzw. „publication bias“), die zufällig starken publiziert. Das verzerrt dann das Gesamtbild zur wahren Effektstärke.

http://datacolada.org/2014/05/01/20-we-cannot-afford-to-study-effect-size-in-the-lab/

At what sample size do correlations stabilize?



https://osf.io/rdasy/

Volltext

Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize? *Journal of Research in Personality*, *47*(5), 609–612. https://doi.org/10.1016/j.jrp.2013.05.009

Die meisten Studien sind „unter-powered“

PLOS ONE | Publish | About | Browse | Search
advanced search

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

The N-Pact Factor: Evaluating the Quality of Empirical Journals with Respect to Sample Size and Statistical Power

R. Chris Fraley, Simine Vazire

Published: October 8, 2014 • DOI: 10.1371/journal.pone.0109019

0 Saves | 6 Citations | 7,177 Views | 166 Shares

Article | Authors | Metrics | Comments | Related Content

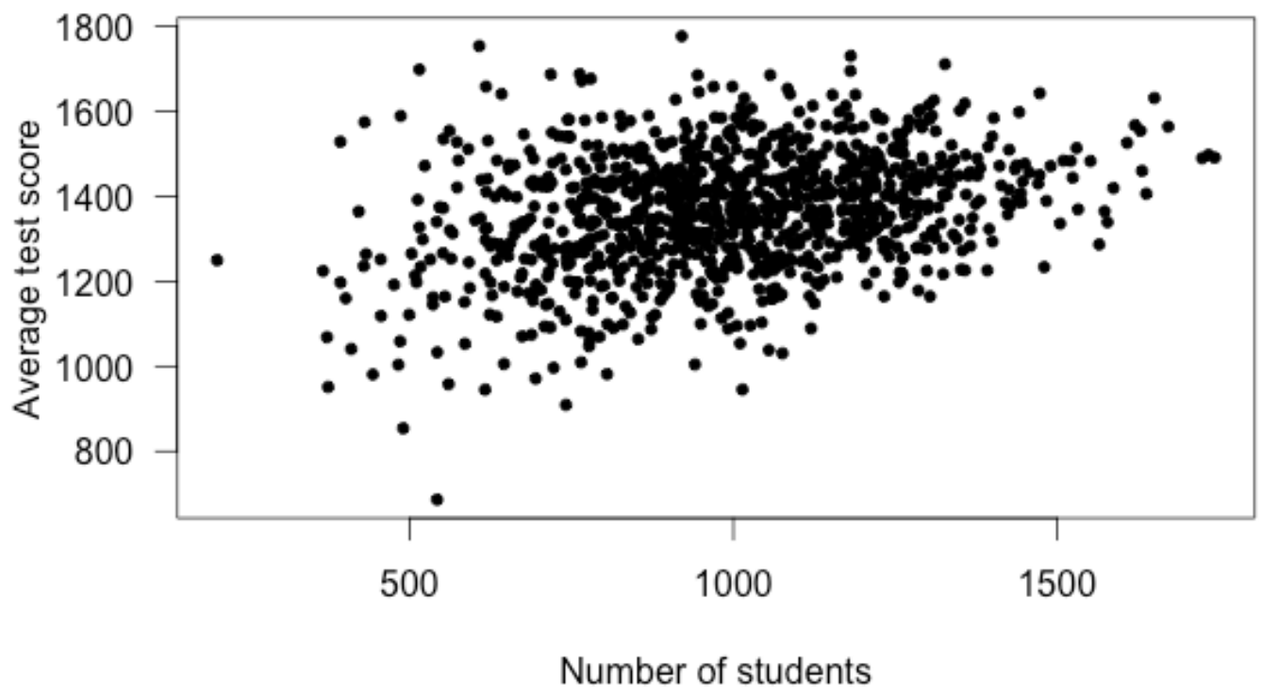
Download PDF | Print | Share

Abstract

The authors evaluate the quality of research reported in major journals in social-personality psychology by ranking those journals with respect to their *N*-pact Factors (NF)–the statistical power of the empirical studies they publish to detect typical effect sizes. Power is a particularly important attribute for evaluating research quality because, relative to studies that have low power, studies that have high power are more likely to (a) to provide accurate estimates of effects, (b) to produce literatures with low false positive rates, and (c) to lead to replicable findings. The authors show that the average sample size in social-personality research is 104 and that the power to detect the typical effect size in the field is approximately 50%. Moreover, they show that there is considerable variation among journals in sample sizes and power of the studies they publish, with some journals consistently publishing higher power studies than others. The authors hope that these rankings will be of use to authors who are choosing where to submit their best work, provide hiring and promotion committees with a superior way of quantifying journal quality, and encourage competition among journals to improve their NF rankings.

Bei großen Gruppen werden die Parameter gut geschätzt

- Stellen wir uns vor, Sie die Qualität von Schulen zu beurteilen; dazu werden „test scores“ ermittelt.
- Sie schauen sich die besten 5 % an: Aha, die kleinsten Schulen schneiden am besten ab. Überrascht mich nicht, ist doch logisch – familiäre Atmosphäre…
- Jetzt schauen Sie sich die untersten 5 % der Schulen an. Ja was ist denn das: Wieder lauter kleine Schulen! Da stimmt doch was nicht!



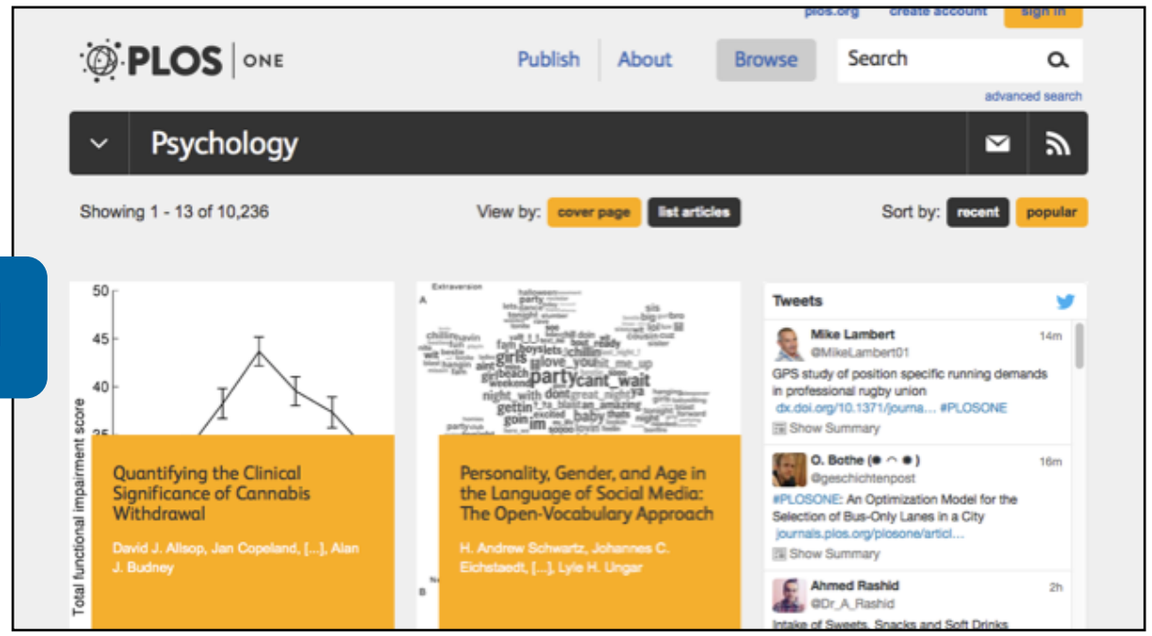
Tatsächlich hat alles seine Ordnung: In großen Schulen wir aufgrund der großen Stichprobengröße genau geschätzt – es resultieren mittlere Werte. Bei kleinen Schulen ist die Schätzgenauigkeit schlechter. Die Qualitätswerte streuen mehr, man findet mehr Extreme. In beide Richtungen.

http://www.statisticsdonewrong.com/regression.html

Erarbeiten Sie das Studiendesign einer Studie!

1

Suchen Sie sich eine empirische Studie.



z. B. hier: www.plosone.org

2

Werten Sie sie mit diesen Fragen aus:

1. Name der Studie
2. Forschungsfrage
3. zentrale Hypothesen (UVs, AVs)
4. Versuchsdesign (Gruppen etc.)
5. Kontrollmechanismen (z. B. Manipulation Check)
6. Nicht kontrollierte Störvariablen
7. Zentrale Ergebnisse

3

Präsentieren Sie Ihre Ergebnisse

so, wie es Ihnen am meisten Spaß macht

Spielregeln

- **Ziel** der Übung ist es, das **Versuchsdesign** (sowie Ergebnisse) zu **verstehen**
- ca. 4 Personen pro Gruppe
- 45 Min. Vorbereitung
- Präsentation (ca. 3-5 Min. präsentieren)
- statistische Analysen und tiefere theoretische Überlegungen sind **nicht** Gegenstand der Übung!