Institut für Angewandte Statistik

LAQM – Linzer Arbeitsgruppe für Quantitative Methoden
04. November 2020

Andreas Quatember

Signifikanz vs. Relevanz – Eine Annäherung

Einleitung

Empirische Untersuchungen wissenschaftlicher Fragestellungen:

- Werden Ehen, die aus Onlinebekanntschaften entstehen, seltener geschieden als andere?
- Sind beim PISA-Test im Kompetenzbereich Mathematik Burschen besser als Mädchen?
- Führen asymptomatische Corona-Infektionen zu geringerer Immunität als symptomatische?

Im Jahr 2013 ergab eine groß angelegte Studie mit mehr als 19000 Teilnehmern unter der Leitung von John T. Cacioppo von der University of Chicago, dass sich Eheleute, die sich online kennen gelernt haben, mit geringerer Wahrscheinlichkeit wieder scheiden lassen ($p < 0{,}002$). Außerdem waren die noch Verheirateten unter ihnen mit ihrer Ehe tendenziell zufriedener als jene, die auf traditionellem Weg zusammengekommen waren ($p < 0{,}001$). Diese Werte sehen eindrucksvoll aus, aber der eigentliche Effekt war winzig: Kennenlernen über das Internet drückte die Scheidungsrate von 7,67 auf 5,96 Prozent und hob die Zufriedenheit mit der Ehe von 5,48 auf 5,64 auf einer Sieben-Punkte-Skala.

Die Signifikanz eines Ergebnisses sage eben nichts über dessen praktische Bedeutung aus, erklärt Geoff Cumming, emeritierter Psychologieprofessor von der La Trobe University in Melbourne (Australien). »Wir sollten uns fragen: ›Wie groß ist der Effekt, mit dem wir es zu tun haben?‹ und nicht: ›Gibt es überhaupt einen Effekt?‹«

Da es sich bei PISA um sehr große Stichproben pro Teilnehmerland handelt, können teilweise auch sehr kleine Unterschiede (beispielsweise zwischen Mädchen und Burschen) als statistisch signifikant nachgewiesen werden. Das heißt, dass die Unterschiede mit großer Wahrscheinlichkeit in der betreffenden Population existieren, aber relativ klein sein können, sodass sie in praktischer Hinsicht nicht relevant sind. Einige Ergebnisdarstellungen beinhalten daher Angaben zur Effektstärke eines Unterschieds[9]. Damit wird zusätzlich zur statistischen Signifikanz auch die Größenordnung von Unterschieden quantifiziert. Eine grobe Faustregel für die Einschätzung der Größenordnung sind die Richtwerte nach Cohen (1988), nach denen Effektstärken ab 0.2 als kleine, ab 0.5 als mittlere und ab 0.8 als große Unterschiede eingestuft werden. Demnach kann ab einer Effektstärke von 0.2 (d. h., der Unterschied macht 20 % der Standardabweichung aus) zusätzlich zu einer statistischen Signifikanz auch von einer praktischen Relevanz ausgegangen werden.

Suchan, B., et al. (Hrsg.) (2019). PISA 2018. Grundkompetenzen am Ende der Pflichtschulzeit im internationalen Vergleich. Leykam, Graz, S. 36

Der Aspekt der praktischen Relevanz

Statistische Signifikanztests können einen faktenbasierten Anhaltspunkt im Rahmen des Entscheidungsprozesses zwischen zwei Hypothesen über einen interessierenden Sachverhalt liefern

▶ **Beispiel:** Bei der Differenz D der Populationsmittelwerte von Burschen und Mädchen waren die Hypothesen offenbar

$H_0: D \leq 0$ und $H_1: D > 0$

Vollerhebung der Schüler*innen: Berechnung des wahren D und korrekte Entscheidung für H_0/H_1 (unter impliziten Annahmen betreffend Operationalisierung, Auswahlrahmen, Nonresponse, Messung, …)

Aber ist jede wahre Differenz $D > 0$ denn auch praktisch relevant? - Wenn nicht, warum hat man dann die Einshypothese so formuliert?

Der statistische Signifikanztest

Allgemeine Handlungslogik des statistischen Testens von Hypothesen auf Stichprobenbasis nach Fisher (1935)[1]:

- Aufstellen von statistischen Hypothesen auf Basis einer theoriegestützten Forschungshypothese
- Berechnung einer der statistischen Fragestellung entsprechenden Teststatistik d in einer Stichprobenerhebung
- Einschätzung von $d = d_0$ als auf dem Signifikanzniveau α schwaches oder starkes Indiz gegen H_0 mittels seines p-Wertes $P(d \geq d_0 | H_0)$, der den Grad der Unvereinbarkeit der Stichprobendaten mit H_0 misst

[1] Fisher, R. A. (1971). *The Design of Experiments*. 8. Auflage. Hafner Publishing Company, New York. Ch. 2 (vgl. http://hdip-data-analytics.com/_media/resources/pdf/fisher-1956.pdf; Zugegriffen: 21.10.2020)

Standardmäßig erfolgt die Hypothesenformulierung vollkommen unabhängig vom jeweiligen Untersuchungsgegenstand durch

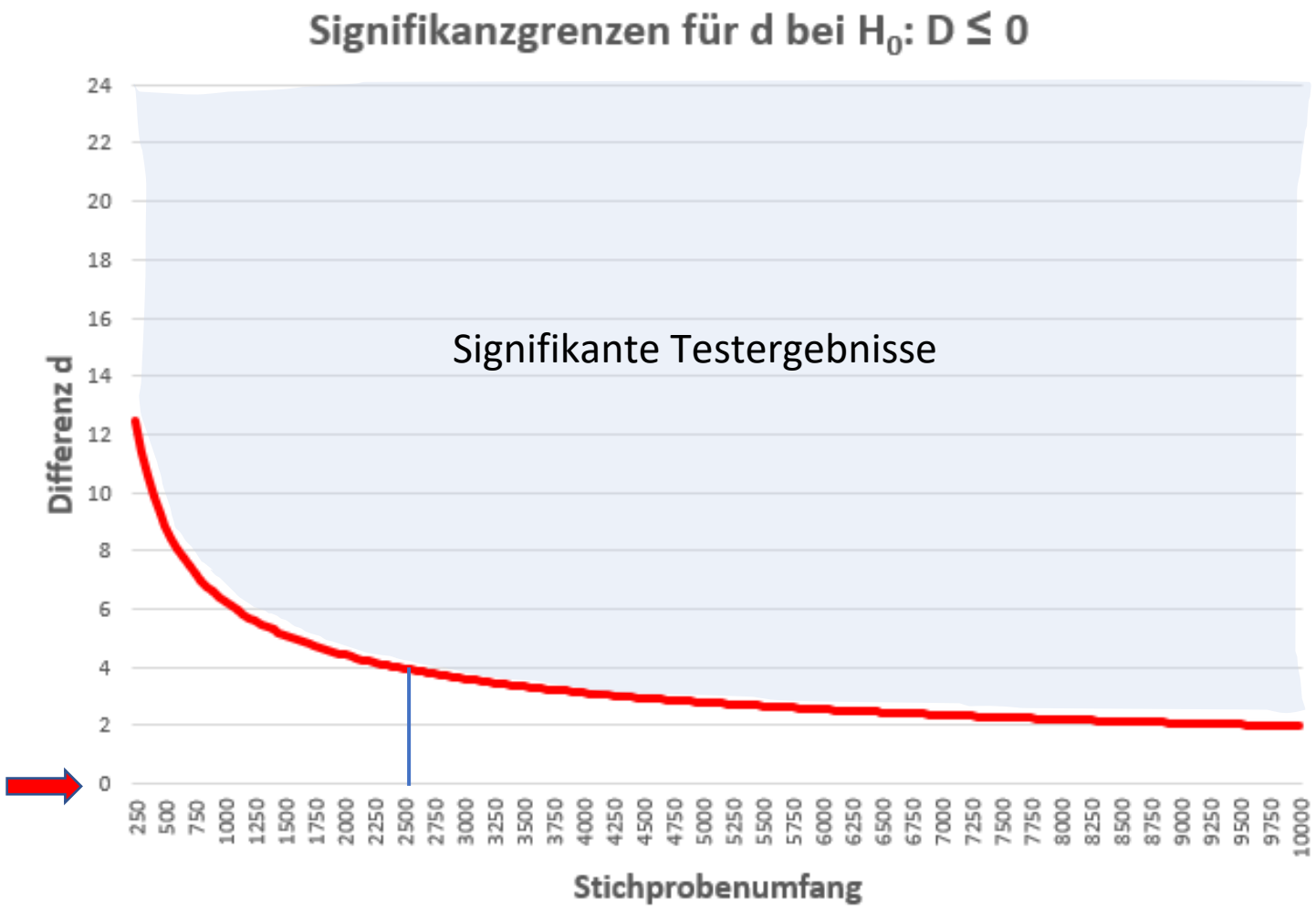
H_0: $D \leq 0$ und H_1: $D > 0$

(siehe PISA) bzw.

H_0: $D = 0$ und H_1: $D \neq 0$

Es wird als Schwäche des statistischen Signifikanztests interpretiert, dass mit zunehmendem Stichprobenumfang auch beliebig gering von H_0 abweichende „Effekte“ $D > 0$ (bzw. $D \neq 0$) mit immer größerer Wahrscheinlichkeit signifikant werden (➔ Erhöhung der *Testpower*)

Problematisch ist dies insbesondere bei zweiseitigen Fragestellungen und speziell im Big Data-Kontext



Obere Schranke: $d_o = 0 + u_{1-\alpha} \cdot \sqrt{\hat{V}(d|D = D_R)}$ (z. B. bei n = 2500: 3,95)

Aber ist ein signifikanter Mittelwertsunterschied von z. B. d = 4 auch praktisch bedeutsam?

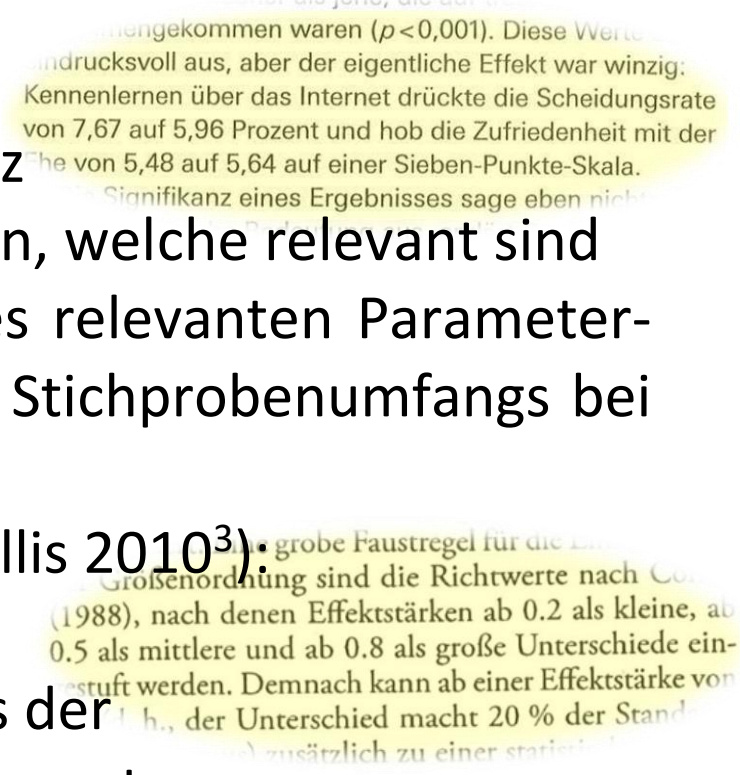
Es hängt vom jeweiligen Untersuchungsgegenstand ab, in welcher Größenordnung Abweichungen von H_0: $D \leq 0$ praktisch bedeutsam sind und welche nicht (Kosten-Nutzen-Abwägung)

Die aufgestellten statistischen Hypothesen *müssen* auch das prüfen, was man prüfen möchte

Dies erfordert Bestimmung einer kontextbezogenen Relevanzgrenze D_R, welche die nichtrelevanten von den relevanten Parameterwerten trennt

Es gibt für die Festlegung von D_R verschiedene denkbare Ansätze:

- **Erfahrungsgeleiteter Ansatz**:
  Expert*innen, die angeben, dass eine Differenz nicht relevant ist, sollten auch angeben können, welche relevant sind
- **Testpowergeleiteter Ansatz**: Verwendung des relevanten Parameterwertes, der bei der Bestimmung des nötigen Stichprobenumfangs bei vorgegebener Testpower zu fixieren ist
- **Konventionsansatz** (vgl. etwa: Cohen 1969[2], Ellis 2010[3]):
  Z. B. Definition von kleinen, mittleren und starken Effekten basierend auf dem Verhältnis der Effektgröße zur Standardabweichung der Messwerte



[2] Cohen, J. (1969). *Statistical Power Analysis for the Behavioral Sciences*. Academic Press, New York.
[3] Ellis, P. D. (2010). *The Essential Guide to Effect Sizes: Statistical power, Meta-Analysis, and the Interpretation of Research Results*. Cambridge University Press, Cambridge.

Verschiedene Vorgehensweisen (vgl. Goodman et al. 2019)[4]:

- **„D > 0-Test“**: Berücksichtigung der Stichprobenschwankung bei kontextunabhängiger H_0: $D \leq 0$; keine Berücksichtigung der kontextbezogenen Relevanz $D_R > 0$
- **Mindesteffektgröße D_R der Teststatistik**: Berücksichtigung der Relevanz; nicht der Stichprobenschwankung bei $D = D_R$

[4] Goodman, W. M:, Spruill, S. E., Komaroff, E. (2019). A Proposed Hybrid Effecrt Size Plus p- Value Criterion: Empirical Evidence Supportuing its Use. The American Statistician. 73(1), 168-185.

GESUNDHEIT

Corona-Studie: Geringere Immunität nach milden Infektionen

Von nachrichten.at/apa 19. Juni 2020 07:20 Uhr

Nur 62,2 Prozent aus der Gruppe ohne Symptome hatten wenige Wochen nach der Infektion noch Kurzzeit-Antikörper im Blut - verglichen mit 78,4 Prozent der symptomatischen Patienten.

https://www.nachrichten.at/meine-welt/gesundheit/corona-studie-geringere-immunitaet-nach-milden-infektionen;art114,3267740 (Zugegriffen: 21.10.2020)

the symptomatic group tested positive for IgG approximately 3–4 weeks after exposure. Moreover, 62.2% (23/37) of the asymptomatic group were positive for IgM, whereas 78.4% (29/37) of the symptomatic group were IgM positive. Interestingly, IgG levels in the

Long, Q.-X. et al. (2020). Clinical and immunological assessment of asymptotic SARS-CoV-2 infections. nature medicine LETTERS.

- **„D > 0-Test“ plus Mindesteffektgröße D_R**: Berücksichtigung der Relevanz; nicht der dazu gehörigen Stichprobenschwankung
- **Konfidenzintervall**: Überdeckungswahrscheinlichkeit entspricht nicht dem Signifikanzniveau (keine p-Werte); nicht immer anwendbar (z. B. beim χ^2- oder beim F-Test)

Signifikanz vs. Relevanz: Eine Annäherung

Die von den Anwendenden aufgestellten statistischen Hypothesen sind häufig unbrauchbar, weil sie nicht die korrekte Übersetzung der Forschungshypothesen in kontextbezogene, vorwissensgeleitete statistische Hypothesen sind

Eine geeignete Teststrategie sollte gleichzeitig die Stichprobenschwankung bei der tatsächlichen H_0 *und* die Relevanz der Testergebnisse berücksichtigen

Soll etwa überprüft werden, ob ein praktisch bedeutsames D der Leistungen von Burschen und Mädchen vorliegt, dann *darf* H_1 nur jene D`s enthalten, die in diesem Kontext praktisch bedeutsam sind, und H_0 nur jene, die praktisch unbedeutend sind

Legt man z. B. D = 10 als Relevanzgrenze D_R fest, dann ergeben sich beim einseitigen Test auf praktisch bedeutsame Differenzen die Hypothesen

H_0: $D \leq 10$ und H_1: $D > 10$

Dieser Schritt führt von – den jeweiligen Kontext nicht beachtenden – „D > 0-Tests“ zu kontextbezogenen Signifikanztests, den Relevanztests

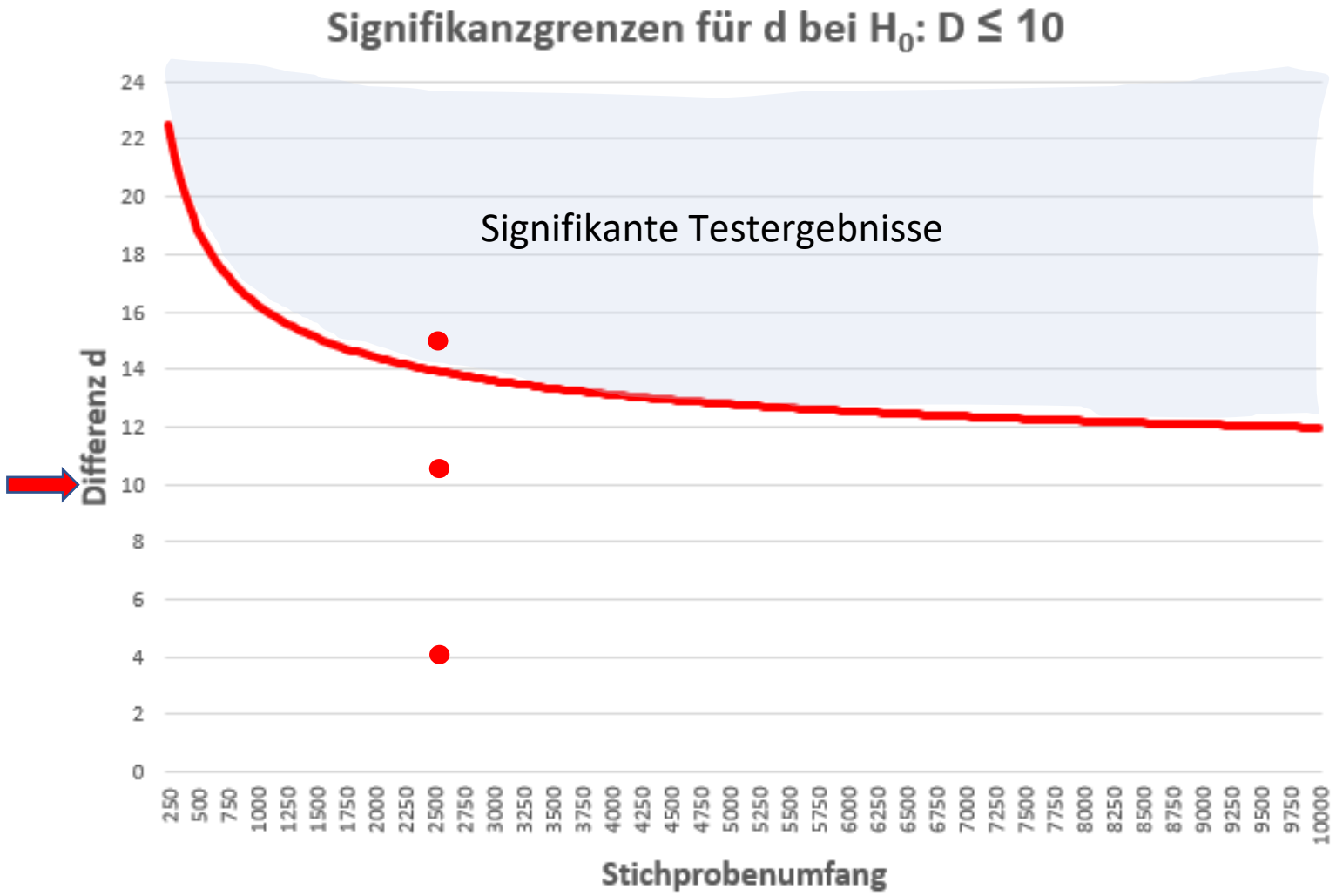
Nur wenn tatsächlich alle D > 0 als relevant betrachtet werden, ergeben sich die Hypothesen von „D > 0-Tests“

Beim Testen von

$$H_0: D \leq D_R \text{ und } H_1: D > D_R$$

erhält man mit einem p-Wert $\leq \alpha$ ein signifikantes praktisch relevantes Testergebnis

Die Erhöhung der Testpower durch Erhöhung des Stichprobenumfangs macht auf dem Signifikanzniveau α keine inhaltlich unbedeutenden Testergebnisse mehr signifikant, sondern ausschließlich relevante



Obere Schranke: $d_o = D_R + u_{1-\alpha} \cdot \sqrt{\hat{V}(d|D = D_R)}$

Bei zweiseitigen Fragestellungen ergeben sich die Hypothesen durch

$$H_0: D \in [D_{Ru} ; D_{Ro}] \text{ und } H_1: D \notin [D_{Ru} ; D_{Ro}]$$

mit D_{Ru} und D_{Ro}, der unteren bzw. oberen Relevanzgrenze

Weitere Beispiele für Relevanzteststrategien[5]

Der konzeptionelle Schritt vom standardmäßigen „D > 0-“ zum inhaltsgetriebenen Relevanztest ist auf alle statistischen Fragestellungen

- durch Festlegung adäquater Relevanzschwellen und
- Adaptierung der statistischen Teststrategie

zu übertragen

[5] vgl. Quatember, A. (2005). Das Signifikanz-Relevanz-Problem beim statistischen Testen von Hypothesen. ZUMA-Nachrichten, 57, Jg. 29, 1-23.

Beispiel 1 Korrelationstests

Festlegung der Relevanzgrenze ρ_R z. B. auf Basis des Bestimmtheitsmaßes

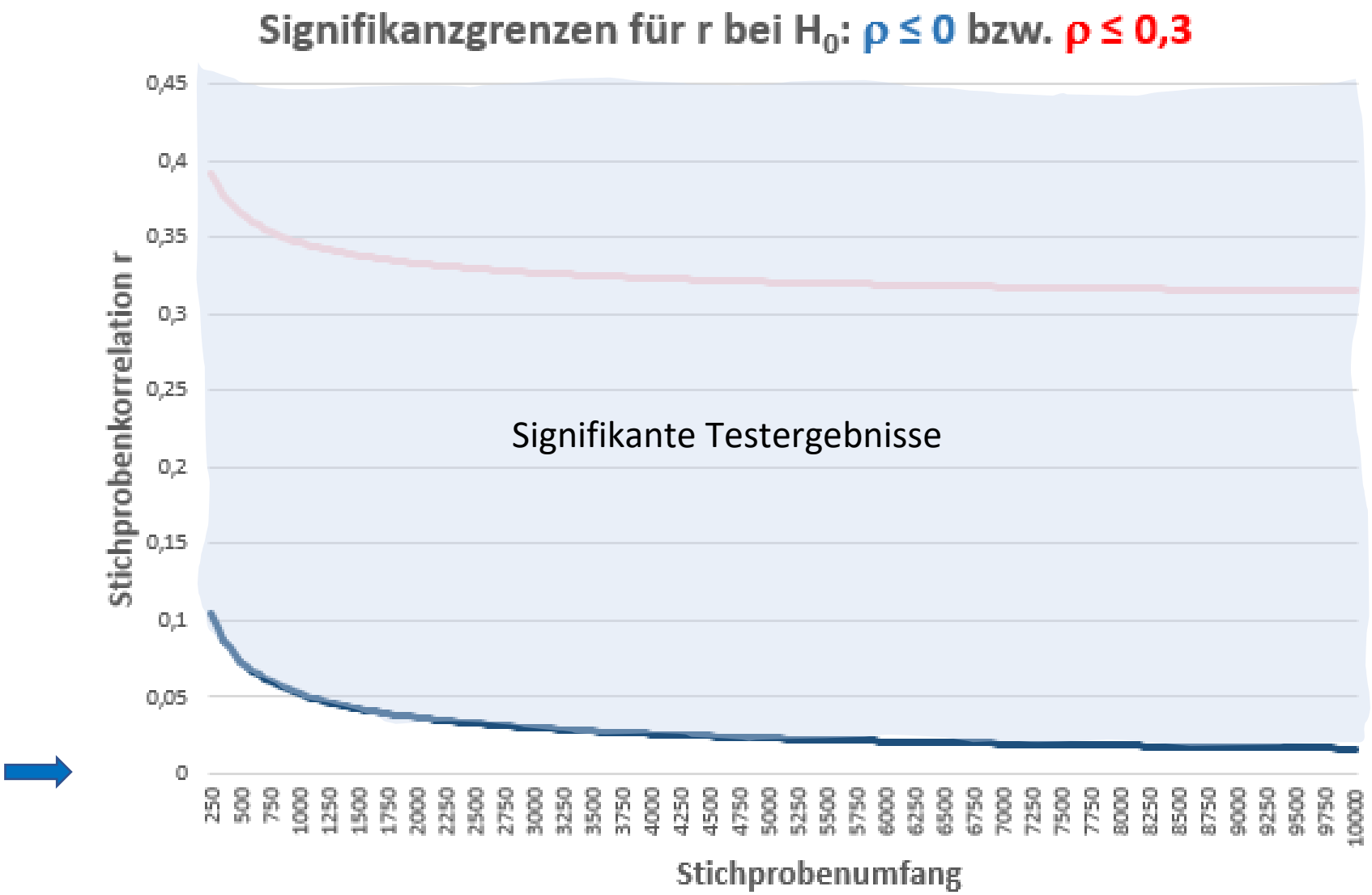
Cohen (1969)[6] definiert z. B. ρ von $\rho_R = 0{,}1$, 0,3 und 0,5 als kleinen, mittleren und großen Effekt

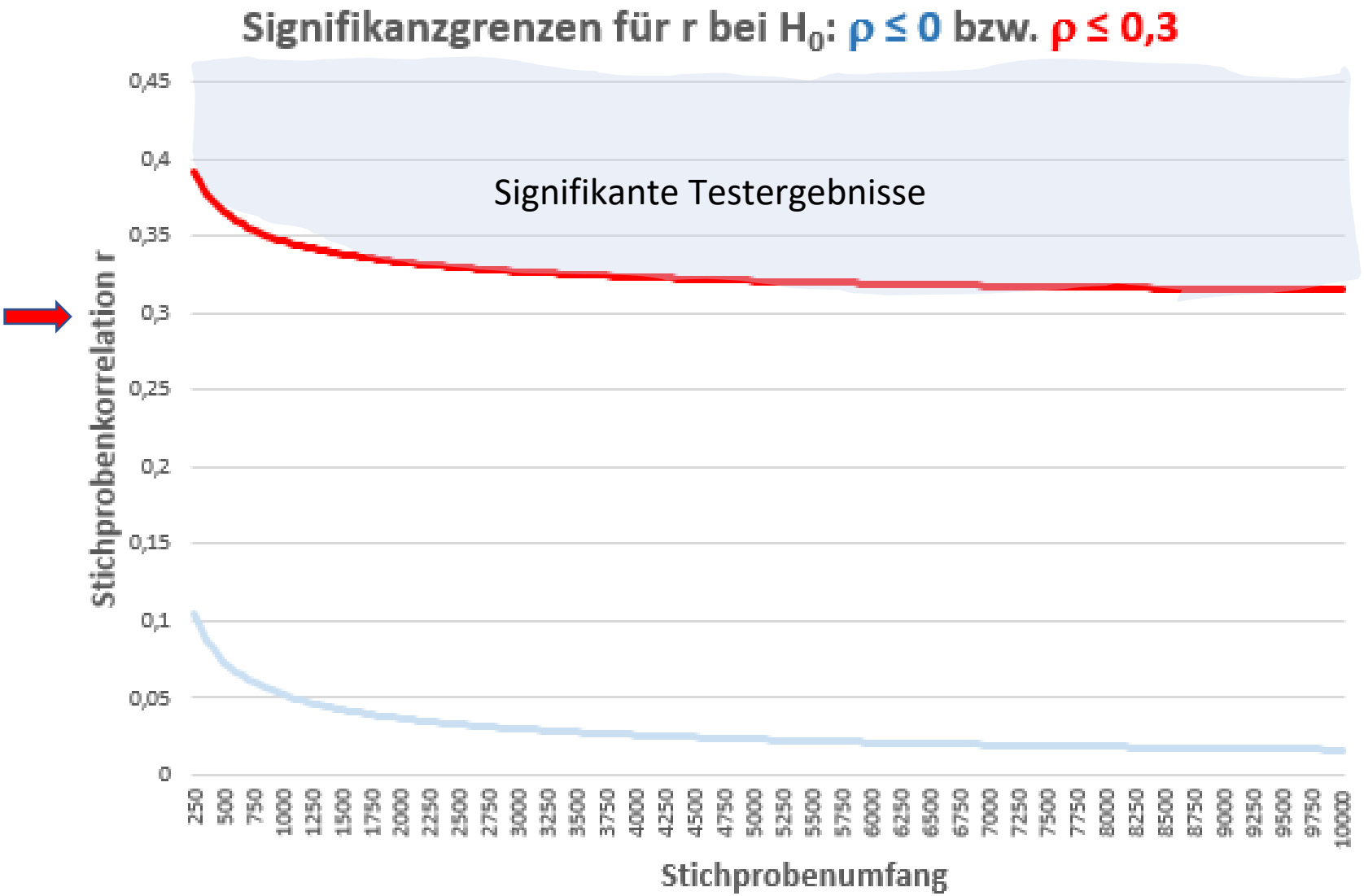
Unterschiedliche Stichprobenverteilungen von r unter H_0: $\rho \leq 0$ und z. B. H_0: $\rho \leq 0{,}3$ erfordern unterschiedliche Teststatistiken

Bei H_0: $\rho \leq 0$ ist die Teststatistik $z = r \cdot \sqrt{\frac{n-2}{1-r^2}}$ standardnormalverteilt,

während es bei H_0: $\rho \leq 0{,}3$ $z = \frac{\sqrt{n-3}}{2} \cdot \left(\ln \frac{1+r}{1-r} - \ln \frac{1+\rho_R}{1-\rho_R} \right)$ ist

[6] Cohen, J. (1969). *Statistical Power Analysis for the Behavioral Sciences*. Academic Press, New York.


Signifikanzgrenzen für r bei H_0: $\rho \leq 0$ bzw. $\rho \leq 0{,}3$
Stichprobenkorrelation r
0,45
0,4
0,35
0,3
0,25
0,2
0,15
0,1
0,05
0
Signifikante Testergebnisse
250
500
750
1000
1250
1500
1750
2000
2250
2500
2750
3000
3250
3500
3750
4000
4250
4500
4750
5000
5250
5500
5750
6000
6250
6500
6750
7000
7250
7500
7750
8000
8250
8500
8750
9000
9250
9500
9750
10000
Stichprobenumfang


Signifikanzgrenzen für r bei H_0: ρ ≤ 0 bzw. ρ ≤ 0,3
Signifikante Testergebnisse
Stichprobenkorrelation r
0,45
0,4
0,35
0,3
0,25
0,2
0,15
0,1
0,05
0
250
500
750
1000
1250
1500
1750
2000
2250
2500
2750
3000
3250
3500
3750
4000
4250
4500
4750
5000
5250
5500
5750
6000
6250
6500
6750
7000
7250
7500
7750
8000
8250
8500
8750
9000
9250
9500
9750
10000
Stichprobenumfang

Beispiel 2 Einfache Varianzanalyse

Festlegung der Relevanzgrenze an Hand des Verhältnisses $\Phi = \frac{\sigma_Z^2}{\sigma_I^2}$

σ_Z^2 … Varianz zwischen den Gruppen

σ_I^2 … Varianz innerhalb der Gruppen

Cohen (1969)[7] definiert für Φ ein Ausmaß von $\Phi_R = 0{,}1$, $0{,}25$ und $0{,}4$ als kleinen, mittleren und großen Effekt

Bei H_0: $\Phi = 0$ ist die Teststatistik f zentral F-verteilt, bei H_0: $\Phi \leq 0{,}1$ non-zentral F-verteilt

[7] Cohen, J. (1969). *Statistical Power Analysis for the Behavioral Sciences*. Academic Press, New York.

Damit ergibt sich beispielsweise bei 4 Gruppen mit Stichprobenumfängen von je 1000 bei H_0: $\Phi = 0$ eine Signifikanzschranke von f = 2,607, während diese bei H_0: $\Phi \leq 0{,}1$ den Wert von 21,963 aufweist

Zusammenfassung und Ausblick

Die Signifikanz-Relevanz-Problematik besteht nur bei falscher Übersetzung der Forschungshypothese in die statistische Einshypothese des Signifikanztests

Ihre korrekte Übersetzung führt zu Relevanztests, welche die inhaltliche Bedeutsamkeit der Testergebnisse und die Stichprobenschwankung berücksichtigen

Gegenüber dem herkömmlichen Handlungsablauf von „D > 0-Tests“ ist dafür eine inhaltlich begründete Relevanzgrenze zu bestimmen

Entsprechende Teststatistiken und Stichprobenverteilungen sind darauf basierend zu verwenden

Weitere Problemstellungen der Anwendung von Signifikanztests:

- Multiples Alles-mit-Allem-Testen („p-hacking“) ohne theoriegetriebene Forschungshypothesen
- Publication Bias
- Anteil zweiseitiger Tests
- Nichtreproduzierbarkeit signifikanter Resultate gefördert durch Außerachtlassung der impliziten Annahmen (bzgl. Stichprobendesign, Nonresponse, …) und expliziten Modelle

Aber das sind andere Geschichten (vgl. Fisher 1935)

Vielen Dank für Ihre signifikante und
für mich äußerst relevante Aufmerksamkeit!