

M24 Statistik 1: Wintersemester 2023/24

Vorlesung 07: Effektstärke

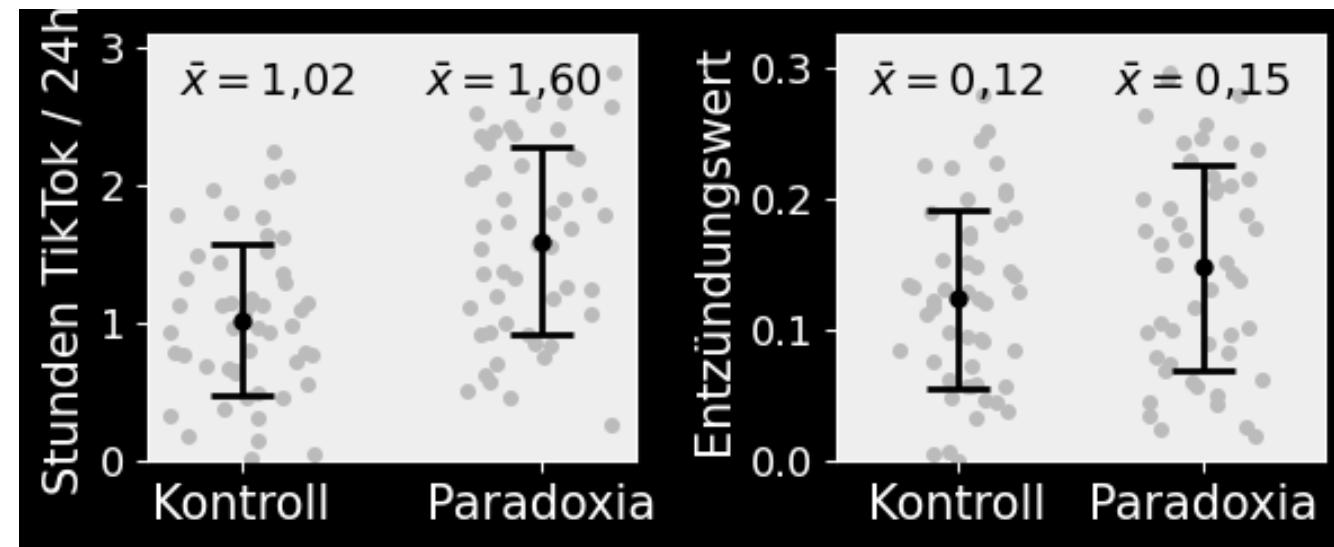
Prof. Matthias Guggenmos

Health and Medical University Potsdam

28.11.2023



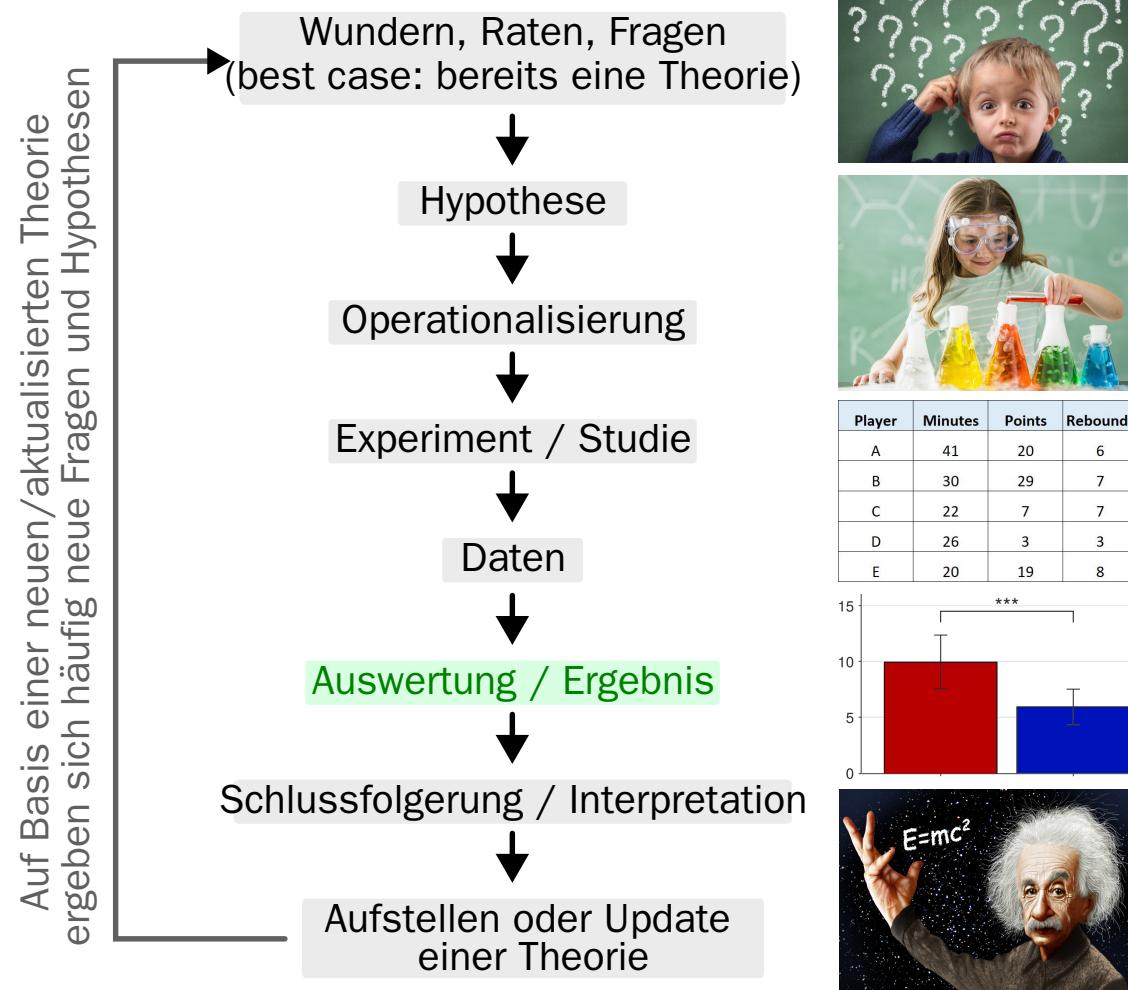
Sie sinnieren weiterehin über das Ergebnis Ihrer Beobachtungsstudie. Paradoxiker verbringen sowohl mehr Zeit auf TikTok, als auch weisen sie höhere Entzündungswerte auf. Zwar ist der Mittelwertsunterschied bei der TikTok-Zeit größer, aber Sie wissen, dass TikTok-Zeit und Entzündungswerte völlig unterschiedliche Skalen und daher nicht vergleichbar sind.



Die Frage lautet: wie kann man die beiden Gruppenunterschiede bezüglich TikTok-Zeit und Entzündungsparametern vergleichbar machen? Wie können wir eine Aussage treffen, welcher der beiden Effekte stärker ist?

Effektstärke

Der Forschungsprozess

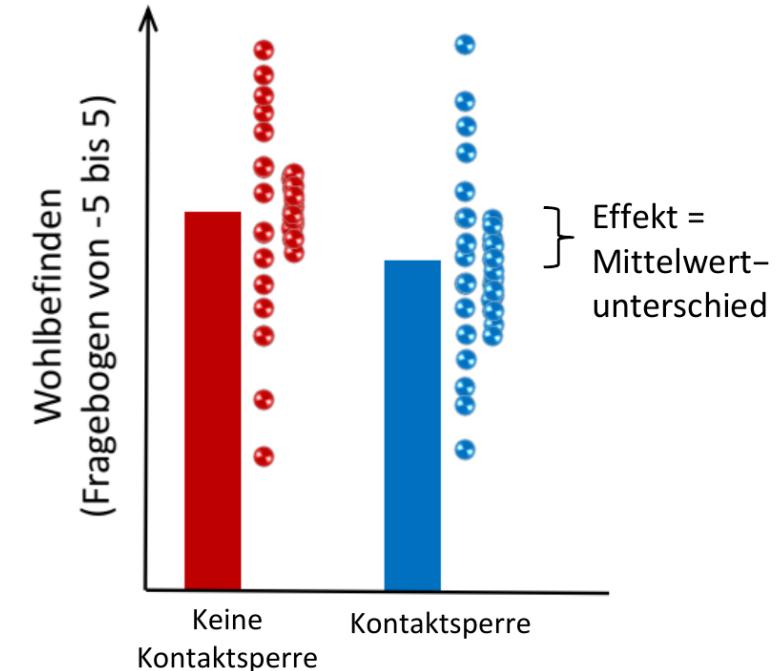


Player	Minutes	Points	Rebounds
A	41	20	6
B	30	29	7
C	22	7	7
D	26	3	3
E	20	19	8



Effektstärke

- In psychologischer Forschung untersuchen wir in den meisten Fällen die Auswirkung von Variablen X_i auf Variablen Y_i .
- Diese Auswirkung ist entweder als **Unterschied** (z.B. wenn X die Gruppenzugehörigkeit angibt) oder als **Zusammenhang** (wenn X und Y metrische Variablen mit einer vermuteten kausalen Wechselwirkung sind) messbar – man spricht auch von **Effekten**.
- Wie kann die Aussagekraft bzw. Bedeutsamkeit von Effekten bestimmt und kommuniziert werden?
 - ⇒ **statistische Signifikanz** (ab Vorlesung 10): kann ein Effekt *allein durch Zufall erklärt werden?*
 - ⇒ **praktische Signifikanz** (Effektstärken): ist die Stärke des Effektes *in der Praxis bedeutsam?*
- Die Stärke eines Effektes im Sinne der praktischen Signifikanz wird als **Effektstärke** oder **Effektgröße** bezeichnet. Wir werden nachfolgend den Begriff Effektstärke verwenden.
- Unterschiedlichen Maße für die Effektstärke werden als **Effektmaße** bezeichnet.



Beispiel: Studie zum Wohlbefinden in Regionen mit und ohne Corona-Kontaktsperre

Unstandardisierte und standardisierte Effektstärken

- Mittelwertsdifferenzen, Kovarianzen und Regressionskoeffizienten sind **unstandardisierte Effektstärken**, weil sie in den Rohwerten der Messung vorliegen.

 **Beispiel** **Beispiel Mittelwertunterschied:** der durchschnittliche Größenunterschied von erwachsenen Männern und Frauen in Deutschland beträgt 16cm¹.

 **Beispiel** **Beispiel unstandardisierter Regressionskoeffizient:** je 0,1 Verbesserung in der **Abiturnote** steigt das monatliche Einstiegseinkommen um durchschnittlich 70 **Euro**².

- In den beiden genannten Beispielen haben die Effektstärken sinnvolle und interpretierbare Einheiten und wären vergleichbar zwischen Studien.
- Gerade in der Psychologie ist dies aber nicht immer gegeben:
 - Fragebögen: Punktzahlen hängen willkürlich vom Kodierungsschema und der Zahl der Items ab
 - Ratingsskalen: Ratingsskalen unterscheiden sich häufig (Wohlbefinden auf einer Skala von 0 bis 100%, Wohlbefinden auf einer Skala von -5 bis 5, usw.)
- Falls die Skala (z.B. der verwendete Fragebogen) neu oder wenig bekannt ist, wie soll dann der Effekt interpretiert werden? Wann kann er als groß und wann als klein gelten?

Standardisierte Effektstärken

- Um Effektstärken unabhängig von der verwendeten Skala zu vergleichen, werden Effektstärken **standardisiert**.
- Die Transformation der **Standardisierung** haben wir bereits bei Variablen kennengelernt (Teilen durch die Standardabweichung s_X) – sie lässt sich analog auch auf Effekte beziehen.
- Im Fall von Zusammenhangsanalysen haben wir die standardisierte Effektstärke bereits kennengelernt: der **Korrelationskoeffizient** (r, r_s, τ, ϕ). Beispiel Pearson-Korrelation:

$$r = \frac{Cov(X, Y)}{s_X s_Y}$$

- Der Nenner $s_X s_Y$ stellt hier die Standardisierung dar.
- Wir können sehen, dass r standardisiert ist, da es keine Einheit hat und eine vergleichbar ist zwischen verschiedenen Skalen und verschiedenen Variablen.



Alle Korrelationskoeffizienten (r, r_s, τ, ϕ) sind bereits Effektstärken.

Verrechnung von Varianzen und Standardabweichungen

Bevor wir uns der Standardisierung von Effektstärken bei Mittelwertsunterschieden widmen, lohnt es sich einen Blick darauf zu werfen, wie Varianzen verschiedener Bedingungen in einer Stichprobe (abhängige Messungen) oder verschiedener Stichproben (unabhängige Messungen) miteinander kombiniert werden können.

Abhängige Messungen in einer Stichprobe

Beispiel: Sie messen die Merkfähigkeit einer Stichprobe vor (Bedingung A) und nach (Bedingung B) einer Schlafphase.

In diesem Fall können wir nicht nur die Mittelwerte \bar{x}_A und \bar{x}_B auf Gruppenebene voneinander abziehen ($\Delta\bar{x} = \bar{x}_A - \bar{x}_B$), sondern bereits die einzelnen Messwerte $x_{A,i}$ und $x_{B,i}$ auf Versuchspersonenebene: $\Delta x_i = x_{A,i} - x_{B,i}$

Die Frage lautet hier: wie groß ist die Variabilität von Δx_i ?

Wir können diese Variabilität mit der gewohnten Formel berechnen, nur dass wir statt x_i die Δx_i einsetzen:

$$\hat{\sigma}_{\Delta}^2 = \frac{1}{n-1} \sum (\Delta x_i - \Delta \bar{x})^2 \quad \text{bzw.} \quad \hat{\sigma}_{\Delta} = \sqrt{\frac{1}{n-1} \sum (\Delta x_i - \Delta \bar{x})^2}$$

Verrechnung von Varianzen und Standardabweichungen

Mit einigen mathematischen Tricks lässt sich zeigen, dass die Formel für die Varianz der Differenzwerte auch wie folgt dargestellt werden kann:

$$\hat{\sigma}_{\Delta}^2 = \hat{\sigma}_A^2 + \hat{\sigma}_B^2 - 2 \hat{Cov}(X_A, X_B) \quad \text{bzw.}$$

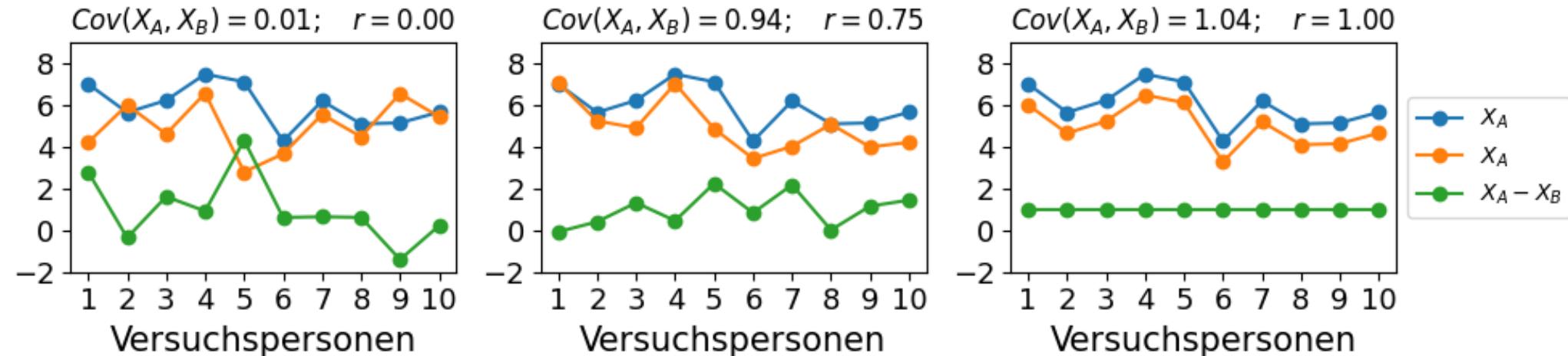
$$\hat{\sigma}_{\Delta} = \sqrt{\hat{\sigma}_A^2 + \hat{\sigma}_B^2 - 2 \hat{Cov}(X_A, X_B)}$$

Diese Formel macht transparent, was passiert, wenn wir zwei Zufallsvariablen voneinander abziehen:

- Sind die Zufallsvariablen *nicht korreliert* ($\hat{Cov}(X_A, X_B) = 0$), so ist die Varianz der Differenz einfach der Summe der Einzelvarianzen in den Bedingungen A und B.
- Sind die Zufallsvariablen *positiv korreliert* ($\hat{Cov}(X_A, X_B) > 0$), so reduziert sich diese die Summe in Abhängigkeit von der Kovarianz.
- Sind die Zufallsvariablen *negativ korreliert* ($\hat{Cov}(X_A, X_B) < 0$), so erhöht sich die Summe in Abhängigkeit von der Kovarianz.
(*Subtraktion* der negativen Kovarianz resultiert in einer Erhöhung – “minus x minus = plus”).

Verrechnung von Varianzen und Standardabweichungen

Der Effekt der Kovarianz auf die Varianz der Differenzwerte ist besonders intuitiv bei einer positiven Korrelation. Betrachten wir zwei Variablen X_A und X_B mit unterschiedlichen Kovarianzen und wie sich dabei jeweils die Variabilität des Differenzwertes entwickelt:



In allen drei Plots gilt $\bar{x}_A = 6$ und $\bar{x}_B = 5$, d.h. $\bar{x}_A - \bar{x}_B = 1$.

Wir sehen: je höher die Korrelation \bar{x}_A und \bar{x}_B , desto geringer die Variabilität der Differenz von $\bar{x}_A - \bar{x}_B$! Ist die Korrelation perfekt ($r = 1$), so ist die Differenz sogar konstant, d.h. ihre Variabilität ist gleich 0.

Verrechnung von Varianzen und Standardabweichungen

Unhängige Messungen in zwei Stichproben

Im Fall zweier unabhängiger Messungen, ist es nicht möglich die Einzelmesswerte $x_{A,i}$ und $x_{B,i}$ auf Versuchspersonenebene voneinander abzuziehen. Es wäre ohnehin völlig unklar, welche Versuchsperson man in Gruppe A mit welcher anderen Versuchsperson in Gruppe B matched.

Die Frage lautet hier: wie groß ist die mittlere Varianz beider Gruppen?

Da die Stichprobengrößen n_A und n_B unterschiedlich sein können, berechnen wir den *an den Stichprobengrößen gewichteten Varianzmittelwert*:

$$\hat{\sigma}_{\text{pooled}}^2 = \frac{(n_A - 1)\hat{\sigma}_A^2 + (n_B - 1)\hat{\sigma}_B^2}{n_A + n_B - 2} \quad \text{bzw.} \quad \hat{\sigma}_{\text{pooled}} = \sqrt{\frac{(n_A - 1)\hat{\sigma}_A^2 + (n_B - 1)\hat{\sigma}_B^2}{n_A + n_B - 2}}$$

$\hat{\sigma}_{\text{pooled}}^2$ wird als **gepoolte Varianz** bezeichnet. Die Subtraktion von 1 von den Stichprobengrößen ist Ausdruck der Besselkorrektur.

Wir halten fest: die gepoolte Varianz zweier unabhängiger Stichproben entspricht der “mittleren Varianz” beider Gruppen.

Herleitung der gepoolten Varianz und Standardabweichung

- Wir starten mit den (nicht bias-korrigierten) **Varianzen** s_A^2 und s_B^2 der Stichproben A und B.
- Die **gepoolte Varianz** ist definiert als der gewichtete Mittelwert der beiden Einzelvarianzen in den beiden Gruppen:

$$s_{\text{pooled}}^2 = \frac{n_A \cdot s_A^2 + n_B \cdot s_B^2}{n_A + n_B}$$

- Da wir von der Stichprobe auf die Population schließen wollen, wird die **Besselkorrektur** verwendet, d.h. $n_A \rightarrow n_A - 1$ und $n_B \rightarrow n_B - 1$ und $s^2 \rightarrow \hat{\sigma}^2$:

$$\hat{\sigma}_{\text{pooled}}^2 = \frac{(n_A - 1) \cdot \hat{\sigma}_A^2 + (n_B - 1) \cdot \hat{\sigma}_B^2}{n_A + n_B - 2}$$

- Analog gilt:

bzw. $\hat{\sigma}_{\text{pooled}} = \sqrt{\frac{(n_A - 1) \cdot \hat{\sigma}_A^2 + (n_B - 1) \cdot \hat{\sigma}_B^2}{n_A + n_B - 2}}$

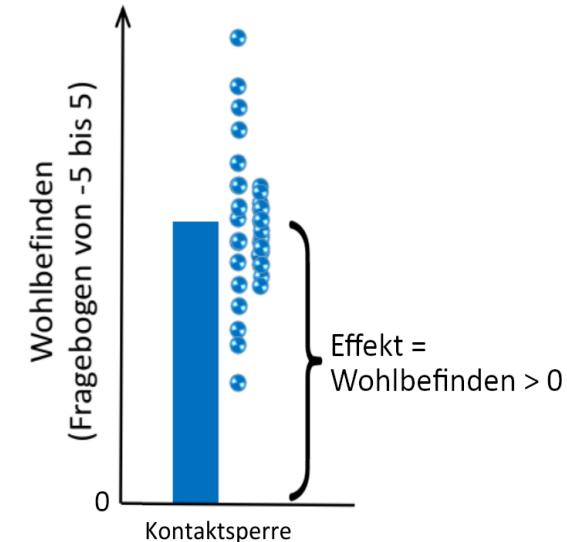


Standardisierte Effektstärken für Mittelwertunterschiede: Einzelmessung

Fall 1: Es gibt nur eine einzelne Messung an einer Gruppe und die Frage ist, ob der Mittelwert \bar{x} bedeutsam über einem Referenzwert μ_0 liegt.

$$\text{Mittelwertsunterschied} = \bar{x} - \mu_0$$

Standardisierte Effektstärke: $d = \frac{\bar{x} - \mu_0}{\hat{\sigma}}$



Beispiel: Studie zum Wohlbefinden – ist das Wohlbefinden *in der Gruppe* mit Kontaktsperrre noch im positiven Bereich ($\bar{x} > \mu_0$ mit $\mu_0 = 0$)?

- Die Verwendung von $\hat{\sigma}$ statt σ bringt zum Ausdruck, dass wir hier auf Basis der Stichprobe die Varianz in der Population schätzen (d.h. Besselkorrektur mit $n - 1$):

$$\hat{\sigma} = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n - 1}}$$

- Die Bezeichnung d für standardisierte Mittelwertsunterschiede stammt von dem Statistiker Jacob Cohen – häufig wird daher auch von **Cohen's d** gesprochen.

Standardisierte Effektstärken für Mittelwertunterschiede: abhängige Messungen

Fall 2: A und B sind abhängige Messungen in derselben Gruppe (d.h. A und B sind zwei Versuchszeitpunkte oder Versuchsbedingungen).

$$\text{Mittelwertsunterschied} = \bar{x}_A - \bar{x}_B$$

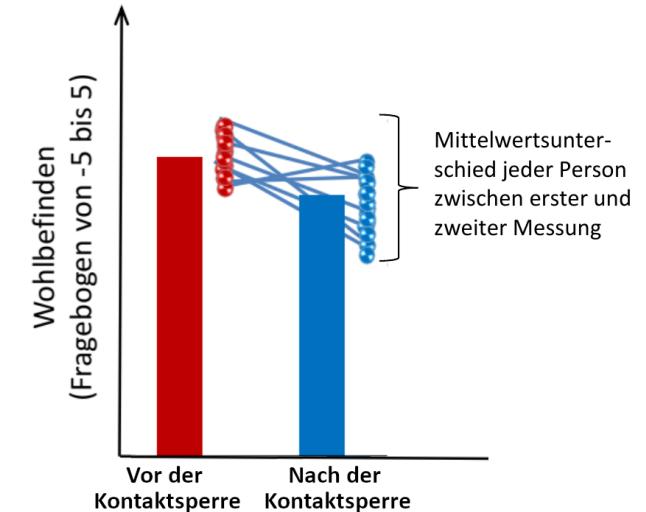
- Klassischerweise wird für die Berechnung von Cohen's d die Standardabweichung $\hat{\sigma}_\Delta$ der Differenz $\Delta X = X_A - X_B$ gewählt:

Standardisierte Effektstärke: $d = \frac{\bar{x}_A - \bar{x}_B}{\hat{\sigma}_\Delta}$

$$\text{mit } \hat{\sigma}_\Delta = \sqrt{\frac{\sum (\Delta x_i - \Delta \bar{x})^2}{n - 1}} \quad \text{und} \quad \Delta x_i = x_i^{(A)} - x_i^{(B)}$$



Handelt es sich um ein Pre-post-Interventionsdesign (Messung vor und nach einer Intervention an derselben Gruppe), bietet es sich eigentlich an, nur die Standardabweichung des Vortestes für die Standardisierung zu nehmen. Grund: die Variabilität vor einer Intervention ist am ehesten zwischen Studien vergleichbar.³ In der Praxis geschieht dies aber selten.



Beispiel: Unterscheidet sich das Wohlbefinden in derselben Gruppe vor und nach einer Kontaktsperrre?

Standardisierte Effektstärken für Mittelwertunterschiede: unabhängige Messungen

Fall 3: A und B sind **unabhängige Gruppen**, d.h. sie sind aus unterschiedlichen Versuchspersonen zusammengesetzt.

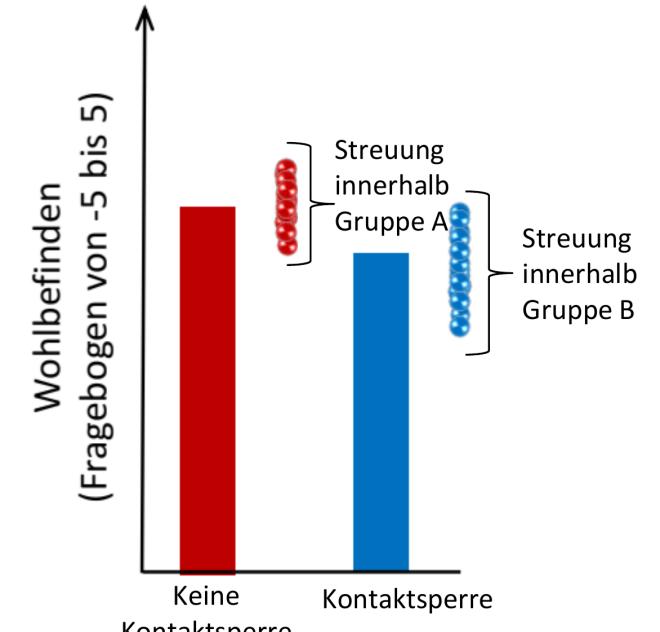
- Auch hier gilt: Mittelwertunterschiede werden immer mit der Streuung *innerhalb der Gruppen* standardisiert.
- Da es zwei Gruppen gibt, wird berechnet, wie die Datenpunkte beider Gruppen im Mittel *um ihren jeweiligen Mittelwerte \bar{x}_A und \bar{x}_B streuen*.
- Diese Art von Standardabweichung ist die bereits eingeführte **gepoolte Standardabweichung $\hat{\sigma}_{\text{pooled}}$** :

$$\hat{\sigma}_{\text{pooled}} = \sqrt{\frac{(n_A - 1)\hat{\sigma}_A^2 + (n_B - 1)\hat{\sigma}_B^2}{n_A + n_B - 2}}$$

Falls $n_A = n_B = n$:

$$\hat{\sigma}_{\text{pooled}} = \sqrt{\frac{\hat{\sigma}_A^2 + \hat{\sigma}_B^2}{2\left(1 - \frac{1}{n}\right)}}$$

Standardisierte Effektstärke: $d = \frac{\bar{x}_A - \bar{x}_B}{\hat{\sigma}_{\text{pooled}}}$



Unterscheidet sich der Mittelwert zweier Gruppen?

Interpretation von Cohen's d

- Die Werte von Cohen's d reichen von $-\infty$ bis $+\infty$. Negative Werte sind möglich, da das Vorzeichen davon abhängt, welcher Mittelwert von welchem abgezogen wird.
 - Es ist Konvention, d so zu berechnen, dass d positiv ausfällt, falls der Effekt in die hypothetisierte Richtung geht.
 - Die Interpretation der Effektstärke macht sich aber am absoluten Wert $|d|$ fest: wenn $d = -0,3$, dann hat der Effekt eine Stärke von $d = 0,3$, aber in eine andere Richtung als erwartet.
- Durch die Standardisierung mit der Standardabweichung gilt umgekehrt, dass Cohen's d ausdrückt, um wie viel Standardabweichungen ein Effekt von einem Nulleffekt abweicht. Diese Interpretation ist am intuitivsten für den Fall einer *Einzelmessung mit Referenzwert*:



Beispiel

Eine Studie untersucht, ob die Schlafdauer in einer Gruppe von Psychologiestudierenden in der Bachelorarbeitsphase geringer ist als die durchschnittliche Schlafdauer in Deutschland (7:45 Stunden). Tatsächlich zeigt sich eine verringerte Schlafdauer mit einem Cohen's d von 0,3.

- Die Effektstärke von 0,3 sagt aus, dass die Schlafdauer um 0,3 Standardabweichungen gegenüber dem Durchschnittswert verringert ist. Die Einheit *Standardabweichung* bezieht sich dabei auf die Standardabweichung $\hat{\sigma}$ der Schlafdauer in der untersuchten Gruppe.
 - Bei mehreren Bedingungen/Gruppen gilt zwar weiterhin die Interpretation im Sinne von *in Einheiten von Standardabweichungen*, allerdings ist die Definition der Standardabweichungen ($s_{\Delta}, s_{\text{pooled}}$) eher kompliziert und damit nicht mehr sonderlich intuitiv.

Interpretation von Effektstärken

- Um Effektstärken besser einordnen und kommunizieren zu können, hat Jacob Cohen folgende Unterteilung vorgeschlagen:

d	r	Interpretation
< 0.2	< 0.1	Trivialer Effekt
ab 0.2	ab 0.1	Kleiner Effekt
ab 0.5	ab 0.3	Mittlerer Effekt
ab 0.8	ab 0.5	Großer Effekt

- Zugleich fügt aber Cohen selbst folgende Qualifizierung an:

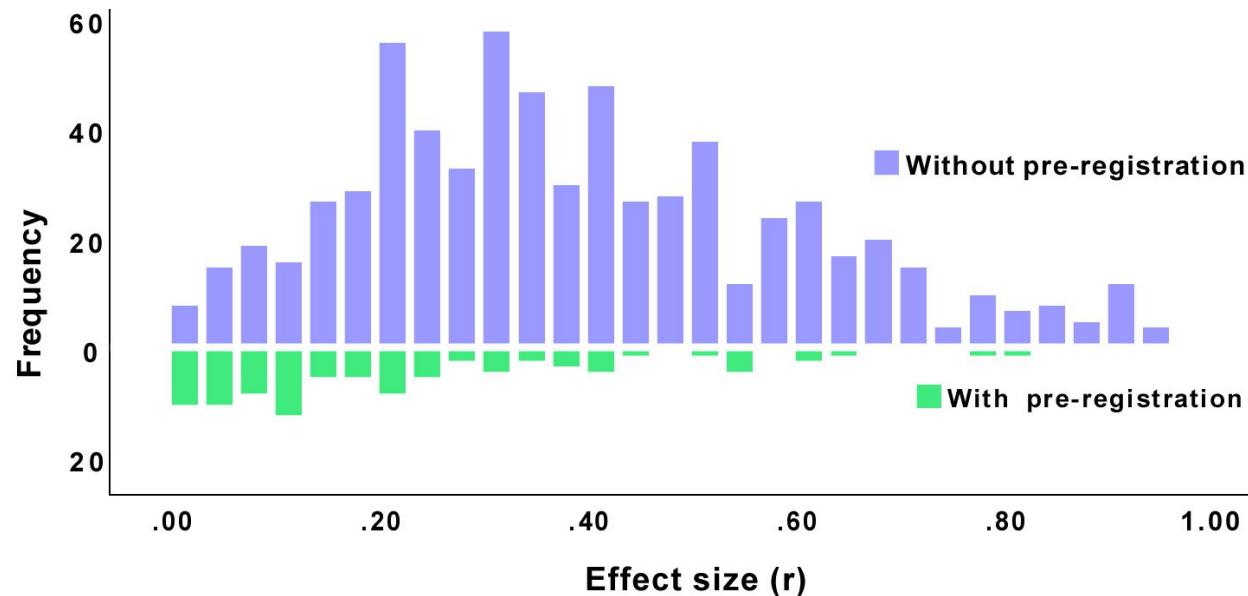


The terms „small,” “medium,” and „large” are relative, not only to each other, but to the area of behavioral science or even more particularly to the specific content and research method.
(Cohen, 1988, p. 25)

- Effektstärken sollten also idealerweise in ihrem jeweiligen Kontext interpretiert werden.
- Beispiel:** Effektstärken bezüglich der Veränderung des Körpergewichts durch Diäten sind erwartbar größer, als Veränderungen bei eher stabilen Merkmalen wie Persönlichkeitsfacetten. ⇒ Ein d-Wert der einer vergleichsweise geringen Veränderung des Körpergewichts entspricht, würde vielleicht in der Persönlichkeitsforschung als starker Effekt gelten.

Interpretation von Cohen's d

- Um zu beurteilen, was als kleine / mittlere / starke Effekte in einem spezifischen Kontext gilt, konsultiert man prinzipiell die entsprechende Fachliteratur nach typischen Referenzeffektstärken.
- Problem: es gibt gute Evidenz, dass *publizierte Effekte* die *wahren Effekte* überschätzen (**Publikationsbias**) ⇒ führt zu falschen Maßstäben



Die Abbildung zeigt, dass Effekte, bei denen Hypothesen und Analysen vorab registriert wurden ("with pre-registration") deutlich geringere Effektstärken aufweisen, als Effekte "without pre-registration". In der Abbildung ist zu beachten, dass die absolute und nicht die relative Häufigkeit aufgetragen ist, wodurch Studien ohne Preregistrierung — von denen es deutlich mehr gibt — visuell dominieren.⁴

- ⇒ Die Interpretation und Einordnung von Effektstärken ist ein nicht-triviales Problem, das viel "domain knowledge" erfordert. Dies gilt für standardisierte wie unstandardisierte Effektstärken.
- Faustregeln finden sich hier: <https://imaging.mrc-cbu.cam.ac.uk/statswiki/FAQ/effectSize>

Effektstärken für Mittelwertunterschiede bei mehr als zwei Messungen

- Gibt es mehr als zwei Experimentalbedingungen oder Gruppen (A, B, C, ...), gibt es zwei Möglichkeiten:
 - Von Interesse sind die **paarweisen** Mittelwertsunterschiede (z.B: $A - B, A - C, B - C$) → in diesem Fall kann wie bisher Cohen's d für jedes Paar angewendet werden.
 - Von Interesse ist, ob sich die Mittelwerte in den Gruppen A, B, C **in ihrer Gesamtheit betrachtet** unterscheiden, d.h. ob die Aufteilung in diese spezifischen Gruppen sinnvoll ist.

Fall 2 ist unser erster Kontakt mit der **Varianzanalyse**, die in Statistik 2 ausführlich behandelt wird. Man kann die Fragestellung in Fall 2 auch folgendermaßen formulieren:

Zu welchem Grad wird die Varianz der gepoolten Daten aller Gruppen bereits erklärt durch die Mittelwerte der Gruppen?

- Auf Basis dieser Formulierung ist nicht mehr überraschend, dass die Effektstärke für Mittelwertunterschiede von mehreren Messungen als *Verhältnis zweier Streuungen* ausgedrückt werden kann:

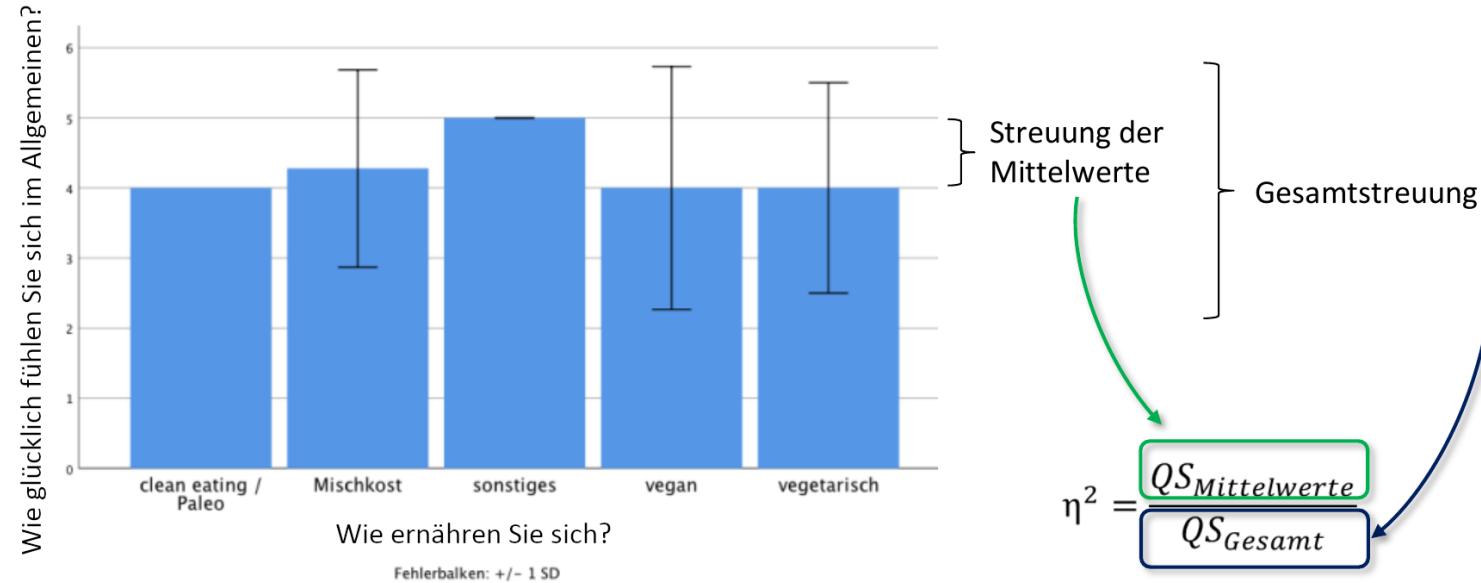
$$\eta^2 = \frac{QS_{\text{Mittelwerte}}}{QS_{\text{Gesamt}}}$$

$$\sigma^2 = \frac{\sum(x_i - \bar{x})^2}{N}$$

Quadratsumme (QS)

Die Quadratsumme entspricht dem Zähler in der Formel für die Varianz.

Effektstärken für Mittelwertunterschiede bei mehr als zwei Messungen



Courtesy of Prof. Thomas Schäfer, Medical School Berlin

- η^2 ("Eta Quadrat") gibt an, wie viel der Gesamtvarianz durch die Varianz der Mittelwerte aufgeklärt wird.
- Es kann zwischen 0 (Mittelwerte erklären keine Varianz) und 1 (Mittelwerte erklären die komplette Varianz) liegen.
- Die Berechnung von QS_{Gesamt} umfasst 1) die Varianz zwischen den Bedingungen, 2) die Varianz zwischen Personen (über Bedingungen hinweg), und 3) wie sehr sich die Varianz der Bedingungen zwischen Personen unterscheidet:

$$QS_{\text{Gesamt}} = QS_{\text{Bedingungen}} + QS_{\text{Personen}} + QS_{\text{Personen} \times \text{Bedingungen}}$$

Effektstärken für Mittelwertunterschiede bei mehr als zwei Messungen

- Es kann argumentiert werden, dass die Varianz, die lediglich die interindividuellen Unterschiede der Versuchspersonen (Varianzanteil 2 bzw. QS_{Personen}) charakterisiert, für die Effektstärke irrelevant ist und nicht zu QS_{Gesamt} gezählt werden sollte, d.h.:

$$QS_{\text{Gesamt}} = QS_{\text{Bedingungen}} + QS_{\text{Personen} \times \text{Bedingungen}}$$

- Wird die Varianz QS_{Personen} nicht berücksichtigt, spricht man vom **partiellen η_p^2** .
- Eine ausführliche Diskussion zum Pro und Kontra von η^2 vs. η_p^2 findet sich z.B. im Buch von Eid, Gollwitzer und Schmitt im Kapitel zur Varianzanalyse.



Warum werden bei der Berechnung von η^2 die Quadratsummen und nicht die Varianzen direkt verwendet? Grund ist, dass die Varianz die *durchschnittliche* (quadrierte) Abweichung vom Mittelwert angibt (daher der Faktor $\frac{1}{n}$), und beim Vergleich verschiedener Variabilitätskomponenten nicht festgestellt werden kann, wie viel der Datenvariabilität *absolut gesehen* durch eine Variabilitätskomponente erklärt wird.



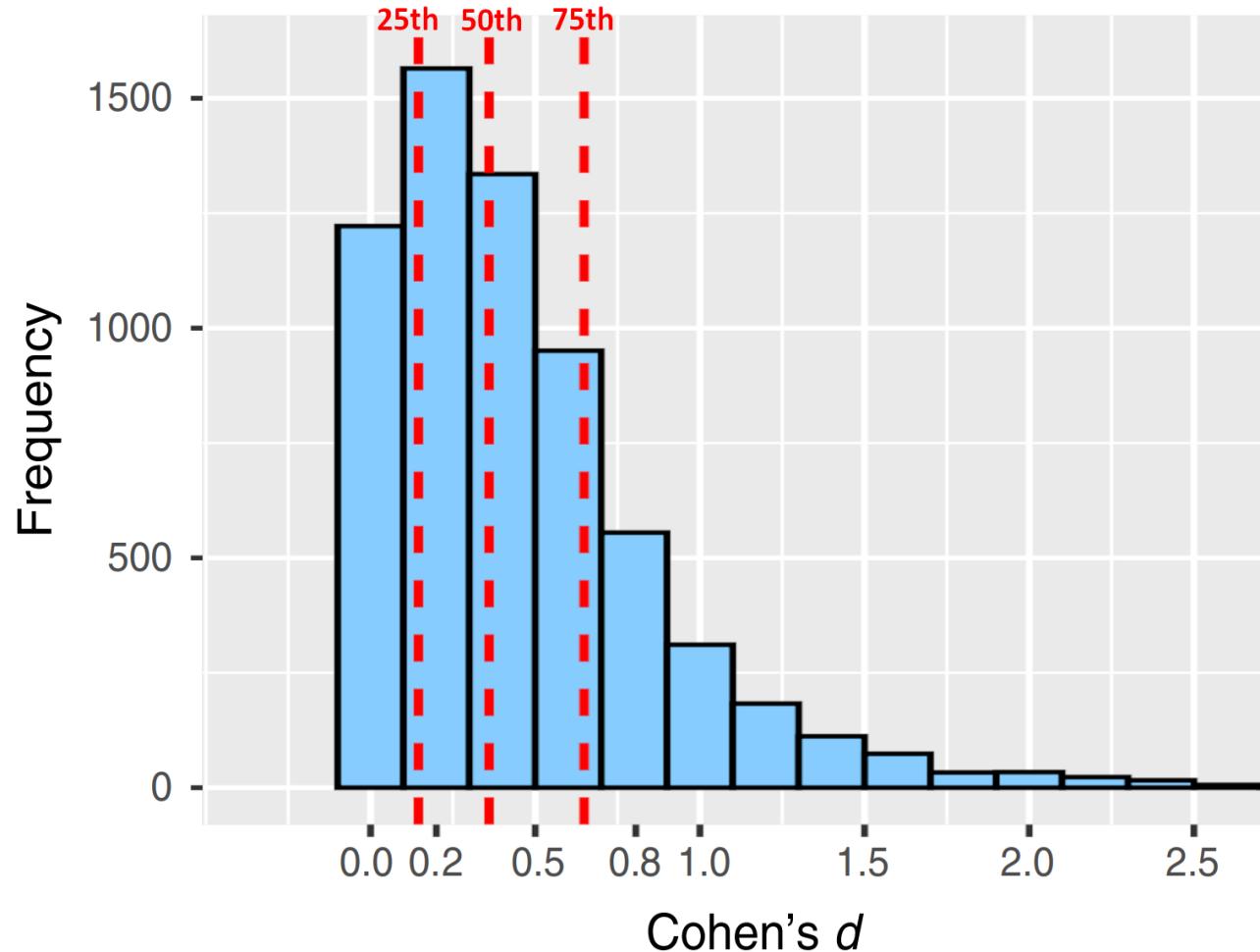
Typische Effektstärken in der Sozialpsychologie

Subgroup	Number of meta-analysis	Number of effect sizes	Median	Mean	SD
Correlation					
Groups	15	998	0.26	0.31	0.25
Interpersonal relationships	12	2,323	0.30	0.32	0.19
Prejudice	10	2,639	0.18	0.21	0.15
Self	10	1991	0.29	0.31	0.20
Attitude	9	2,352	0.26	0.29	0.20
Social cognition	9	1,248	0.27	0.33	0.24
Gender differences	5	585	0.23	0.27	0.20
Cohen's d					
Gender differences	12	1,261	0.22	0.30	0.31
Prejudice	10	1,370	0.34	0.44	0.40
Self	10	884	0.48	0.59	0.56
Interpersonal relationships	9	1,075	0.28	0.39	0.41
Social cognition	9	750	0.50	0.58	0.52
Attitude	5	428	0.39	0.47	0.38

Typische Effektstärke in der sozialpsychologischen Literatur.⁵



Tatsächliche Verteilung von Effektstärken (Sozialpsychologie)



Trotz Publikationsbias sind die tatsächlich in der Literatur berichteten Effektstärken geringer als bei der Einteilung nach Cohen angenommen. Auf Basis dieser Studie wären die korrekten Grenzen für Cohen's d 0,15, 0,36, und 0,65. Bild adaptiert von Lovakov & Agadulina (2021)⁶.



<https://rpsychologist.com/d3/cohend/>

R ← PSYCHOLOGIST

xEA English ▾ About Posts Visualizations     

Interpreting Cohen's d Effect Size

An Interactive Visualization

Created by [Kristoffer Magnusson](#)

Share  

The Cohen's d effect size is immensely popular in psychology. However, its interpretation is not straightforward and researchers often use general guidelines, such as small (0.2), medium (0.5) and large (0.8) when interpreting an effect. Moreover, in many cases it is questionable whether the standardized mean difference is more interpretable than the unstandardized mean difference.

In order to aid the interpretation of Cohen's d , this visualization offers these different representations of Cohen's d : visual overlap, Cohen's U_3 , the probability of superiority, percentage of overlap, and the number needed to treat. It also lets you change the standard deviation and displays the

Standardisierte oder unstandardisierte Effektstärken?

- Die Frage ob Effektstärken in standardisierter oder unstandardisierter Form berichtet werden sollten wird durchaus kontrovers diskutiert^{7 8}.
- Standardisierte Effektstärken ermöglichen eine bessere Vergleichbarkeit zwischen unterschiedlichen Skalen, gleichzeitig wird die intuitive Bedeutung von *Effektstärke* aber verwässert.



Beispiel Es wird berichtet, dass ein Coronaimpfstoff die Viruslast bei einer Infektion reduziert. Die Effektstärke wird mit Cohen's $d = 0.4$ angegeben.

- Aus diesem Beispiel wird klar, dass die Effektstärke zum einen wenig intuitiv ist (in jedem Fall für Nicht-Wissenschaftler:innen), zum anderen ist nicht ersichtlich, ob die Viruslast nennenswert reduziert wurde oder ob der Rückgang eher klein war, aber die Standardabweichung in der Gruppe so gering, dass dennoch ein hoher d -Wert erreicht wurde.
- Darüber hinaus hängt die Standardabweichung einer Variable häufig mit eher nebensächlichen Details eines experimentellen Designs (within-subject vs. between-subject) oder einer Stichprobe (nur Psychologiestudierende oder heterogeneres Sample der Allgemeinbevölkerung?) zusammenhängt, die für die Effektstärke wenig relevant sind.
- Aus diesen Gründen sollte für **Effekte, die in interpretierbaren Einheiten vorliegen, immer (auch) die unstandardisierte Effektstärke** angegeben werden (z.B. Notenstufen, Einkommen, IQ-Punkte, Größe- Gewichtsangaben, Zeitangaben).

Umrechnung von Cohen's d und Korrelationskoeffizient r

Fassen Metaanalysen sowohl Studien zusammen, die Effekte als Korrelation berichten (Effektmaß r), als auch Studien, die Effekte als Mittelwertsunterschiede zwischen Bedingungen/Gruppen berichten (Effektmaß d), entsteht die Notwendigkeit r und d ineinander umzurechnen.

Fall 1: eine der Variablen in der Korrelation ist eine natürliche binäre Variable (z.B. männl./weibl.)

In diesem Fall gilt:

$$d = \frac{2r}{\sqrt{1 - r^2}} \quad \hat{se}(d) = \frac{2}{\sqrt{(n - 1)(1 - r^2)}}$$

wobei $\hat{se}(d)$ der Standardfehler der Effektstärke d ist, der zusätzlich angegeben werden sollte.

Umgekehrt gilt:

$$r = \frac{d}{\sqrt{d^2 + 4}}$$



Umrechnung von Cohen's d und Korrelationskoeffizient r

Fall 2: beide Variablen in der Korrelation sind kontinuierlich, oder eine Variable ist binär, aber entstanden durch Dichotomisierung einer kontinuierlichen Variable (advanced!)

In diesem Fall muss eine Variable als unabhängige Variable X definiert werden. Falls eine der Variablen durch Dichotomisierung binär ist, ist diese Variable in jedem Fall die unabhängige Variable.

Es gilt:

$$d = \frac{kr}{\hat{\sigma}_X \sqrt{1 - r^2}} \quad \hat{se}(d) = |d| \sqrt{\frac{1}{r^2(n-3)} + \frac{1}{2(n-1)}}$$

wobei $\hat{se}(d)$ der Standardfehler der Effektstärke d ist, der zusätzlich angegeben werden sollte.

Interpretation: d entspricht der durchschnittlichen Zunahme der standarisierten Y -Variable mit jeder Zunahme von X um k (Rohwert)Einheiten – k muss vom Forscher gewählt werden.

Hier wird klar, dass die Formel in Fall 1 implizit $k = 2\hat{\sigma}_X$ annimmt. Wählt man dieses k auch in Fall 2, ist die Berechnung von d identisch zu Fall 1 – der Standardfehler unterscheidet sich allerdings weiterhin! Quelle:⁹

Umgekehrt gilt:

$$r = \frac{d\hat{\sigma}_X}{\sqrt{d^2\hat{\sigma}_X^2 + k^2}}$$



Absolute Risikoreduktion (ARR)

Sind beide Variablen dichotom, sind weder der Korrelationskoeffizient noch Cohen's d intuitive Effektmaße.



Beispiel Sie untersuchen, ob ein neues Medikament die Heilungsrate (Erfolgsrate) einer Krankheit verbessert. Die Treatmentgruppe erhält das Medikament, die Kontrollgruppe Placebo.

Eine sinnvolles Effektmaß ist hier, *um wie viel* die Erfolgsrate in der Treatmentgruppe die Erfolgsrate in der Kontrollgruppe übersteigt. Dies lässt sich einfach aus einer Vierfeldertafel ableiten:

	Geheilt	Nicht Geheilt
Treatment	A	B
Kontrolle	C	D

$$ARR = \frac{\frac{A}{A + B} - \frac{C}{C + D}}{\frac{A}{A + B}}$$

Erfolgsrate in der Treatmentgruppe

Erfolgsrate in der Kontrollgruppe

- ARR ist die **Absolute Risikoreduktion**.

Numbers Needed to Treat (NNT)

- Noch gängiger als die Absolute Risikoreduktion ist die inverse Größe, die als **Numbers Needed to Treat (NNT)** bezeichnet wird.

Definition **Number Needed to Treat:** Anzahl der Personen, die behandelt werden müssten, damit eine zusätzliche Person einen Nutzen hat.

Mathematisch ist *NNT* das Inverse der *ARR*:

$$NNT = \frac{1}{ARR}$$

- Da es um Personen geht, wird die NNT immer aufgerundet.

Beispiel

	Geheilt	Nicht Geheilt
Treatment	90	10
Kontrolle	35	35

$$ARR = \frac{90}{90+10} - \frac{35}{35+35} = 0,9 - 0,5 = 0,4 \rightarrow NRR = \frac{1}{0,4} = 2,5$$

⇒ Drei weitere Personen müssten behandelt werden, damit eine zusätzliche Person einen Nutzen hat (d.h. die andernfalls nicht geheilt würde).

Numbers Needed to Treat (NNT)

- Auch wenn das Ziel von *NNT* eine einfache laienverständliche Kommunikation der Treatmenteffizienz ist, darf angezweifelt werden, ob dies immer der Fall ist, wie durch folgendes Beispiel demonstriert¹⁰:

Consider a situation in which drug versus placebo response rates are 12% versus 1%, respectively; the advantage for the drug is 11%, and the NNT is 9. Consider another situation in which the drug versus placebo response rates are 99% versus 88%, respectively; the NNT is again 9. These 2 situations are strikingly different. In the first situation, there is almost no placebo response, and medication is associated with a relatively large treatment gain. In the second situation, there is a large placebo response, and medication is associated with a relatively small treatment gain. Yet, the NNT is the same in the 2 situations. So, it is really important for clinicians to know not only what the unique contribution of the drug is (NNT) but also what the placebo response and nonresponse rates are.

Odds Ratio (OR)

- Das **Odds Ratio (OR)** vergleicht das *Heilerfolgsverhältnis in der Treatmentgruppe* zum *Heilerfolgsverhältnis in der Kontrollgruppe*:

	Geheilt	Nicht Geheilt
Treatment	A	B
Kontrolle	C	D

$$\text{OR} = \frac{\frac{A}{B}}{\frac{C}{D}} = \frac{A \cdot C}{B \cdot D}$$

Beachte: Als Heilerfolgsverhältnis wird hier das Verhältnis der Zahl der geheilten Patienten gegenüber der Zahl der nicht geheilten Patienten verstanden.

- Hat das Treatment keine Auswirkung, so sind die Heilerfolgsverhältnisse in beiden Gruppen gleich, d.h. $OR = 1$.
- Ist das Treatment erfolgreich, ist das Heilerfolgsverhältnis in der Treatmentgruppen höher als in der Kontrollgruppe, d.h. $OR > 1$.
- Ist das Treatment sogar nachteilig, ist das Heilerfolgsverhältnis in der Treatmentgruppen *kleiner* als in der Kontrollgruppe, d.h. $OR < 1$.



Als kleine Übung kann man das Odds Ratio für die beiden hypothetischen Beispiele auf der vorherigen Folie berechnen.
Spoiler: das Odds Ratio ist für beide Fälle gleich ($OR = 13.5$)!

Übersicht Effektmaße

Effekte	
Unterschiede	Zusammenhänge
2 unabhängige Messungen d	Intervalldaten r
2 abhängige Messungen d	Ordinaldaten ρ / τ
mehr als 2 unabhängige Messungen η^2	Vierfeldertafel φ
mehr als 2 abhängige Messungen η^2	Vierfeldertafel für Erfolg/Risiko NNT / OR

Referenz

Empfehlungen nach Lakens (2013)¹¹:

d Familie

ES	Standardizer	Use
Cohen's d_{pop}	σ (population)	Independent groups, use in power analyses when population σ is known, σ calculated with n
Cohen's d_s	Pooled SD	Independent groups, use in power analyses when population σ is unknown, σ calculated with $n-1$
Hedges' g	Pooled SD	Independent groups, corrects for bias in small samples, report for use in meta-analyses
Glass's Δ	SD pre measurement or control condition	Independent groups, use when experimental manipulation might affect the SD
Hedges' g_{av}	$(SD_1 + SD_2)/2$	Correlated groups, report for use in meta-analyses (generally recommended over Hedges' g_{rm})
Hedges' g_{rm}	SD difference scores corrected for correlation	Correlated groups, report for use in meta-analyses (more conservative than Hedges' g_{av})
Cohen's d_z	SD difference scores	Correlated groups, use in power analyses

r Familie

ES (Biased)	ES (Less Biased)	Use
eta squared (μ^2)	omega squared (ω^2)	Use for comparisons of effects within a single study
eta squared (μ_p^2)	omega squared (ω_p^2)	Use in power analyses, and for comparisons of effect sizes across studies with the same experimental design.
Generalized eta squared (μ_G^2)	Generalized omega squared (ω_G^2)	Use in meta-analyses to compare across experimental designs



Standardisierte Effektstärken für Mittelwertunterschiede: aktuelle Forschung

Ein Problem der vorgestellten Effektmaße für abhängige Messungen, ist, dass sie nicht vergleichbar sind mit Effektmaßen für unabhängige Messungen. Es sind streng genommen **unterschiedliche Skalen**.

Eine aktuelle Forschungsarbeit¹² bietet dafür eine Lösung mit folgender (komplizierterer) Formel:

$$d = \sqrt{\frac{2(1-r)}{n}} \cdot t'_\nu(\lambda) \quad \text{mit} \quad \lambda = \sqrt{\frac{n}{2(1-r)}} \cdot \frac{\bar{x}_A - \bar{x}_B}{\hat{\sigma}}$$

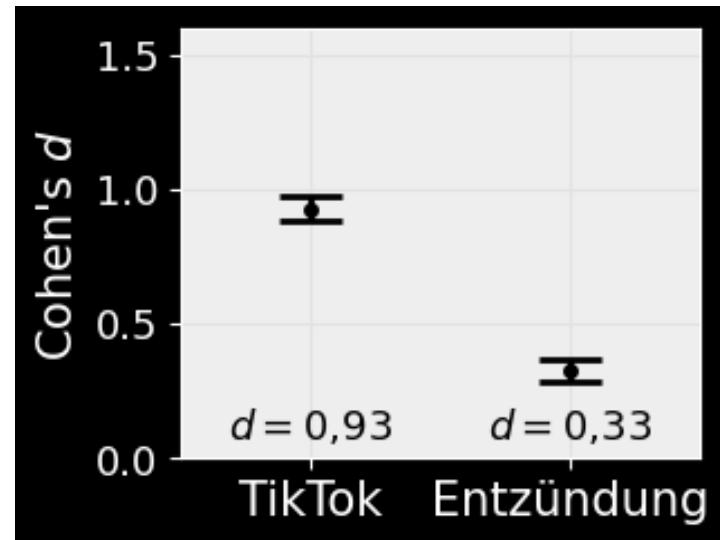
wobei r die Pearson-Korrelation zwischen den beiden abhängigen Messungen ist und t' die nichtzentrale t -Verteilung mit $\nu = 2(n-1)/(1+r^2)$ Freiheitsgraden. Zu beachten ist, dass die Formel nur gilt, falls gleiche Standardabweichungen in den beiden Bedingungen angenommen werden können ($\sigma = \sigma_A = \sigma_B$).

Laut dem Autor der Forschungsarbeit ist diese Formel sowohl auf abhängige als auch unabhängige Messungen anwendbar.

Weiterer nützlicher Link: ¹³



Sie berechnen nun Cohen's d für die beiden Gruppenunterschiede hinsichtlich der TikTok-Zeit und Entzündungswerte:



Hinweis: in der Abbildung wurde nicht nur die Effektstärke selbst aufgetragen, sondern auch ein Streuungsmaß der Effektstärke (Standardfehler der Effektstärke). Dazu kommen wir in Vorlesung 08.

Langsam schärft sich das Bild: während die Entzündungswerte höchstens eine mittlere Effektstärke aufweisen ($d = 0,33$), ist der TikTok-Effekt beeindruckend groß: $d = 0,93$.

Fußnoten

1. <https://www.laenderdaten.info/durchschnittliche-koerpergroessen.php>
2. Arbeitsberichte Dresdner Soziologie Nr. 21, <https://tud.qucosa.de/api/qucosa%3A24622/attachment/ATT-0/>
3.
Cumming G (2013) Cohen's d needs to be readily interpretable: Comment on Shieh (2013). *Behav Res* 45:968–971.
4.
Schäfer T, Schwarz MA (2019) The Meaningfulness of Effect Sizes in Psychological Research: Differences Between Sub-Disciplines and the Impact of Potential Biases. *Frontiers in Psychology* 10 Available at:
<https://www.frontiersin.org/articles/10.3389/fpsyg.2019.00813>.
5.
Lovakov A, Agadullina ER (2021) Empirically derived guidelines for effect size interpretation in social psychology. *Eur J Soc Psychol* 51:485–504.
6.
Lovakov A, Agadullina ER (2021) Empirically derived guidelines for effect size interpretation in social psychology. *Eur J Soc Psychol* 51:485–504.
7. <https://twitter.com/ceptional/status/1687577019629142017>
8.
Baguley T (2009) Standardized or simple effect size: What should be reported? *British Journal of Psychology*