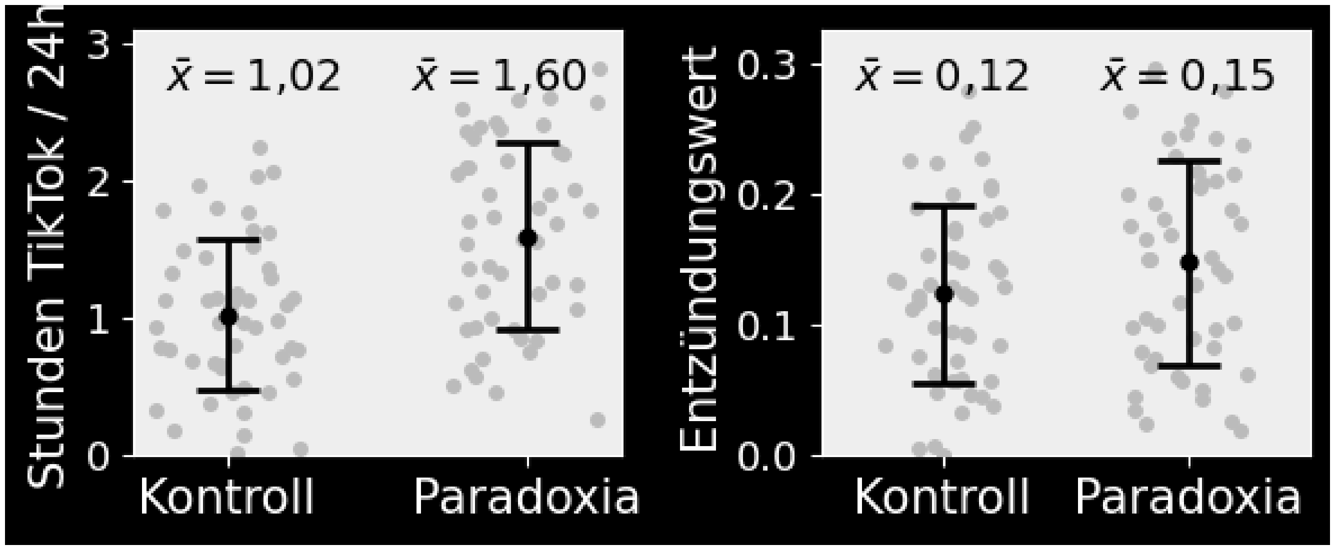
M24 Statistik 1: Wintersemester 2024 / 2025

Vorlesung 11: Effektstärke

Prof. Matthias Guggenmos

Health and Medical University Potsdam

Sie sinnieren weiterhin 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.

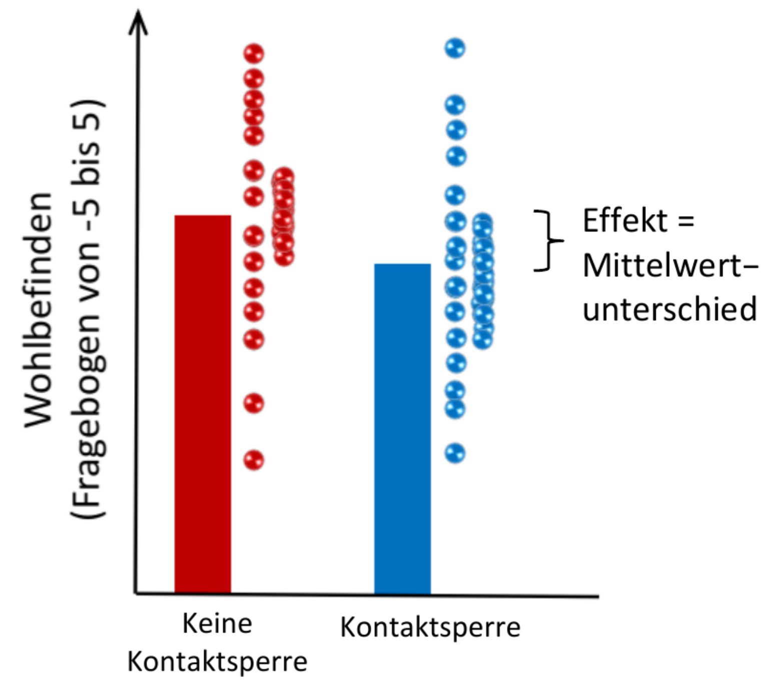


Die Frage lautet: wie kann man die beiden Gruppenunterschiede bezüglich TikTok-Zeit und Entzündungsparametern vergleichbar machen? Wie können wir eine Aussage treffen, welcher der beiden Effekte <u>stärker</u> ist?

Effektstärke

Effektstärke

- In psycholgischer 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?
  - $\Rightarrow$ **statistische Signifikanz** (ab Vorlesung 10): kann ein Effekt *allein durch Zufall erklärt werden*?
  - $\Rightarrow$ **praktische Signifikanz** (Effektstärken): ist die Stärke des Effektes *in der Praxis bedeutsam*?



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

- 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.

Unstandardisierte 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 16**cm**[1].

Beispiel

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

- In beiden 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
  - Ratingskalen: Ratingskalen 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 Zufallsvariablen kennengelernt: *Teilen durch die Standardabweichung*.
  - Dieser Ansatz lässt sich analog auch auf Effekte beziehen
  - Ein Beispiel für einen Effekt ist die Mittelwertdifferenz — hier gilt:
    standardisierte Effektstärke = Mittelwertdifferenz / Standardabweichung
  - Besteht der Effekt aus einem Produkt zweier Variablen (wie bei der Kovarianz), muss der Effekt durch die Standardabweichung beider Variablen geteilt werden, um die Einheiten herauszukürzen.
- In der Folge sind standardisierte Effekte einheitslos, da die zugrundegelegten Standardabweichungen die gleiche Einheit wie die Effekte haben.
- Zu beachten ist, dass das Teilen durch die Standardabweichung der resultierenden Effektstärke eine spezifische Interpretation zuschreibt: *ein Effekt wird als "stärker" gewertet, wenn der unstandardisierte Effekt groß ist im Vergleich zur zugrundegelegten Standardabweichung*.
- Es gibt zwei wesentliche Funktionen / Einsatzzwecke von Effektstärken:
  - Einordnung der Stärke eines Effektes in einer *einzelnen Studie*
  - Vergleichbarmachung von Effekten zwischen *verschiedenen Studien*

Mittelwertunterschiede

Cohen's d

- Es werden drei Fälle von Mittelwertunterschieden unterschieden:
  - Fall 1: eine Stichprobe + Einzelmessung: Differenz zwischen dem Mittelwert einer Messung und einem Referenzwert (z.B. IQ=100).
  - Fall 2: eine Stichprobe + abhängige Messungen: Differenz der Mittelwerte zweier Messungen in derselben Gruppe (z.B. IQ morgens und IQ abends).
  - Fall 3: zwei Stichproben + unabhängige Messungen: Differenz der Gruppenmittelwerte (z.B. IQ in einer blonden versus brünetten Gruppe).
- Die standardisierte Effektstärke für alle drei Fälle eines Mittelwertunterschieds berechnet sich als

$$d = \frac{\text{Mittelwertdifferenz}}{\text{Standardabweichung}}$$

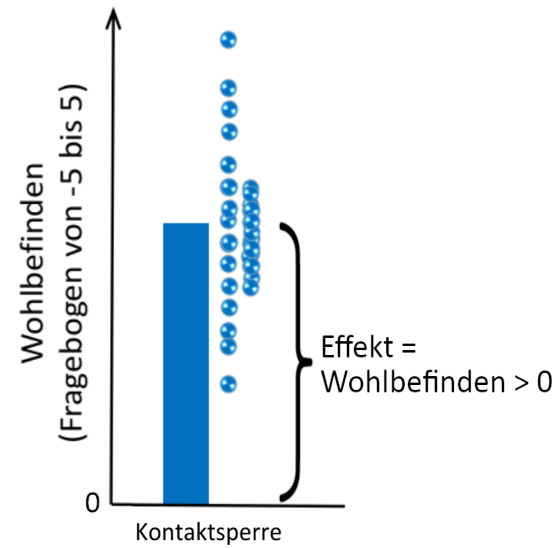
- Die Bezeichnung d für standardisierte Mittelwertunterschiede stammt von dem Statistiker Jacob Cohen — häufig wird daher auch von **Cohen's d** gesprochen.
- Während die Mittelwertdifferenz eindeutig ist, ist die weniger triviale Frage: Standardabweichung *von was*?

Fall 1: eine Stichprobe + Einzelmessung

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.

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

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



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

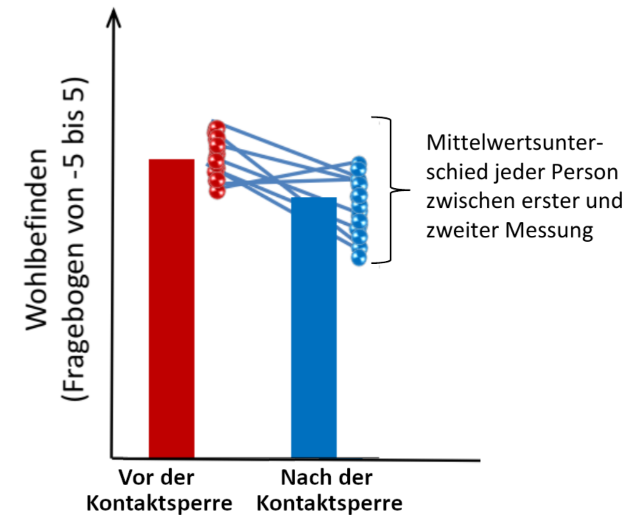
- Die Wahl der Standardabweichung bereitet hier keine Kopfzerbrechen — es ist schlicht die Standardabweichung der Variable X in der Stichprobe.
- Die Formel für die Standardabweichung ist die bekannte Formel mit Besselkorrketur:

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

Fall 2: eine Stichprobe + zwei abhängige Messungen

Es wird die Mittelwertdifferenz $\bar{x}_A - \bar{x}_B$ zweier abhängiger Messungen in einer Stichprobe berechnet und die Frage ist, ob diese Differenz bedeutsam ist.

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



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

Standardisierte Effektstärke:

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

- $\hat{\sigma}_\Delta$ ist die Standardabweichung der Differenzenvariable $\Delta X = X_A - X_B$:

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

- Wir bezeichnen diese als **Differenzenvarianz**.

Fall 2: andere Darstellung der Differenzenvarianz

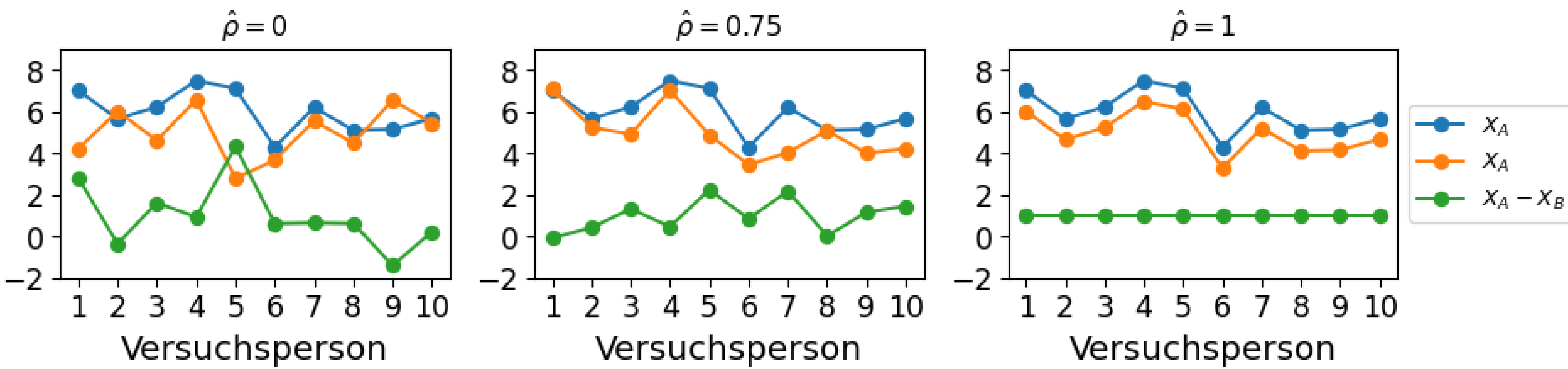
- Mit einigen mathematischen Tricks lässt sich zeigen, dass die Standardabweichung der Differenzwerte auch wie folgt dargestellt werden kann:

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

- $\hat{\rho}$ ist der Pearson-Korrelationskoeffizient zwischen den Messungen X_A und X_B.
- Diese Formel macht transparent, dass die Differenzenvarianz vom Zusammenhang zwischen X_A und X_B abhängt:
  - Sind die Zufallsvariablen *nicht korreliert* ($\hat{\rho} = 0$), so ist die Varianz der Differenz einfach der Summe der Einzelvarianzen.
  - Sind die Zufallsvariablen *positiv korreliert* ($\hat{\rho} > 0$), so reduziert sich die Summe in Abhängigkeit von der Kovarianz.
  - Sind die Zufallsvariablen *negativ korreliert* ($\hat{\rho} < 0$), so erhöht sich die Summe in Abhängigkeit von der Kovarianz. (*Subtraktion* der *negativen* Kovarianz resultiert in einer Erhöhung — “minus x minus = plus”).

Fall 2: andere Darstellung der Differenzenvarianz

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. $\bar{x}_A - \bar{x}_B = 1$.

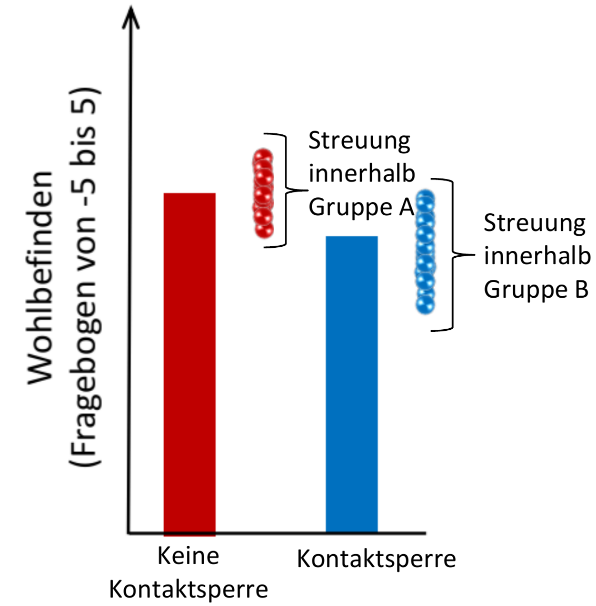
Wir sehen: je höher die Korrelation $\bar{x}_A$ und $\bar{x}_B$, desto geringer die Variabilität der Differenz von $\bar{x}_A - \bar{x}_B$! Ist die Korrelation perfekt ($\hat{\rho} = 1$), so ist die Differenz sogar konstant, d.h. ihre Variabilität ist gleich 0.

Fall 3: zwei Stichproben + unabhängige Messungen

Es wird die Mittelwertdifferenz $\bar{x}_A - \bar{x}_B$ zweier unabhängiger Gruppen berechnet und die Frage ist, ob diese Differenz bedeutsam ist.

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

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



Unterscheidet sich der Mittelwert zweier Gruppen?

- Im Nenner von Cohen's d wird hier die **gepoolte Varianz** verwendet (Details siehe Bonuscontent Gepoolte Varianz am Ende der Folien).
- Es handelt sich dabei um die mittlere Varianz über beide Gruppen hinweg, entsprechend einem an den Stichprobengrößen n_A und n_B gewichteten Mittelwert:

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

Fall 3: zwei Stichproben + unabhängige Messungen

- Streng genommen gilt die angebene Formel für Cohen's bei unabhängigen Stichproben nur, wenn die Gruppenvarianzen $\hat{\sigma}_A$ und $\hat{\sigma}_B$ ähnlich sind.
  - Gängige Faustregel: Varianzen sollen sich höchstens um Faktor 2 unterscheiden: $0.5 < \frac{\sigma_A^2}{\sigma_B^2} < 2$
- Da sich jedoch bislang kein Verfahren für den den Gruppenvergleich mit unähnlichen Varianzen etabliert hat, behandeln wir diesen Fall im Rahmen von Statistik 1 nicht.
  - Bei Interesse finden Sie eine mögliche Formel für diesen Fall im Bonuscontent Cohen's d bei unterschiedlichen Gruppenvarianzen.

Interpretation

Interpretation von Cohen’s d: Vorzeichen

- Die Werte von Cohen’s d reichen von $-\infty$ bis $+\infty$. Negative Werte sind möglich, da das Vorzeichen davon abhängt, ob bei der Mittewertdifferenz im Zähler Mittelwert A von Mittelwert B abgezogen wird, oder umgekehrt.
  - Für die Interpretation der Effektstärke ist ausschlaggebend, was mit “Effekt” gemeint ist “wie rum” die Mittelwertdifferenz gebildet wurde.
  - Es ist prinzipiell eine sinnvolle Konvention, d so zu berechnen, dass d positiv ausfällt, falls der Effekt in die hypothesierte Richtung geht.

Beispiel

Eine Studie untersucht den “Albtraum-Effekt”, also die Hypothese, dass Albträume die Konzentrationsfähigkeit X am nächsten Tag in Mitleidenschaft ziehen. Es zeigt sich, dass

$$d = \frac{\bar{x}_{\text{Albtraum}} - \bar{x}_{\text{Keine Albtraum}}}{\hat{\sigma}_{\Delta}} = -0.8$$

Das Ergebnis zeigt, dass die Konzentrationsfähikeit tatsächlich geringer nach Albtraum-Nächten ist. Cohen’s d ist hier negativ, allerdings entspricht das Vorzeichen der Richtung der Hypothese und der “Albtraum-Effekt” würde als ein starker Effekt entsprechend $d = 0.8$ gewertet werden (siehe übernächste Folie).

Hier wäre es ggf. sinnvoll, die Mittelwertdifferenz direkt so zu bilden, dass d positiv ist, wenn die Hypothese zutrifft.

Interpretation von Cohen’s d: Magnitude / Stärke

- Cohen’s d drückt aus, **um wie viel Standardabweichungen ein Effekt von einem Nulleffekt abweicht**.
- Auch hier ist die Frage “Standardabweichungen von *Was*?”
- Die 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.

- Die Interpretation der Effektstärke bei Mittelwertdifferenzen *zweier* abhängiger oder unabhängiger Messungen ist im Prinzip identisch, jedoch ist die Frage “Standardabweichungen von *Was*?” weniger intuitiv zu beantworten (→ $\hat{\sigma}_\Delta$, $\hat{\sigma}_{\text{pooled}}$).

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

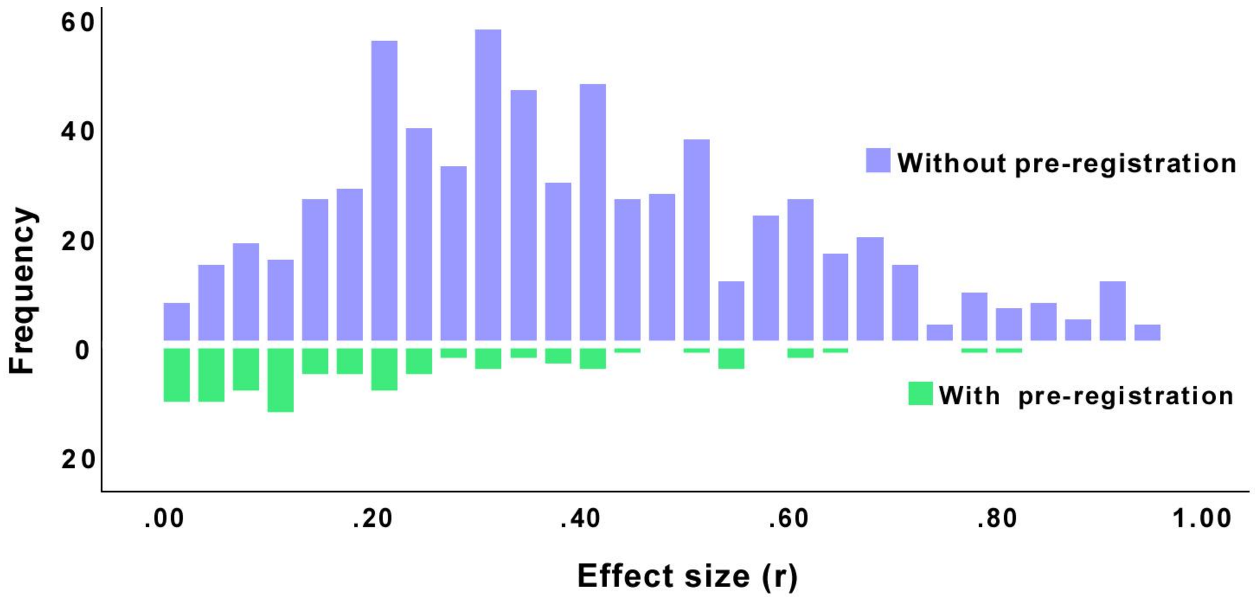
- Zugleich fügt aber Cohen selbst folgende Qualifizierung an:

> 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. $\Rightarrow$ 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 die Verteilung von Effekten bei präregistrierten Studien (grün) deutlich zu geringeren Effektstärken hin verschoben ist, verglichen mit Studien ohne Präregistrierung (lila).[3]

⇒ 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

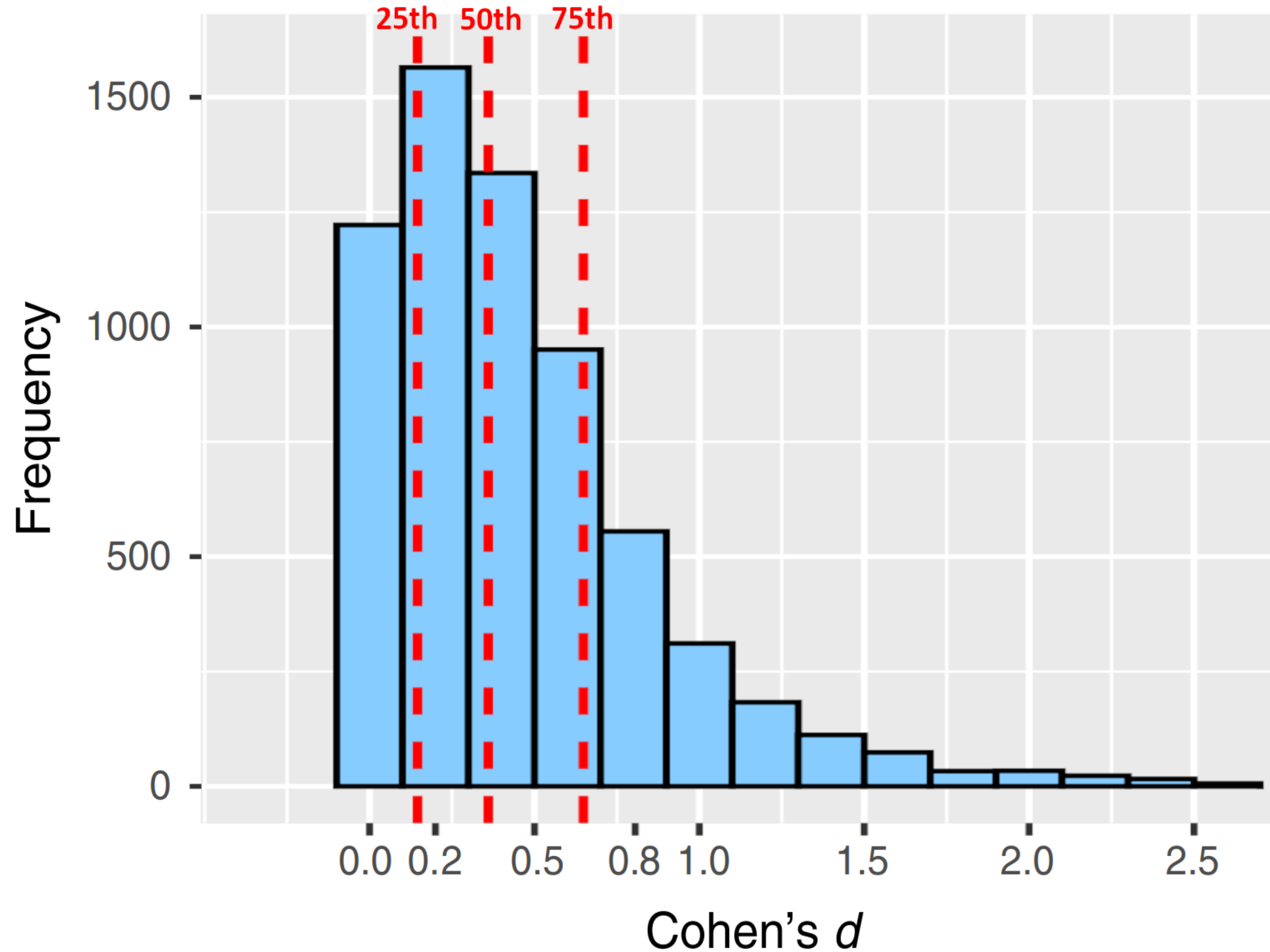
Typische Effektstärken in der Sozialpsychologie

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[4].

KEIN KLAUSUR STOFF

Typische Effektstärken in der 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)[5].

KEIN KLAUSUR STOFF

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

R <- PSYCHOLOGIST

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?

- Ob Effektstärken in standardisierter oder unstandardisierter Form berichtet werden sollten wird kontrovers diskutiert[6] [7].
- Standardisierte Effektstärken ermöglichen eine bessere Vergleichbarkeit zwischen unterschiedlichen Skalen, gleichzeitig wird die intuitive Bedeutung von *Effektstärke* aber verwässert.

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

- Auf Basis der standardisierten Effektstärke $d = 0.4$ ist nicht ersichtlich, ob die Viruslast nennenswert reduziert wurde oder, ob der Rückgang tatsächlich 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).

Effektstärke bei Zusammenhängen

Korrelationskoeffizient

- Im Fall von Zusammenhangsanalysen haben wir bereits standardisierte Effektstärke kennengelernt — die **Korrelationskoeffizienten** ($\hat{\rho}$, $\hat{\rho}_s$, $\hat{\tau}$, $\hat{\phi}$).
- Beispiel Pearson-Korrelation:

$$\hat{\rho} = \frac{\hat{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

- Der Nenner $\sigma_X \sigma_Y$ stellt hier die Standardisierung dar.

Alle Korrelationskoeffizienten ($\hat{\rho}$, $\hat{\rho}_s$, $\hat{\tau}$, $\hat{\phi}$) sind bereits standardisierte Effektstärken.

Zusammenhang von Cohen's d und Korrelationskoeffizient $\hat{\rho}$

- Fassen Metaanalysen sowohl Studien zusammen, die Effekte als Korrelation berichten (Effektmaß $\hat{\rho}$), als auch Studien, die Effekte als Mittelwertsunterschiede zwischen Bedingungen/Gruppen berichten (Effektmaß d), entsteht die Notwendigkeit $\hat{\rho}$ und d ineinander umzurechnen.
- Folgende Formel findet häufig Anwendung:

$$d = \frac{2\hat{\rho}}{\sqrt{1-\hat{\rho}^2}}$$

- Nach dem Korrelationskoeffizient $\hat{\rho}$ aufgelöst:

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

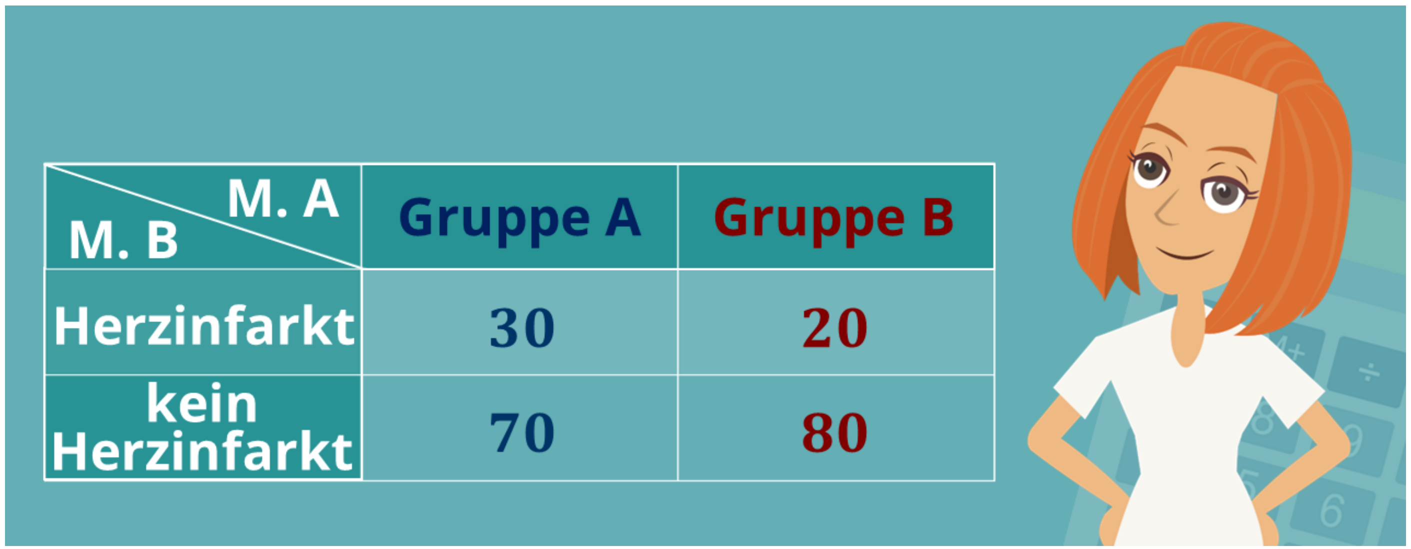

Nota Bene

Aufgrund ihrer Einfachheit finden die obigen Formeln vielfach Anwendung in der Praxis. Weniger bekannt ist, dass die Formeln implizit annehmen, dass einer der beiden Variablen der Korrelation als *unabhängige Variable* X definiert wird, und dass die Umrechnungsformel für Cohen's d de facto die Effektstärke in der Variablen Y schätzt, die mit einer Erhöhung der unabhängigen Variablen X um zwei Standardabweichungen σ_X verbunden ist. Es lässt sich zeigen, dass die Formeln lediglich Spezialfall einer allgemeineren Formel sind (siehe Bonuscontent Umrechnung von Cohen's d zu $\hat{\rho}$: ein Nachtrag).

Weitere Effektmaße

Weitere Effektmaße

- Im Bereich klinischer Studien finden sich weitere Effektmaße, die in der Regel zum Ziel haben, die Wirksamkeit einer Intervention (Behandlung) zu quantifizieren.
- Folgende drei Effektmaße aus diesem Kontext betrachten wir:
  - Absolute Risikoreduktion (ARR)
  - Numbers Needed to Treat (NNT)
  - Odds ratio (OR)



M. A / M. B	Gruppe A	Gruppe B
Herzinfarkt	30	20
kein Herzinfarkt	70	80

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:

	male	female	Σ
nonsmoker	a	b	a+b
smoker	c	d	c+d
Σ	a+c	b+d	n

Erfolgsrate in der Treatmentgruppe

$$ARR = \boxed{\frac{a}{a+b}} - \boxed{\frac{c}{c+d}}$$

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.

Defin ition

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

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

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

Odds Ratio (OR)

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

	male	female	Σ
nonsmoker	a	b	a+b
smoker	c	d	c+d
Σ	a+c	b+d	n

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

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. $OR < 1$.

ARR, NNT, OR: Caveat

- Auch wenn das Ziel der Effektmaße ARR, NNT, OR eine einfache laienverständliche Kommunikation der Treatmenteffizienz ist, darf angezweifelt werden, ob dies immer der Fall ist, wie durch folgendes Beispiel demonstriert[8]:

> 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.

- Hinweis: ARR und OR haben das gleiche Problem.

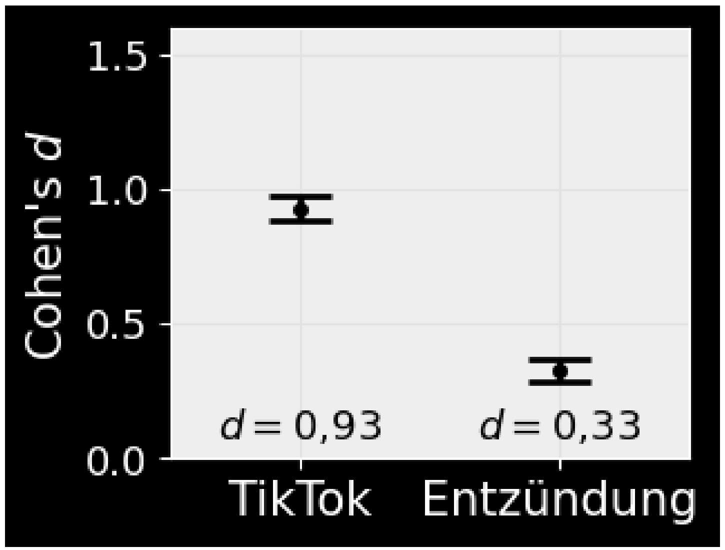
Zusammenfassung: Übersicht Effektmaße

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

[[Zusammenfassung]]

- Die praktische Signifikanz von Effekten wird als **Effektstärke** bezeichnet.
- Effektstärken sind **standardisierte einheitslose Maße**, die **zwischen Studien verglichen** werden können.
- Während **Korrelationskoeffizienten bereits standardisiert** sind, müssen **Mittelwertsunterschiede** erst in standardisierte Effektmaße überführt werden: **Cohend's d**.
- Die Berechnung von Cohen's d hängt ab vom Design ab: Standardisierung mit der **Differenzenvarianz** bei abhängigen Messungen und mit der **gepoolten Varianz** bei unabhängigen Messungen.
- Im medizinischen Kontext sind weitere Effekmaße geläufig: absolute Risikoreduktion (ARR), Numbers Needed to Treat (NNT) und Odds Ratio (OR).

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: $d = 0{,}93$.

Bonuscontent

Gepoolte Varianz

- Bei unabhängigen Messungen ist es nicht (wie in Fall 2) möglich, die Einzelmesswerte $x_{A,i}$ und $x_{B,i}$ direkt voneinander abzuziehen. Nicht zuletzt wäre es völlig unklar, welche Versuchsperson aus A jeweils von welcher Versuchsperson aus B subtrahiert wird.
- Damit kann auch nicht die Varianz der individuellen Effekte der Versuchspersonen berechnet werden ("Differenzenvarianz"), sondern lediglich eine **gepoolte Varianz** $\hat{\sigma}^2_{\text{pooled}}$.
- Die Annahme ist dabei, dass beide Gruppen eine ähnliche Streuung haben, z.B. weil sie randomisiert aus derselben Population gezogen wurden und man nicht erwartet, dass die Gruppenzugehörigkeit die Varianz beeinflusst.
- Der Name leitet sich von der Vorstellung ab, die Datenpunkte beider Gruppen in einen gemeinsamen "Pool" zu werfen, und dann die Varianz auf allen Daten zu berechnen.
- Vor dem "Wurf in den Pool" müssen die Mittelwerte der Datenpunkte aus den jeweiligen Gruppen gleichgesetzt werden — im einfachsten Fall so, dass jeweils der Gruppenmittelwert von den Datenpunkten abgezogen wird (die Daten werden **zentriert** — beide Gruppen haben danach den Mittelwert 0):

$$X'_A = X_A - \bar{X}_A \qquad und \qquad X'_B = X_B - \bar{X}_B$$

$$X_{\text{pooled}} = [X'_A; X'_B]$$

$$\hat{\sigma}^2_{\text{pooled}} = \hat{Var}(X_{\text{pooled}})$$

Gepoolte Varianz

- Es lässt sich mathematisch recht einfach zeigen, dass die gepoolte Varianz identisch dem an den Stichprobengrößen n_A und n_B gewichteten Varianzmittelwert ist:

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

Die Subtraktion von 1 von den Stichprobengrößen n_A und n_B ist Ausdruck der Besselkorrektur.

- Die Gewichtung mit den Stichprobengrößen grenzt die gepoolte Varianz von der einfach gemittelten Varianz $\hat{\sigma}^2_{\text{av}}$ ab:

$$\hat{\sigma}^2_{\text{av}} = \frac{\hat{\sigma}^2_A + \hat{\sigma}^2_B}{2} \quad \text{bzw.} \quad \hat{\sigma}^2_{\text{av}} = \sqrt{\frac{\hat{\sigma}^2_A + \hat{\sigma}^2_B}{2}}$$

- $\hat{\sigma}^2_{\text{av}}$ wird auch als **ungepoolte Varianz** bezeichnet, weil hier nicht alle Datenpunkte in einen Pool geworfen werden, sondern die Varianzen $\hat{\sigma}^2_A$ und $\hat{\sigma}^2_B$ einzeln berechnet und anschließend ohne Berücksichtigung unterschiedlicher Stichprobengrößen gemittelt werden.

Gepoolte Varianz

- Der Nachteil der gepoolten Varianz ist, dass sie nicht mehr der interindividuellen Variabilität des Effektes (hier Mittelwertdifferenz) entspricht, sondern der Variabilität des gemessenen Merkmals X.
  - Die Mittelwertdifferenz im Zähler wird also streng genommen nicht mehr ins Verhältnis zu ihrer eigenen Variabilität gesetzt.
- Die gepoolte Varianz hat jedoch auch einen Vorteil: es wird de facto nur eine Varianz geschätzt (die Popuationsvarianz des Merkmals X) und diese Schätzung basiert auf $n_A + n_B$ Versuchspersonen. Sie kann also recht präzise geschätzt werden. Die Präzision der Varianz bestimmt wiederum direkt und wesentlich die Präzision des Effektstärkenmaßes selbst.
- Wenn Effektstärken in erster Linie zur Vergleichbarmachung mit anderen Studien dienen (Stichwort **Metaanalyse**), ist eine a) *präzise* und b) *zwischen Studien vergleichbare* Schätzung der Varianz ein wichtiger Aspekt — ggf. wichtiger als die Interpretierbarkeit des resultierenden Effektmaßes.

Beispiel

Bei einem Pre-post-Interventionsdesign (Messung vor und nach einer Intervention an derselben Gruppe), kann es vorteilhaft sein, nur die *Standardabweichung des Vortestes* für die Standardisierung zu nehmen. Grund: die Variabilität *vor* einer Intervention ist am ehesten zwischen Studien vergleichbar — damit werden auch die Effektstärken besser vergleichbar.[9] In der Praxis geschieht dies aber selten (nicht immer ist die statistisch beste auch die populärste Praxis).

Cohen’s d bei unterschiedlichen Gruppenvarianzen

- Eine wichtige Voraussetzung für die Verwendung der gepoolten Varianz bei Cohen’s d mit unabhängigen Messungen ist, dass die Einzelvarianzen $\hat{\sigma}_A^2$ und $\hat{\sigma}_A^2$ ähnlich sind.
- Warum? Die Idee ist, die Varianzschätzung des Merkmals durch Nutzung der Datenpunkte *beider* Gruppen zu verbessern. Sind die Varianzen in beiden Gruppen jedoch unterschiedlich, so ist der Fall gegeben, dass es gar nicht *die eine* Varianz des Merkmals gibt, sondern die Varianz stark von der Gruppenzugehörigkeit abhängt.
- Eine gängige Faustregel besagt, dass die Varianzen sich höchstens um den Faktor 2 unterscheiden sollten, d.h. $0.5 < \frac{\hat{\sigma}_A^2}{\hat{\sigma}_B^2} < 2$.
- Diese Einschränkung bei der Verwendung der gepoolten Varianz wird jedoch häufg missachtet, mutmaßlich auch, weil sich keine alternative Lösung für den Fall unabhängiger Messungen in großer Breite etabliert hat.
- Eine häufig empfohlene Variante im Fall unabhängiger Messungen und unterschiedlicher Varianzen ist die Effektstärkengröße *Hedges*’ g* auf Basis der **ungepoolten Varianz** $\hat{\sigma}_{\text{av}}$ [10]:

$$\text{Hedges' } g^* \approx \left(1 - \frac{3}{4(n_A + n_B) - 9}\right) \cdot \frac{\bar{x}_A - \bar{x}_B}{\hat{\sigma}_{\text{av}}} \quad \text{mit} \quad \hat{\sigma}_{\text{av}} = \sqrt{\frac{\hat{\sigma}_A^2 + \hat{\sigma}_B^2}{2}}$$

Effektstärkenvergleich zwischen abhängigen und unabhängigen Designs

Problem: häufig sollen Effektstärken zwischen Designs mit *abhängigen* und *unabhängigen* Messungen verglichen werden — jedoch lässt sich in Fall 2 (unabhängige Messungen) keine Differenzenvarianz berechnen.

- In diesem Fall wird manchmal folgende Variante von Cohen's d bei abhängigen Messungen verwendet, die auf der **gepoolten Varianz** basiert:

$$d = \frac{\bar{x}_A - \bar{x}_B}{\hat{\sigma}_{\text{pooled}}} \quad \text{mit} \quad \hat{\sigma}_{\text{pooled}} = \sqrt{\frac{\hat{\sigma}_A^2 + \hat{\sigma}_B^2}{2}}$$

- Beachte, dass hier die gepoolte Varianz $\hat{\sigma}^2_{\text{pooled}}$ identisch mit der ungepoolten (einfach gemittelten) Varianz $\hat{\sigma}^2_{\text{av}}$ ist, da bei abhängigen Messungen immer gilt: $n = n_A = n_B$.

Effektstärkenvergleich zwischen abhängigen und unabhängigen Designs

- Laut einer aktuellen Forschungsarbeit[11] sind die Werte von Cohen's d auf Basis der gepoolten Varianz nicht exakt vergleichbar zwischen abhängigen und unabhängigen Messungen (auch dann, wenn die Varianzen ähnlich sind).
- Die Arbeit bietet dafür folgende Modifikation von Cohen's d für abhängige Messungen an:

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

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. Beachte, dass die Formel nur gilt, falls ähnliche Varianzen in den beiden Bedingungen A und B angenommen werden können.

- Laut dem Autor Denis Cousineau der Forschungsarbeit ist damit Cohen's d exakt vergleichbar zwischen abhängigen und unabhängigen Messungen.
- Weiterer nützlicher Link zu Effektstärken bei abhängigen Messungen: [12]

Umrechnung von Cohen's d zu $\hat{\rho}$: ein Nachtrag

- Wie erwähnt, impliziert die Formel $d = \frac{2\hat{\rho}}{\sqrt{1-\hat{\rho}^2}}$ eigentlich, dass eine der beiden Variablen als *unabhängige Variable* festgelegt wurde. Cohen's d bemisst dabei die Effektstärke in der Variablen Y, die mit einer Erhöhung der unabhängigen Variablen X um zwei Standardabweichungen σ_X verbunden ist.
- Tatsächlich ist die angebene Formel lediglich ein Spezialfall folgender allgemeiner Formel[13]:

$$d = \frac{k\hat{\rho}}{\sigma_X\sqrt{1-\hat{\rho}^2}} \qquad \text{bzw.} \qquad \hat{\rho} = \frac{d\sigma_X}{\sqrt{d^2\sigma_X^2 + k^2}}$$

- Hier gilt allgemein: d entspricht der durchschnittlichen Zunahme der standarsierten Y-Variable mit jeder Zunahme von X um k (Rohwert)Einheiten; k muss vom Forschenden gewählt werden.
- Wählt man $k = 2\sigma_X$, so ergibt sich tatsächlich die vereinfachte Formel.
- Allerdings ist $k = 2\sigma_X$ zum einen ein ziemlich extremer Kontrast und zum anderen von der Studien- oder Populations-spezifischen Standardabweichung σ_X abhängig. Daher scheint in Metaanalysen eigentlich die Festlegung auf einen kleineren und konstanten Wert von k sinnvoll.

Effektstärken für Mittelwertunterschiede bei mehr als zwei Messungen

- Gibt es mehr als zwei Experimentalbedingungen oder Gruppen (A, B, C, ...), gibt es zwei Möglichkeiten:
  1. Von Interesse sind die **paarweisen** Mittelwertunterschiede (z.B: $A-B, A-C, B-C$) → in diesem Fall kann wie bisher Cohen's d für für jedes Paar angewendet werden.
  2. 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 Mittlwertunterschiede von mehreren Messungen als *Verhältnis zweier Streuungen* ausgedrückt werden kann:

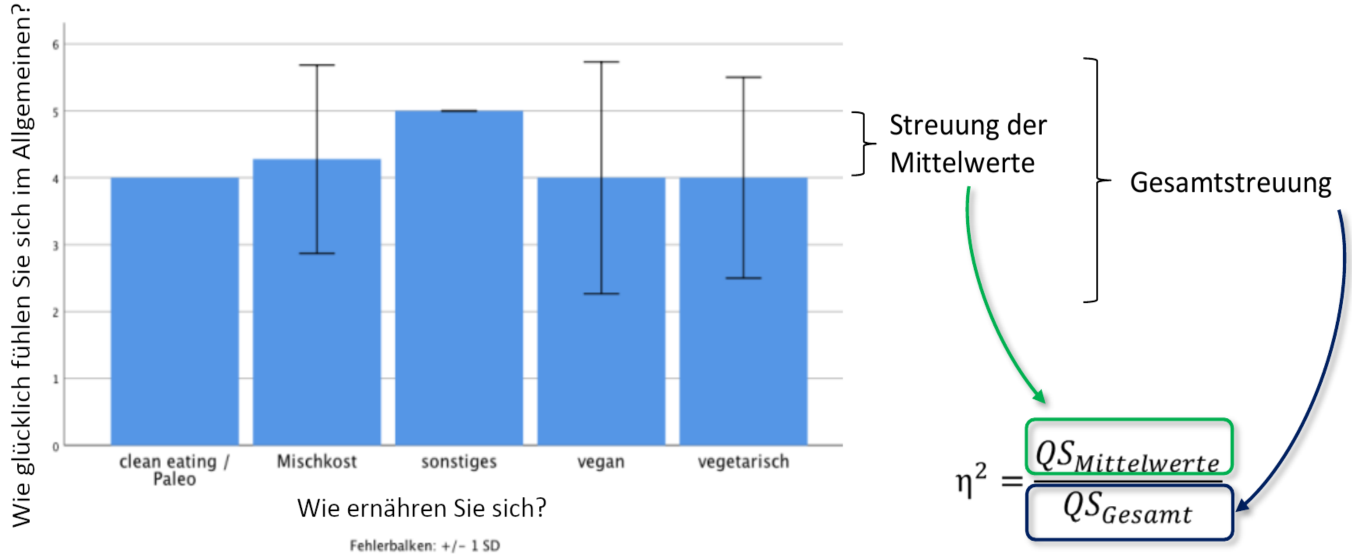
$$\eta^2 = \frac{QS_{\text{Mittelwerte}}}{QS_{\text{Gesamt}}}$$

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

Quadrat-
summe (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 zwischen Personen (über Bedingungen hinweg), und 3) wie sehr sich die Varianz der Bedingungen zwischen Personen unterscheidet:

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

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.:

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

- 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.


Nota Bene

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.

Empfehlungen nach Lakens (2013)[14]

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 then 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

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. 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.
4. Lovakov A, Agadullina ER (2021) Empirically derived guidelines for effect size interpretation in social psychology. Eur J Soc Psychol 51:485–504.
5. Lovakov A, Agadullina ER (2021) Empirically derived guidelines for effect size interpretation in social psychology. Eur J Soc Psychol 51:485–504.
6. https://twitter.com/ceptional/status/1687577019629142017
7. Baguley T (2009) Standardized or simple effect size: What should be reported? British Journal of Psychology 100:603–617.
8. Andrade C (2015) The Numbers Needed to Treat and Harm (NNT, NNH) Statistics: What They Tell Us and What They Do Not: (Practical Psychopharmacology). J Clin Psychiatry 76:e330–e333.
9. Cumming G (2013) Cohen’s d needs to be readily interpretable: Comment on Shieh (2013). Behav Res 45:968–971.
10. Delacre M, Lakens D, Ley C, Liu L, Leys C (2023) Why Hedges’ g*s based on the non-pooled standard deviation should be reported with Welch’s t-test. Open Science Framework. Available at: https://osf.io/tu6mp. Hinweis: der Terminus “ungepoolte” Varianz meint hier, dass die Daten nicht implizit in einen Pool geworfen werden und dann die Gesamtvarianz berechnet wird; stattdessen wird unabhängig von möglicherweise unterschiedlich großen Gruppengrößen n_A und n_B der Mittelwert der beiden Einzelvarianzen berechnet.
11. Cousineau D (2020) Approximating the distribution of Cohen’s d_p in within-subject designs. TQMP 16:418–421.
12. http://jakewestfall.org/blog/index.php/2016/03/25/five-different-cohens-d-statistics-for-within-subject-designs/
13. Mathur MB, VanderWeele TJ (2020) A Simple, Interpretable Conversion from Pearson’s Correlation to Cohen’s d for Continuous Exposures. Epidemiology 31:e16–e18.
14. Lakens D (2013) Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. Frontiers in Psychology 4 Available at: https://www.frontiersin.org/articles/10.3389/fpsyg.2013.00863