Jahrbuch Musikpsychologie
jbdgm.psychopen.eu | 2569-5665

Spots

Die Berechnung des Konfidenzintervalls für die Effektgröße Cohen's d

Computing the Confidence Interval for the Effect Size Cohen's d

Viola Pausch*[a]

[a] Hochschule für Musik, Theater und Medien Hannover, Hannover, Deutschland.

Zusammenfassung

Mit Hilfe der Effektgröße Cohen's d kann ein Effekt quantitativ und metrikfrei geschätzt werden. Dieser Effekt kann z. B. durch die Abweichung eines Mittelwertes von einem bestimmten Wert oder durch den Mittelwertsunterschied zwischen zwei Stichproben zustande kommen. Die Breite eines Konfidenzintervalls für die Effektgröße Cohen's d gibt an, wie genau diese Schätzung ist. In diesem Beitrag soll gezeigt werden, warum nichtzentrale t-Verteilungen eine große Rolle in der präzisen Berechnung der Konfidenzintervalle für Cohen's d spielen und wie diese berechnet werden können. Auf der Online-Plattform Open Science Framework stehen zwei Programme in R frei zur Verfügung, die ein solches Konfidenzintervall für Cohen's d für eine bzw. für zwei Stichproben ausgehend von den folgenden Eingabeparametern berechnen: Konfidenzniveau (z. B. 95%), Stichprobengröße(n), Mittelwert(e) und Standardabweichung(en). Am Ende dieses Beitrags wird das Vorgehen an einem Beispiel illustriert.

Schlüsselwörter: Effektgrößen, Cohen's d, Konfidenzintervall, nichtzentrale t-Verteilung, R, Statistik

Abstract

The effect size Cohen's d allows for a quantitative and metric-free estimation of an effect. This effect can be the result of the deviation of a mean value from a certain value or the mean difference between two samples. The precision of this estimation is given by the width of a confidence interval for the effect size Cohen's d. The aim of this article is to show the importance of noncentral t distributions for a precise estimation of confidence intervals for Cohen's d and to explain how to compute them. On the Open Science Framework online platform, two programs in R are freely available that calculate the confidence intervals for Cohen's d for one or two samples based on the following input variables: confidence level (e.g. 95%), sample size(s), mean(s) and standard deviation(s). The article concludes by illustrating the discussed approach with an example.

Keywords: Effect size, Cohen's d, confidence interval, noncentral t distribution, R, statistics

Jahrbuch Musikpsychologie, 2018, Vol. 28: Musikpsychologie — Musik und Bewegung, Artikel e29, https://doi.org/10.5964/jbdgm.2018v28.29

Eingereicht: 2018-01-02. Akzeptiert: 2018-09-30. Publiziert (VoR): 2019-03-07.

*Korrespondenzanschrift: Hochschule für Musik, Theater und Medien Hannover, Neues Haus 1, 30175 Hannover, Deutschland. E-Mail: pauschv@stud.hmtm-hannover.de

Einleitung

Die Effektgröße Cohen's *d* dient dazu, die Abweichung des Gruppenmittelwerts einer Stichprobe von einem bestimmten Wert μ_0 bzw. den Mittelwertsunterschied von zwei unabhängigen Stichproben im Sinne von anteiligen Standardabweichungen metrikfrei zu beschreiben. Da Cohen's *d* standardisiert ist, ist ein Vergleich von mehreren Werten für Cohen's *d* aus verschiedenen Messinstrumenten bzw. Studien zum Beispiel bei Metaanalysen möglich (Thompson, 2002, S. 27). Um die Größe und damit die Bedeutung eines gefundenen Effekts beurteilen

zu können, halten sich viele Forscherinnen und Forscher an die Richtwerte von Cohen (Cohen, 1988, S. 25f.; s. auch Ellis, 2010, S. 41), der eine gemessene Effektgröße in klein (d = 0,2), mittel (d = 0,5) oder groß (d = 0,8) einteilt. Die Arbeitsgruppe für statistische Inferenz der American Psychological Association (APA) betont in ihren *Guidelines and Explanations* (Wilkinson and the Task Force on Statistical Inference, 1999, S. 599), dass es für gute Forschung unabdingbar sei, Effektgrößen im Kontext von aus der Literatur bereits bekannten Effektgrößen zu berichten und zu interpretieren. Dadurch könnten Leserinnen und Leser von Forschungsberichten beurteilen, ob die Ergebnisse über Stichproben, Designs und Analysen hinweg stabil sind. Dennoch wird dieser dringenden Empfehlung bis heute in der Fachliteratur kaum nachgekommen (Platz, Kopiez & Lehmann, 2012).

Ein Konfidenzintervall für Cohen's d gibt Aufschluss über die Genauigkeit, mit der der Populationsparameter δ durch die Stichprobenstatistik d geschätzt wurde[i]. Ein 95%-Konfidenzintervall für den Populationsparameter δ bedeutet, dass bei wiederholter Ziehung einer gleich großen Stichprobe 95% der 95%-Konfidenzintervalle den tatsächlichen Populationsparameter der Effektgröße δ enthalten.

In dem in dieser Reihe erschienenen Artikel *Statistische Poweranalyse als Weg zu einer 'kraftvolleren' Musikpsychologie im 21. Jahrhundert* (Platz et al., 2012) wurde im Rahmen einer Post-hoc-Analyse mit Hilfe der Standardnormalverteilung das 95%-Konfidenzintervall für die Effektgröße d = 1,71 berechnet, welches Werte von 0,18 bis 3,24 umfasst. In diesem Beitrag wird eine Weiterentwicklung der bisherigen Vorgehensweise vorgeschlagen und es soll gezeigt werden, wie Konfidenzintervalle für Cohen's d mit Hilfe von nichtzentralen t-Verteilungen präziser berechnet werden können. Die hier vorgeschlagene methodische Vorgehensweise unterscheidet sich von der in Platz et al. (2012) gezeigten: Das mit Hilfe der nichtzentralen t-Verteilung ermittelte 95%-Konfidenzintervall für die Effektgröße d = 1,71 ist [0,09; 3,25] und damit breiter als das von Platz et al. berechnete Konfidenzintervall [0,18; 3,24]. Dies lässt darauf schließen, dass wegen des geringen Stichprobenumfangs (n = 4 und m = 5) in der von Platz et al. reanalysierten Studie die Schätzung noch unpräziser ist als von Platz et al. bereits angenommen.

Vorteile der nichtzentralen t-Verteilung und Vorgehensweise

Im Folgenden soll begründet werden, warum nichtzentrale t-Verteilungen zur Bestimmung von Konfidenzintervallen für Cohen's d benötigt werden und wie diese schließlich berechnet werden. Dazu betrachte man zunächst die im zweiten Kasten in Abbildung 1 angegebenen Formeln für Cohen's d für eine (linke Seite) und für zwei unabhängige Stichproben (rechte Seite). Hierbei fällt auf, dass die Verteilungen der Teststatistiken von d ($\frac{\overline{X}-\mu_0}{S}$ bzw. $\frac{\overline{X}-\overline{Y}}{S_{A,B}}$) von zwei Verteilungen, nämlich der Verteilung von $\overline{X}$ bzw. von $\overline{X}-\overline{Y}$ und der Verteilung von S bzw. $S_{A,B}$ abhängen. Dies steht im Gegensatz zur Teststatistik $\overline{X}$, die nur von einer Verteilung abhängt. Daher können Konfidenzintervalle für die Effektgröße Cohen's d nicht auf dieselbe Art und Weise wie Konfidenzintervalle für den Erwartungswert μ berechnet werden.

Die Idee für die Bestimmung eines Konfidenzintervalls für Cohen's d (Cumming & Finch, 2001, S. 550f.; Smithson, 2003, S. 34 ff.) besteht darin, ein Konfidenzintervall $[\Delta_L, \Delta_U]$ für den sog. Nichtzentralitätsparameter Δ zu suchen und dieses mit Hilfe der Beziehung $\Delta = \delta\sqrt{n}$ bzw. $\Delta = \sqrt{\frac{n \cdot m}{n+m}}\,\delta$ wie in Abbildung 1 verdeutlicht in ein Konfidenzintervall für δ umzuwandeln.

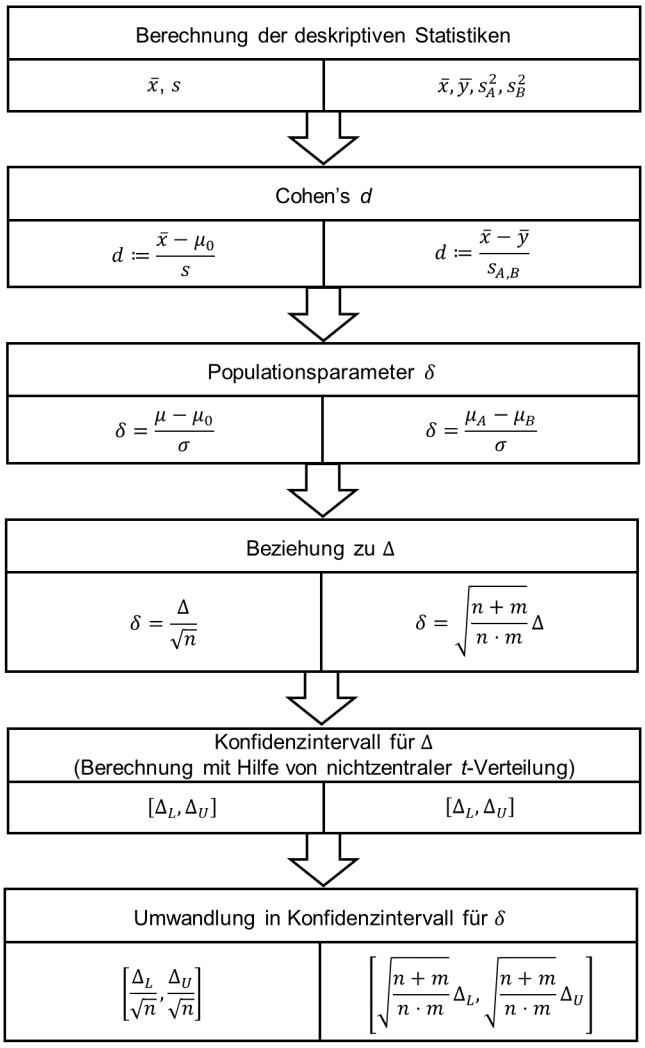

Abbildung 1. Von den deskriptiven Statistiken zum Konfidenzintervall für Cohen's *d*. Die linke Seite des Flussdiagramms stellt das Vorgehen für eine Stichprobe und die Abweichung des Mittelwerts $\bar{x}$ von einem bestimmten Wert μ_0 (z. B. $\mu_0 = 0$) dar. Die rechte Seite zeigt das Vorgehen für zwei unabhängige Stichproben. Die Indizes L bzw. U stehen für die untere (lower) bzw. obere (upper) Intervallgrenze. Die Standardabweichung s und die zusammengefasste Standardabweichung $s_{A,B}$ seien folgendermaßen definiert: $s \coloneqq \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$ und $s_{A,B} \coloneqq \sqrt{\frac{(n-1)s_A^2 + (m-1)s_B^2}{n+m-2}}$, wobei *n* und *m* die Größen der Stichproben A und B sind.

Die Teststatistik von Δ ($T \coloneqq \frac{\overline{X} - \mu_0}{S/\sqrt{n}}$ bzw. $T_{A,B} \coloneqq \sqrt{\frac{n \cdot m}{n+m}}\frac{\overline{X} - \overline{Y}}{S_{A,B}}$) hat eine nichtzentrale *t*-Verteilung mit *n* - *1* bzw. *n* + *m* - 2 Freiheitsgraden und Nichtzentralitätsparameter Δ = $\frac{\mu - \mu_0}{\sigma/\sqrt{n}}$ bzw. Δ = $\sqrt{\frac{n \cdot m}{n+m}}\frac{\mu_A - \mu_B}{\sigma}$, wenn von einer Normalverteilung mit Streuung σ der identisch verteilten, paarweise stochastisch unabhängigen, reellen Zufallsvariablen $X_1, ..., X_n$ und $Y_1, ..., Y_m$ ausgegangen wird. Warum dies so ist, soll hier nicht näher erläutert werden. Für detaillierte Informationen sei stattdessen auf Smithson (2003, S. 33 ff.) verwiesen. Wegen der nichtzentralen *t*-Verteilung ist der Umgang mit der Teststatistik von Δ komplexer. Es gibt beispielsweise keine Tabellen für bestimmte *t*-Quantile wie für die zentrale *t*-Verteilung. Abhilfe schaffen die beiden Programme *CI1.R* und *CI2.R* in R, mit deren Hilfe die Grenzen des Konfidenzintervalls für δ bei einer (*CI1.R*) bzw. bei zwei (*CI2.R*) unabhängigen Stichproben bestimmt werden können. Der jeweils erste Teil der beiden Programme stammt hierbei von

Smithson (Smithson, 2019). Der Code wurde leicht verändert. Die Programme benötigen als Eingabewerte das Konfidenzniveau (z. B. 95%), die Stichprobengröße(*n*), den Mittelwert bzw. die Mittelwerte und die Standardabweichung(en). Die Software ist als ergänzendes Material verfügbar.

Anwendungsbeispiel

Das Vorgehen soll nun an einem Beispiel verdeutlicht werden, in dem das 95%-Konfidenzintervall für Cohen's *d* für den Mittelwertsunterschied der Variable *mean_NEO* zwischen zwei unabhängigen Stichproben (Frauen und Männer) berechnet wird. Der Datensatz hierfür stammt aus den Beispieldatensätzen, welche das Statistik-Programm *JASP* zur Verfügung stellt und heißt *Kitchen Rolls - A nice t-test data set*. (Die vollständigen Beispieldateien für *JASP* sind als ergänzende Materialien verfügbar.) Die Daten wurden in einer Studie von Wagenmakers et al. (2015) erhoben. Hierbei wurde mittels zwölf Items der „Openness to experience"-Subskala aus dem Neurotizismus-Extraversion-Offenheits-Persönlichkeitsinventar (NEO PI-R; Costa & McCrae, 1992) die Vorliebe für neuartige Erfahrungen und Aktivitäten gemessen. Die Variable *mean_NEO* wird durch den Mittelwert der Antworten auf diese zwölf Items gebildet. Die Mittelwerte und Standardabweichungen von *mean_NEO* sind in Tabelle 1 dargestellt.

Tabelle 1

Mittelwerte und Standardabweichungen der beiden Gruppen (Frauen und Männer)

	mean_NEO		
Gruppe	***N***	***M***	***SD***
Frauen	77	$\bar{x}$ = 0,592	s_A = 0,474
Männer	25	$\bar{y}$ = 0,947	s_B = 0,412

Das Programm *CI2.R* berechnet als Effektgröße $d = \frac{\bar{x} - \bar{y}}{s_{A,B}} = \frac{0{,}592 - 0{,}947}{0{,}460} = -0{,}77$. Das vom Programm ermittelte 95%-Konfidenzintervall für Δ ist [-5,36; -1,33]. Daraus folgt schließlich [-1,23; -0,31] für das gesuchte 95%-Konfidenzintervall für δ. Die negativen Werte kommen dadurch zustanden, dass der Mittelwert der Frauen (0,592) kleiner als der der Männer (0,947) und daher die Differenz negativ ist. Für die Interpretation benötigt man den absoluten Wert der Effektgröße, also *d* = 0,77. Laut Cohens Benchmarks (Cohen, 1988, S. 25f.; s. auch Ellis, 2010, S. 41) ist dies ein mittlerer bis großer Effekt. Das Konfidenzintervall [0,31; 1,23] lässt jedoch vermuten, dass ein kleiner bis großer Effekt vorliegt. Die Schätzung des Populationsparameters δ ist demnach nicht besonders präzise.

Schlussfolgerungen

Die nichtzentrale *t*-Verteilung wurde Smithson (Smithson, 2003, S. 41) zufolge noch in den 1980er-Jahren außer Acht gelassen, weil sie ohne passende Software nicht verwendbar war. Ihre Wiederentdeckung in den 1990er-Jahren und die Verfügbarkeit benutzerfreundlicher Software für ihre Berechnung (Smithson, 2003, S. 41) wie beispielsweise die hier vorgestellten Programme *CI1.R* und *CI2.R* sollten Forscherinnen und Forscher nun darin bestärken, mehr Gebrauch von nichtzentralen *t*-Verteilungen zu machen. Damit können sie von

dem Vorteil profitieren, dass die Schätzung des Populationsparameters δ mit Hilfe der nichtzentralen *t*-Verteilung vor allem bei kleinen Stichproben mathematisch exakter ist als mit Hilfe der zentralen *t*-Verteilung.

Anmerkungen

i) Im Grunde genommen handelt es sich hier nicht um das Konfidenzintervall für Cohen's *d*, sondern um das Konfidenzintervall für den wahren, aber unbekannten Populationsparameter von Cohen's *d*, der mit δ bezeichnet wird. In der Literatur werden beide Formulierungen verwendet.

Finanzierung

Die Autorin hat keine Finanzierung für das Forschungsprojekt erhalten.

Interessenkonflikte

Die Autorin hat erklärt, dass keinerlei konkurrierende Interessen bestehen.

Danksagung

Mein Dank gilt Prof. Dr. Reinhard Kopiez (HMTM Hannover) und Dr. Björn Böttcher (Institut für Mathematische Stochastik, TU Dresden) für ihre Unterstützung.

Ergänzende Materialien

Zu diesem Artikel sind die folgenden ergänzenden Materialien verfügbar:

- CI1.R und CI2.R: R Skripte zur Berechnung des Konfidenzintervalls für δ bei einer (CI1.R) bzw. bei zwei (CI2.R) unabhängigen Stichproben.
- Datensatz KitchenRolls.csv (Wagenmakers et al., 2019) des Anwendungsbeispiels.

Quellenverzeichnis der ergänzenden Materialien

Pausch, V. (2019). Materialien zu "Berechnung des Konfidenzintervalls für Cohen's *d*". PsychOpen. Abrufbar im PsychArchives Repositorium: https://doi.org/10.23668/psycharchives.2368

Wagenmakers, E.-J., Morey, R. D., Etz, A. & Steingroever, H. (2019). Materialien zu "Turning the hands of time again: A purely confirmatory replication study and a Bayesian analysis". Open Science Framework. Abrufbar im OSF Repositorium (im Ordner *t-test*): https://osf.io/6zr98/

Literatur

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ, USA: Erlbaum.

Costa, P. T. & McCrae, R. R. (1992). *NEO personality inventory professional manual*. Odessa, FL, USA: Psychological Assessment Resources.

Cumming, G. & Finch, S. (2001). A primer on the understanding, use, and calculation of confidence intervals that are based on central and noncentral distributions. *Educational and Psychological Measurement, 61*, 532-574. https://doi.org/10.1177/0013164401614002

Ellis, P. (2010). *The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results*. Cambridge, United Kingdom: Cambridge University Press.

Platz, F., Kopiez, R. & Lehmann, M. (2012). Statistische Poweranalyse als Weg zu einer 'kraftvolleren' Musikpsychologie im 21. Jahrhundert. In W. Auhagen, C. Bullerjahn & H. Höge (Eds.), *Musikpsychologie Bd. 22* (pp. 165-179). Göttingen, Germany: Hogrefe Verlag.

Smithson, M. (2003). *Sage university papers series on quantitative applications in the social sciences: Vol. 07-140. Confidence intervals*. Thousand Oaks, CA, USA: Sage.

Smithson, M. (2019). Noncent.ssc [Computer program]. Retrieved from http://www.michaelsmithson.online/stats/CIstuff/Nonct.ssc

Thompson, B. (2002). What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher, 31*(3), 25-32. https://doi.org/10.3102/0013189X031003025

Wagenmakers, E.-J., Beek, T. F., Rotteveel, M., Gierholz, A., Matzke, D., Steingroever, H., . . . Pinto, Y. (2015). Turning the hands of time again: A purely confirmatory replication study and a Bayesian analysis. *Frontiers in Psychology, 6*, Article 494. https://doi.org/10.3389/fpsyg.2015.00494

Wilkinson, L. & the Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*, 594-604. https://doi.org/10.1037/0003-066X.54.8.594