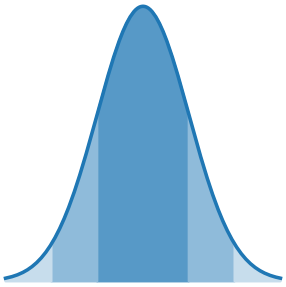

Konfidenzintervall-Rechner

Anleitung

21. Dezember 2020

1 Autoren

Dipl.-Psych. Jonas Schemmel, M.Sc. Rechtspsychologie

(Psychologische Hochschule Berlin (PHB) - Professur Rechtspsychologie)

Johannes Wiesner, M.Sc. Psychologie

(Zentralinstitut für seelische Gesundheit, Mannheim)

2 Vorbemerkung

Die folgende Anleitung basiert im Wesentlichen auf Ziegler & Bühner (2012) und Bühner (2010). Beide Autoren sind Mitglieder im Diagnostik- und Testkuratorium (DTK), das am 18.10.2017 im Auftrag der deutschen Psychologenvereinigungen (BDP/DGPs) Qualitätsstandards für psychologische Gutachten verabschiedet und darin auch die Angabe von Konfidenzintervallen bei psychometrischen Tests vorgegeben hat.

3 Nutzungsbedingungen und Lizenz

4 Warum Konfidenzintervalle berichtet werden sollten

Noch immer existieren psychodiagnostische Instrumente, in denen der Bericht von Konfidenzintervallen nicht standardmäßig enthalten ist. In Ziegler und Bühner (2012, S. 129 ff.) heißt es dazu:

> Das Ergebnis eines psychometrischen Tests oder Fragebogens ist zunächst erst einmal ein Rohwert. Dieser ergibt sich meist als die Summe der richtig gelösten Items (im Test) bzw. die Summe der angekreuzten Kategorien (im Fragebogen). Die Verwendung von Normwerten ermöglicht es dann, den Wert so zu transformieren, dass er in Bezug auf eine Vergleichsgruppe interpretiert werden kann. Die Interpretation des Normwerts hat nun mindestens zwei problematische Aspekte. Zum einen ist der Normwert ohne Kenntnis der Vergleichsgruppe wenig aussagekräftig. Das zweite Problem hängt mit der Messgenauigkeit zusammen. Wenn wir sagen, dass eine Person im Vergleich zur Norm einen IQ von 97 hat, dann legt diese Aussage nahe, dass wir Intelligenz sehr genau, ja geradezu perfekt, erfassen könnten. Wie sonst könnten wir den numerischen Wert so genau ausdrücken? Dies ist jedoch ein Irrglaube, da psychologisch-diagnostische Verfahren immer auch mit einem Messfehler behaftet sind . . . . Vereinfachend lässt sich sagen, dass jede Person eine tatsächliche Ausprägung (wahrer Wert, T) auf der zu messenden Dimension besitzt. Allerdings kann die Messung dieses Wertes durch unsystematische Einflüsse (Messfehler, E) wie beispielsweise Ermüdung verzerrt sein. Würden wir eine Person mit demselben Verfahren unendlich oft messen können, ergäbe sich hypothetisch eine Normalverteilung der Messwerte dieser Person mit dem wahren Wert als Mittelwert. Die Klassische Testtheorie (siehe Bühner, 2010) beschäftigt sich ausführlich mit dieser Thematik.

5 Konfidenzintervall-Rechner

Der Konfidenzintervall-Rechner ist ein Programm, was die Berechnung und Veranschaulichung von Konfidenzintervallen ermöglicht. Das Programm wurde vollständig in Python 3.6.2 geschrieben (Foundation, 2017). Zur Berechnung des Konfidenzintervalls müssen sechs Parameter eingegeben werden:

1. Reliabilitätskoeffizient
2. Standardabweichung des Normwertes
3. Normmittelwert
4. Sicherheitswahrscheinlichkeit (99%, 95%, 90%, 80%)
5. Seitigkeit (einseitiges oder zweiseitiges Testen)
6. Hypothese (Äquivalenzhypothese, Regression zur Mitte)

Der Konfidenzintervall-Rechner soll ein Hilfsmittel für Sachverständige sein, um softwaregestützt Konfidenzintervalle berechnen zu können. Er kann außerdem bei der diagnostischen Auswertung genutzt werden, da die graphische Veranschaulichung automatisch zeigt, in welchem Bereich das Konfidenzintervall liegt (durchschnittlich, über- oder unterdurchschnittlich, weit über- oder unterdurchschnittlich). Diese Anleitung soll das notwendige Wissen liefern, um die Funktionen des Konfidenzintervall-Rechners verstehen und nutzen zu können.

6 Berechnung des Konfidenzintervalls

Die allgemeine Gleichung für Konfidenzintervalle (KI) lautet:

$$KI = Normwert \pm SD * \sqrt{(1 - Reliabilität)} * z \quad (1)$$

Dabei ist der Normwert der normierte Testwert der Person. Er kann den Normtabellen im Testmanual entnommen werden. Den Term $SD * (1 - Reliabilität)$ bezeichnet man als *Standardmessfehler*. Daraus folgt:

$$KI = Normwert \pm Standardmessfehler * z \quad (2)$$

Die im Standardmessfehler enthaltene Standardabweichung *SD* ergibt sich logisch aus der verwendeten Normwertskala (z.B. SD_z = 1, SD_{IQ} = 15, SD_T = 10, SD_{ST} = 2). Sowohl die Standardabweichung als auch der Normwert stehen daher schon fest und müssen nicht durch die oder den Sachverständige/n definiert werden.

Entscheidungsspielraum gibt es bezüglich der Reliabilität und des z-Wertes. Hier müssen vier Entscheidungen getroffen werden, die Auswirkungen auf die Höhe der beiden Werte haben:

1. Welcher *Reliabilitätskoeffizient* wird benutzt?
2. Soll der Nomwert zur Mitte korrigiert werden? In diesem Fall wird eine Korrektur des Normwertes durchgeführt und der *Standardmessfehler* durch den *Standardschätzfehler* ersetzt.
3. Welche *Sicherheitswahrscheinlichkeit* wird zugrunde gelegt? Dies hat Auswirkungen auf den z-Wert.
4. Wird *ein- oder zweiseitig* getestet? Dies wirkt sich ebenfalls ebenfalls auf den z-Wert aus.

Jede dieser Entscheidungen wirkt sich auf die Breite des Konfidenzintervalls aus. Dabei gibt es keine verbindlichen Regelungen. Es liegt in der Verantwortung der Sachverständigen, wie die vorliegende Fragestellung am besten beantwortet werden kann. Im Folgenden wird jede dieser Entscheidungen erläutert.

7 Parameter

7.1 Reliabilitätskoeffizient

Testmanuale sollten immer Angaben zur Reliabilität enthalten. Meistens handelt es sich hierbei um Schätzer der internen Konsistenz (häufig Cronbach's α) oder der Retest-Reliabilität. Die interne Konsistenz ist, vereinfacht gesprochen, die durchschnittliche Korrelation der Items zu einem Messzeitpunkt. Die Retest-Reliabilität wiederum ist die Korrelation der Items zwischen zwei Messzeitpunkten in einem bestimmten Zeitraum. Sowohl die interne Konsistenz als auch die Retest-Reliabilität sind für die Beurteilung der Testgüte von zentraler Bedeutung. Sachverständige müssen entscheiden, welchen der beiden Koeffizienten sie zur Berechnung der Konfidenzintervalle verwendet. Ziegler und Bühner (2012) empfehlen, immer dann Maße der internen Konsistenz als Reliabilitätsschätzer zu verwenden, wenn eine Statusdiagnostik durchgeführt werden soll und Implikationen der Testung auf spätere Zeitpunkte eher irrelevant sind. Liegt eine prognostische Fragstellung vor, empfehlen Ziegler und Bühner (2012) die Verwendung der Restest-Reliabilität.

7.1.1 Interne Konsistenz

In Bezug auf die interne Konsistenz ist zu beachten, dass insbesondere Cronbach's α auch von anderen Faktoren, wie zum Beispiel der Zahl der Items abhängt (Cortina, 1993). Zudem fällt die interne Konsistenz in manchen Verfahren relativ niedrig aus, da manche Konstrukte bewusst breit gefasst sind, sodass die Items verhältnismäßig niedrig korrelieren. Sachverständigen sollte daher bewusst sein, dass Konfidenzintervalle umso größer ausfallen, je breiter das gemessene Konstrukt definiert wird und je weniger Items in der Skala enthalten sind. Andersherum ausgesprochen: Eine niedrige interne Konsistenz muss nicht notwendigerweise für eine „schlechte" Verlässlichkeit des Instruments stehen, solange die Höhe der internen Konsistenz in einem inhaltlich sinnvollen Zusammenhang mit der Breite des gemessenen Konstrukts und der Anzahl der Items steht.

7.1.2 Retest-Reliabilität

Auch die Höhe der Retest-Reliabilität sollte in einem sinnvollen Zusammenhang mit dem Konstrukt stehen, was durch das Instrument erfasst werden soll. Es kann argumentiert werden, dass auch bei nicht explizit prognostischen Fragestellungen eine ausreichend hohe Retest-Reliabilität gewährleistet sein sollte. Persönlichkeitsfragebögen bauen z.B. alle mehr oder weniger auf dem zugrundeliegenden Paradigma auf, dass es sich bei Persönlichkeitseigenschaften um zeitlich mittelfristig stabile Konstrukte handelt (Asendorpf & Neyer, 2012). Auch diese Fragebögen sollten daher eine ausreichend hohe Retest-Reliabilität haben, obwohl sie im Kern der Beantwortung einer statusdiagnostischen Frage dienen. Oder anders herum formuliert: Wenn das Instrument nicht gerade explizit sogenannte *State*-Konstrukte wie zum Beispiel Angst (Cattell & Scheier, 1961) erfasst, bei denen eine niedrige Restest-Reliabiltität inhaltlich begründet ist, stellt eine ausreichend hohe Retest-Reliabilität im Hinblick auf die allgemeine Verlässlichkeit des Instruments in der Regel eine *notwendige aber nicht hinreichende Bedingung* dar.

7.1.3 Praxis

In der Praxis wird im Zuge der Testkonstruktion zur Berechnung der Reliabilität häufig die interne Konsistenz verwendet, da sie mit geringerem Aufwand berechnet werden kann (es bedarf nicht zweier Messzeitpunkte). Sie ist daher öfter in Manualen zu finden. Im Hinblick auf den vorherigen Absatz soll an dieser Stelle allerdings darauf hingewiesen werden, dass es aus Sichtweise der Sachverständigen wünschenswert ist, bei der Testkonstruktion auch die Retest-Reliabilität zu berechnen und diese anschließend im Manual zu berichten.

7.1.4 Konfirmatorisches Hypothesentesten

Im Idealfall haben Sachverständige bereits vor der Testung ihre Entscheidungsstrategien festgelegt und können somit die Wahl des Tests daran ausrichten, ob die erwünschte Reliabilität angegeben und zufriedenstellend ist. Dies entspricht auch einer wissenschaftlich korrekten Herangehensweise, da alle Entscheidungen zur Auswertung des

Tests *a priori* getroffen werden sollen. Erst während der Berechnung des Konfidenzintervalls Entscheidungen über den Reliabilitätskoeffizienten und andere Parameter zu treffen, sollte nach Möglichkeit vermieden werden, da dies zu *konfirmatorischen Hypothesentesten* führen kann.

7.2 Art des Konfidenzintervalls

7.2.1 Äquivalenzhypothese

Für den Fall, dass angenommen werden kann, dass der Testwert ein guter Schätzer des wahren Wertes einer Person ist, wird unter Annahne der sogenannten *Äquivalenzhypothese* getestet. Das Konfidenzintervall wird dann wie in der ersten Formel berechnet und es wird der *Standardmessfehler* verwendet. Da die Reliabilität schon in Abschnitt 7.1 definiert wurde, muss unter der Äquivalenzhypothese also nur noch der z-Wert (siehe 7.2.5) definiert und abgelesen werden.

7.2.2 Regressionshypothese

Individuelle Testwerte einer Person schwanken aufgrund des Messfehlers um den wahren Wert der Person. Manche Testergebnisse geben Anlass zu der Vermutung, dass es sich bei dem gemessenen Testwert um einen Extremwert handelt und somit um keine gute Schätzung des wahren Wertes. Die lässt sich z.B. aus Vorinformationen schließen oder aus dem klinischen Eindruck während der Begutachtung. In diesem Fall kann bzw. sollte das Konfidenzintervall unter der Regressionshypotheseberechnet werden. Dabei wird der individuelle Normwert des Probanden zunächst zur Mitte korrigiert, bevor das Konfidenzintervall berechnet wird. Außerdem wird der Berechnung des Konfidenzintervalls der *Standardschätzfehler* und nicht der *Standardmessfehler* verwendet. Der Standardschätzfehler schätzt die zu erwartende Schwankung der Mittelwerte, die man erhalten würde, wenn mehrmals Stichproben aus der Gesamtpopulation gezogen würden. Für eine ausführlichere Darstellung und Erläuterung der folgenden drei Formeln sei auf Bühner (2010, Kap. 4.8) verwiesen.

Die Formel zur Berechnung des korrigierten Normwertes lautet:

$$Normwert_{korrigiert} = Reliabilität * Normwert_{beobachtet} + Normmittelwert * (1 - Reliabilität) \quad (3)$$

Da die Reliabilität schon in Abschnitt 7.1 bestimmt wurde, ist hier lediglich der *Normmittelwert* neu. Dieser ergibt sich genau wie die Standardabweichung logisch aus der verwendeten Normierung und ist daher schon festgelegt (z.B. M_z = 0, M_{IQ} = 100, M_T = 50, M_{ST} = 5). Die Formel zur Berechnung des Standardschätzfehlers lautet:

$$Standardschätzfehler = SD * \sqrt{Reliabilität * (1 - Reliabilität)} \quad (4)$$

Daraus folgt für die Berechnung des Konfidenzintervals unter Annahme der Regressionshypothese:

$$KI_{korrigiert} = Normwert_{korrigiert} \pm Standardschätzfehler * z \quad (5)$$

Nachdem der korrigierte Normwert und der Standardschätzfehler berechnet wurden, muss also genau wie im Fall der Äquivalenzhypothese noch der z-Wert bestimmt werden. Hierfür sind zwei letzte Entscheidungen notwendig: die *Sicherheitswahrscheinlichkeit* muss definiert und die *Seitigkeit* festgelegt werden.

7.2.3 Sicherheitswahrscheinlichkeit

Prinzipiell gilt, dass Konfidenzintervalle immer breiter werden, je höher die Sicherheitswahrscheinlichkeit ausfällt. Eine sehr hohe Sicherheitswahrscheinlichkeit wird somit immer mit relativ ungenauen Schätzungen „erkauft", wohingegen eine niedrige Sicherheitswahrscheinlichkeit das Risiko erhöht, dass der wahre Wert einer Person außerhalb eines verhältnismäßig engen Intervalls liegt. Der sogenannte *Fehler 1. Art* bzw. α-Fehler bestimmt die Höhe der Sicherheitswahrscheinlichkeit.

$$Sicherheitswahrscheinlichkeit = 1 - \alpha \quad (6)$$

Der oder die Sachverständige muss abwägen, welche Risiken mit einer jeweiligen Fehlentscheidung verbunden wäre. Dies soll kurz an zwei Beispielen erläutert werden:

Beispiel 1: Bei einem 9-jährigen Kind besteht der Verdacht auf eine Hochbegabung, weswegen eine Intelligenztest durchgeführt wird. Ein hoher Wert in diesem Test kann mit der Empfehlung verbunden sein, einen Schulwechsel durchzuführen, um eine bessere Förderung zu ermöglichen. Abgesehen davon, dass Schulwechsel im Kindesalter an sich schon eine Anforderung an das Kind darstellen (neue Umgebung, Verlust alter Freundschaften, etc.), kann ein überschätzter IQ-Wert zu einer Leistungsüberforderung des Kindes und somit zu einer Verschlechterung seiner Situation führen. Die oder der Sachverständige könnte bei der Berechnung der Konfidenzintervalle also eine hohe Sicherheitswahrscheinlichkeit, z.B. 95%, zugrunde legen. Damit wäre zwar die Gefahr verbunden, dass das Konfidenzintervall so breit gerät, dass nicht eindeutig eine Hochbegabung diagnostiziert werden kann, diesem Problem könnte jedoch mit einer mehrfachen Testung entgegen gewirkt werden (sog. *Aggregationsprinzip*).

Beispiel 2: Anders verhält es sich z.B. bei Persönlichkeitstestungen im Rahmen von *Glaubhaftigkeitsbegutachtungen*. In der Regel besitzen diese Verfahren eher informatorischen Wert für die oder den Sachverständigen. Ein möglicher Zusammenhang zwischen bestimmten Persönlichkeitseigenschaften und der Aussagequalität wurde bislang empirisch nicht insoweit belegt, als dass ein Rückgriff auf individuelle Testwerte eine erhebliche Relevanz für die Gesamtbegutachtung hätte. Insofern wäre hier eine niedrigere Sicherheitswahrscheinlichkeit, z.B. 80%, angemessen, da für den Fall, dass der wahre Wert eines Probanden außerhalb des errechneten Konfidenzintervalles liegt, kaum größerer Schaden entstehen würde. Ziegler und Bühner (2012) empfehlen generell eine Sicherheitswahrscheinlichkeit von 80% bei Persönlichkeitstests.

7.2.4 Seitigkeit

Ungerichtete Hypothesen (”Wie fällt der Testwert von Person x im Vergleich zur Normstichprobe aus?”) erfordern einen zweiseitigen Test, gerichtete Hypothesen wie in Beispiel 1 (”Hat das Kind einen IQ von über 130 und ist damit hochbegabt?”) einen einseitigen Test. Für die Beantwortung von gerichteten Hypothesen ist nur eine Seite des Konfidenzintervalles relevant, bei ungerichteten Hypothesen sind beide Seiten des Kon-

fidenzintervalls von Interesse. Die oder der Sachverständige muss dies berücksichtigen, indem sie oder er im Falle einer einseitigen Testung eine Sicherheitswahrscheinlichkeit von $1 - \alpha$ verwendet und im Falle einer zweiseitigen Testung eine Sicherheitswahrscheinlichkeit von $1 - \frac{\alpha}{2}$. Dadurch das im Falle der zweiseitigen Testung α durch 2 geteilt wird, fällt das Konfidenzintervall größer aus als bei der einseitigen Testung. Soll eine einseitige Hypothese überprüft werden, ist außerdem die Richtung entscheidend, nach der abgesichert werden soll. Zur Beantwortung der Frage wie in Beispiel 1 muss beispielsweise streng genommen nur die untere Grenze des Konfidenzintervals berechnet und veranschaulicht werden, die obere Grenze ist hier irrelevant für die Fragestellung.

7.2.5 z-Wert

Der z-Wert wird für das jeweilige α und die daraus resultierende Sicherheitswahrscheinlichkeit aus einer beliebigen z-Wert-Tabelle (z.B. Tabelle 1) abgelesen. Noch einfacher ist es, ein Programm zu verwenden, welches z-Werte auf der Basis der Wahrscheinlichkeiten ausrechnet (und umgekehrt) und dabei berücksichtigen kann, ob ein- oder zweiseitig getestet werden soll (z.B.: http://eswf.uni-koeln.de/glossar/surfstat/normal.htm.). Auch der Konfidenzintervall-Rechner bestimmt nach Eingabe der Sicherheitswahrscheinlichkeit und der gewünschten Seitigkeit automatisch den richtigen z-Wert. Wie man sieht, ist der z-Wert bei der einseitigen Testung immer kleiner als bei der zweiseitigen Testung. Außerdem werden die z-Werte immer größer, umso eine höhere Sicherheitswahrscheinlichkeit festgelegt wird.

Sicherheitswahrscheinlichkeit	z-Wert (einseitig)	z-Wert (zweiseitig)
99%	2.33	2.58
95%	1.64	1.96
90%	1.28	1.64
80%	0.84	1.28

Tabelle 1: Gängige Sicherheitsbereiche und die jeweiligen ein- und zweiseitigen z-Werte

8 Abschließende Bemerkung

Nicht immer ist die eigene Berechnung von Konfidenzintervallen notwendig, da diese in vielen Testmanualen bereits enthalten sind. Allerdings sollten Sachverständige auch in diesem Fall prüfen, ob die im Testmanual verwendeten Parameter zur Berechnung der Konfidenzintervalle mit den eigenen, auf den jeweiligen Begutachtungsfall zugeschnittenen Entscheidungen korrespondieren. Für den WISC-V wurden z.B. Konfidenzintervalle unter der Regressionshypothese mit Reliabilitätsschätzern der internen Konsistenz und einer Sicherheitswahrscheinlichkeit von 90% und 95% berechnet. Inwieweit dies für die jeweilige Fragestellung passend ist, bleibt Entscheidung der Sachverständigen. Im Zweifel können eigene Konfidenzintervalle berechnet werden. In jedem Fall sollten Konfidenzintervalle nicht einfach aus einem Manual übernommen werden, sondern inhaltlich und untersuchungsökonomisch begründet werden im Sinne der obigen Entscheidungen.

Literatur

Asendorpf, J. B. & Neyer, F. J. (2012). Sechs Paradigmen der Persönlichkeitspsychologie: Eigenschaftsparadigma. In J. B. Asendorpf & F. J. Neyer (Hrsg.), *Psychologie der Persönlichkeit* (S. 23–32). Berlin, Heidelberg: Springer Berlin Heidelberg.

Bühner, M. (2010). *Einführung in die Test- und Fragebogenkonstruktion* (3. Aufl.). München: Pearson.

Cattell, R. B. & Scheier, I. H. (1961). *The Meaning and Measurement of Neuroticism and Anxiety*. New York: Ronald.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, *78* (1), 98–104. doi: 10.1037/0021-9010.78.1.98

Foundation, P. S. (2017). *Python.* Zugriff auf `https://www.python.org`

Ziegler, M. & Bühner, M. (2012). *Grundlagen der Psychologischen Diagnostik*. Wiesbaden: VS Verlag für Sozialwissenschaften. doi: 10.1007/978-3-531-93423-5