

Konfidenzintervalle als Alternative zu Signifikanztests

Eduard Brandstätter¹
Johannes Kepler Universität Linz

Zusammenfassung

Die Arbeit argumentiert, Signifikanztests durch Konfidenzintervalle zu ersetzen. Richtig interpretiert, vermeiden Konfidenzintervalle die Probleme klassischer Signifikanztests, sind logisch korrekt, benötigen weder a-priori Hypothesen noch bringen sie triviale Aussagen hervor. Der erste Teil präsentiert Kritik zu klassischen Signifikanztests, der zweite Teil weist Kritik an der Verwendung von Konfidenzintervallen zurück und zeigt deren Vorteile gegenüber Signifikanztests auf.

Einführung

Statistische Signifikanztests haben sich als fester Bestandteil des psychologischen Methodenrepertoires etabliert. Ungeachtet des häufigen Gebrauchs wurde in letzter Zeit vermehrt Kritik an Signifikanztests laut und löste lebhaft Diskussionen aus (siehe Baril & Cannon, 1995; Cohen, 1994; Cohen, 1995; Cortina, 1997; Frick, 1995; Frick, 1996; Gigerenzer, 1993; Gigerenzer et al., 1989; Hagen, 1997; Harlow, Mulaik, & Steiger, 1997; Hubbard, 1995; Kleiter, 1969; McCraw, 1995; Parker, 1995; Schmidt, 1996; Sedlmeier, 1996; Svyantek & Ekeberg, 1995). Um die Probleme von Signifikanztests zu umgehen, wurden verschiedene Verfahren wie graphische Datenanalysen (Cohen, 1994; Tukey, 1977), Metaanalysen (Schmidt, 1996), Replikationen von Untersuchungen und Konfidenzintervalle (Cohen, 1994;

¹ Kontaktadresse:

Johannes-Kepler-Universität; Institut für Pädagogik und Psychologie; Tel: ++43/+732/2468/578 - Fax: /228; mail: e.brandstaetter@jk.uni-linz.ac.at; Altenbergerstr. 69; A-4040 Linz, Österreich

Sedlmeier, 1996) als Alternativen zu Signifikanztests vorgeschlagen. Während die ersten drei Verfahren auf wenig Gegenkritik in der Literatur stießen, lösten Konfidenzintervalle als Ersatz für Signifikanztests sowohl Gegenkritik (z. B. Frick, 1996; Hagen, 1997), als auch Zustimmung (z. B. Schmidt, & Hunter, 1997; Steiger & Fouladi, 1997) aus. Konfidenzintervalle unterlägen – so die Kritiker – den gleichen logischen Fehlinterpretationen, welche auch für Signifikanztests zuträfen. Somit könnten Konfidenzintervalle die Probleme von Signifikanztests nicht lösen. Die Befürworter wiederum betonen, daß Konfidenzintervalle keine a-priori Hypothesen benötigten und somit die Prüfung trivialer Nullhypothesen vermieden.

Die vorliegende Arbeit versucht die verschiedenen pro und kontra Standpunkte in bezug auf Konfidenzintervalle auf ihre Validität zu prüfen. Die Frage, ob Konfidenzintervalle einen geeigneten Ersatz für Signifikanztests darstellen ist entscheidend, da ein unangemessener Gebrauch von Konfidenzintervallen zukünftiger Forschung mehr schaden als nutzen würde. Andererseits würde ein unbegründeter Verzicht auf Konfidenzintervalle oder die Abschaffung von Signifikanztests das Methodenrepertoire der Wissenschaft unnötig einengen.

Im ersten Teil der Arbeit liste ich die wichtigsten Kritikpunkte an Signifikanztests auf. Diese Auflistung erhebt keinen Anspruch auf Vollständigkeit (siehe dazu z. B. Harlow et al. 1997). Im zweiten Teil gehe ich auf Konfidenzintervalle als Alternative zu Signifikanztests ein. Darin erörtere ich (i) die Interpretation, (ii) Kritik und Mißverständnisse an, sowie (iii) Widerstände gegen Konfidenzintervalle. Weiters behandle ich (iv) die Frage, ob Konfidenzintervalle Signifikanztests ersetzen sollen und können, und schließlich weise ich (v) auf die Gefahr möglicher Fehlinterpretationen von Effektstärken (als die Zentren von Konfidenzintervallen) hin.

Kritik an Signifikanztests

Signifikanztests liefern – so die Kritik – (1) irrelevante Informationen und beruhen (2) auf trivialen Nullhypothesen. Beide Kritikpunkte werden nun erörtert.

(1) Signifikanztests liefern irrelevante Informationen

Eine Grundregel richtigen logischen Schließens ist der modus tollens. Nach dem modus tollens folgt aus der Schlußfolgerung Wenn A dann B die Schlußfolgerung „Wenn Non-B dann Non-A“. Falsch hingegen ist die Schlußfolgerung „Wenn B dann A“. Dazu folgendes Beispiel:

Eine Person die in Lichtenstein wohnt, wohnt auch in Europa (Wenn A dann B).

Daraus folgt:

Eine Person die nicht in Europa wohnt, wohnt nicht in Lichtenstein (Wenn Non-B dann Non-A)

Während die Schlußfolgerung

Eine Person die in Europa wohnt, wohnt auch in Lichtenstein (Wenn B dann A)

logisch falsch ist. Cohen (1994) zeigte, daß der modus tollens invalide ist, wenn eine sichere wenn-dann-Abfolge durch eine wahrscheinliche ersetzt wird. Zum Beispiel:

Wenn man würfelt, erscheint *wahrscheinlich* eine Zahl größer als 1 (Wenn A dann wahrscheinlich B)

Die Folgerung

Wenn die Zahl 1 erschienen ist, hat man wahrscheinlich nicht gewürfelt (Wenn Non-B dann wahrscheinlich Non-A)

ist jedoch logisch inkorrekt. Genau dieses Muster scheint aber irrtümlicherweise die häufigste Interpretation eines Ergebnisses eines Signifikanztests zu sein (Cohen, 1994), wie die folgende Schlußfolgerung verdeutlicht:

Wenn die H_0 wahr ist, ist diese Differenz der Stichprobenmittelwerte (Signifikanz) unwahrscheinlich.

Diese Differenz ist aufgetreten.

Daher ist die H_0 wahrscheinlich falsch.

Zur Erinnerung, die Irrtumswahrscheinlichkeit alpha gibt – richtig interpretiert – die Wahrscheinlichkeit an, mit der unter Gültigkeit der H_0 ein bestimmtes oder extremeres Ergebnis (z.B. eine bestimmte oder extremere Differenz zwischen zwei Stichprobenmittelwerten) auftritt, also p (Ergebnis / H_0)¹. Signifikanztests sagen somit nichts über die Wahrscheinlichkeit der Gültigkeit der H_0 aus. Somit testen Signifikanztests strenggenommen auch überhaupt keine Hypothesen. Zusammengefaßt, Signifikanztests geben Auskunft über die Wahrscheinlichkeit, ein bestimmtes oder extremeres Ergebnis unter Gültigkeit der Nullhypothese zu erlangen. Über die Wahrscheinlichkeit der Gültigkeit einer Nullhypothese sagen Signifikanztests nichts aus. Dem nächsten Kritikpunkt zufolge sei dies aber ohnehin

¹ Mit „Ergebnis“ sind sowohl das Stichproben-, als auch alle extremeren Ergebnisse gemeint

nicht sehr relevant, da Nullhypothesen nur triviale Sachverhalte ausdrücken würden.

(2) Signifikanztests beruhen auf trivialen Nullhypothesen

Entsprechend dieser Kritik seien Nullhypothesen an sich irrelevant, da zwischen zwei Mittelwerten praktisch immer irgendwelche Unterschiede bestünden. Anschaulich legt dies eine Untersuchung von Bakan (1966) dar. Bakan unterteilte 60,000 Personen nach zufälligen Kriterien wie etwa, ob der Wohnort östlich oder westlich des Mississippi-Flusses lag und fand in allen Fragen des Fragebogens signifikante Unterschiede. Tukey (1991) faßt die Kritik dazu prägnant zusammen: „It is foolish to ask 'Are the effects of A and B different?' They are always different – for some decimal place., (S. 100). Nahezu ebenso eindeutig äußert sich Nunnally (1960, zit. n. Gigerenzer et al., 1989, S. 210): „If the null hypothesis is not rejected, it usually is because the N [sample size] is too small. If enough data is gathered, the hypothesis will generally be rejected. If rejection of the null hypothesis were the real intention in psychological experiments, there usually would be no need to gather data“.

Diese Kritik geht noch einen Schritt weiter. Demnach sei es eigentlich egal, ob Signifikanztests richtig oder falsch interpretiert würden, da die H_0 in wissenschaftlicher Hinsicht uninteressant wäre. Der Verzicht auf Signifikanztests würde demnach keinerlei Verlust bedeuten.

Als Ausweg aus diesem Dilemma berücksichtigt der erweiterte Ansatz des klassischen Signifikanztests neben einer Nullhypothese H_0 eine zusätzliche Alternativhypothese H_1 . Zur Erinnerung, der am öftesten eingesetzte zweiseitige Signifikanztest geht lediglich von einer Nullhypothese H_0 aus, während die Alternativhypothese H_1 den Rest aller anderen Hypothesen bildet. Der erweiterte Ansatz legt neben der Nullhypothese eine spezifische Alternativhypothese H_1 fest, welche die Bestimmung der Teststärke $(1 - \beta)$ und der notwendigen Stichprobenumfänge zur Planung einer Untersuchung erlauben. Aber selbst dieser erweiterter Ansatz ist mit Problemen verbunden: Erstens läßt sich in der konkreten Forschungspraxis – zumindest in der Psychologie – selten eine genaue, inhaltlich begründete Alternativhypothese angeben. Ob beispielsweise ein Korrelationskoeffizient $\rho = .3$ oder $\rho = .4$ ist, läßt sich a-priori schwer festlegen. Auch wenn Schätzungen für ρ aus bisherigen Studien vorliegen, bleibt dieser Vorgang immer mit einer gewissen Willkür behaftet. Somit folgt, daß Hypothesentestungen ein in-

haltliches Problem darstellen, welches – auf den Punkt gebracht – in der Festlegung einer spezifischen, inhaltlich begründeten Alternativhypothese besteht.

Zweitens, und noch wichtiger, die Zerlegung aller möglichen Hypothesen in eine H_0 und eine spezifische H_1 löst das Problem der Verletzung des modus tollens nicht. Zur Erinnerung, der modus tollens wird invalide, wenn eine sichere „wenn-dann-Abfolge“ durch eine wahrscheinliche ersetzt wird. Der erweiterte Ansatz des klassischen Signifikanztests legt jedoch ebenfalls eine Fehlinterpretation eines signifikanten Ergebnisses nahe: Von den Wahrscheinlichkeiten des Auftretens eines bestimmten oder extremeren Ergebnisses (α und β Fehler) wird unter Annahme einer Null- und Alternativhypothese irrtümlicherweise auf die Gültigkeit dieser Hypothesen geschlossen. Auch hier gilt analog: Signifikanztests testen keine Hypothesen sondern geben Auskunft über die Wahrscheinlichkeiten, ein bestimmtes oder extremeres Ergebnis unter Gültigkeit einer Null-, als auch einer Alternativhypothese zu erlangen. Über die Wahrscheinlichkeiten der Gültigkeit dieser Hypothesen erfährt man nichts. Zusammengefaßt, die Formulierung einer spezifischen Alternativhypothese H_1 kann die anstehenden Probleme ebenfalls nicht zufriedenstellend lösen und ist mit zwei Nachteilen behaftet: Erstens fehlt häufig eine solide inhaltliche Begründung für die Festlegung einer spezifischen H_1 , zweitens erfährt man nichts über die Gültigkeit der H_0 und der H_1 . Der nächste Abschnitt versucht einen möglichen Ausweg aus diesem Dilemma zu geben.

Konfidenzintervalle als Alternative

Die Interpretation von Konfidenzintervallen

Als Lösung aus dem Dilemma schlug Cohen (1994) Konfidenzintervalle vor. Was aber besagen Konfidenzintervalle – etwa in bezug auf die Differenz zweier Populationsmittelwerte? Repliziert man eine Studie unendlich oft und berechnet jedesmal ein 95% Konfidenzintervall um die empirisch ermittelte Mittelwertsdifferenz, liegt die wahre Mittelwertsdifferenz der Population in 95% aller Replikationen innerhalb dieser Intervalle (Bleymüller, Gehlert, Gülicher, 1988; Cohen, 1995). In 5% aller Replikationen liegt demnach die wahre Populationsmittelwertsdifferenz außerhalb dieser Intervalle. Konfidenzintervalle sind somit Zufallsvariablen. Die Größe und Lage dieser Konfidenzintervalle wird somit von Replikation zu Replikation variieren.

Um bei dem erwähnten Beispiel zu bleiben. Ein Forscher interessiert sich für die wahre Differenz zweier Populationsmittelwerte. Dazu zieht er zwei hinreichend große ($n_1, n_2 > 30$), unabhängige Stichproben mit den Mittelwerten \bar{x}_1, \bar{x}_2 und den Streuungen s_1 und s_2 . Ziel ist nun, ein Konfidenzintervall für die Differenz der wahren Populationsmittelwerte ($\mu_1 - \mu_2$) anzugeben. Da – allgemein gesprochen – ein wahrer Populationsparameter aufgrund von Stichprobendaten geschätzt werden soll, spricht man auch von einem *Schätzintervall* (Menges, 1969; Witte, 1980) für den entsprechenden wahren Populationsparameter.² Konkret berechnet sich für die gesuchte wahre Differenz der Populationsmittelwerte das 95% Schätzintervall:

$$I_{\text{Schätz}} \equiv [(\bar{x}_1 - \bar{x}_2) - 1.96 * \sigma_{\text{Diff}}, (\bar{x}_1 - \bar{x}_2) + 1.96 * \sigma_{\text{Diff}}], \text{ oder}$$

$$P[(\bar{x}_1 - \bar{x}_2) - 1.96 * \sigma_{\text{Diff}} \leq (\mu_1 - \mu_2) \leq (\bar{x}_1 - \bar{x}_2) + 1.96 * \sigma_{\text{Diff}}] = .95$$

mit $\sigma_{\text{Diff}} = (\sigma_1^2/n_1 + \sigma_2^2/n_2)^{1/2}$; σ_1 und σ_2 werden im Regelfall durch die Stichprobenstreuungen geschätzt. Die Maßzahl $1 - \alpha$ wird als Konfidenzkoeffizient, α als Konfidenzniveau bezeichnet.

Forscher sind aber im allgemeinen nicht am Prozentsatz der Konfidenzintervalle interessiert, welche den wahren Populationsparameter enthalten sondern wollen wissen, mit welcher Wahrscheinlichkeit sich der wahre Populationsparameter in dem Konfidenzintervall der eben gezogenen Stichprobe befindet. Mit anderen Worten, eine praktisch tätige Forscherin fragt nach der Wahrscheinlichkeit des *singulären* Ereignisses, daß sich der wahre Populationsparameter in einem bestimmten Konfidenzintervall, berechnet aus den Daten einer Stichprobe, befindet.

An diesem Punkt wird unsere praktisch tätige Forscherin von der Wissenschaft leider im Stich gelassen, da über die Zuordnung von einer Wahrscheinlichkeit zu einem *einzelnen* Ereignis in der Statistik zwei verschiedene Lehrmeinungen exi-

² Vom Schätzintervall für wahre Populationswerte ist das *Prognoseintervall* für Stichprobenrealisationen zu unterscheiden:

$$I_{\text{Prognose}} \equiv [(\mu_1 - \mu_2) - 1.96 * \sigma_{\text{Diff}}, (\mu_1 - \mu_2) + 1.96 * \sigma_{\text{Diff}}], \text{ mit}$$

$$P[(\mu_1 - \mu_2) - 1.96 * \sigma_{\text{Diff}} \leq (\bar{x}_1 - \bar{x}_2) \leq (\mu_1 - \mu_2) + 1.96 * \sigma_{\text{Diff}}] = .95$$

Das Prognoseintervall geht von einer Hypothese ($\mu_1 - \mu_2$) aus. Liegt eine Stichprobenrealisation ($\bar{x}_1 - \bar{x}_2$) außerhalb des Prognoseintervalls, weichen die Daten signifikant von der Hypothese ab. Diese Strategie entspricht jener des Signifikanztests. Anstatt der Berechnung des Prognoseintervalls hat sich durch die Computerisierung die genaue Berechnung von p -Werten durchgesetzt. Im Gegensatz zum Schätzintervall, welches eine Zufallsvariable ist, handelt es sich beim Prognoseintervall um einen festen, nicht-zufälligen Konfidenzbereich (sofern σ_1 und σ_2 zur Berechnung von σ_{Diff} bekannt sind).

stieren. Während Anhänger der „Klassischen Wahrscheinlichkeitstheorie“ („frequentists“) die Zuordnung einer Wahrscheinlichkeit zu einem Einzelereignis rigoros auf zwei Werte einschränken (Kendall & Stuart, 1979; Mulaik, Raju & Harshman, 1997), gestatten die Anhänger der „Subjektiven Wahrscheinlichkeitstheorie“ (DeFinetti, 1971; Wright & Ayton, 1994) die Zuordnung aller möglichen Wahrscheinlichkeiten zwischen 0 bis 1 (siehe Reichardt & Gollob, 1997). Die Klassische Wahrscheinlichkeitstheorie definiert den Wahrscheinlichkeitsbegriff auf der Basis wiederholbarer Ereignisse. Demnach ist Wahrscheinlichkeit die (asymptotische) relative Häufigkeit eines Ereignisses, welches unendlich oft wiederholt wird – und zwar unter identischen Bedingungen, welche sich nur durch den Zufall unterscheiden (Reichardt & Gollob, 1997). Für die Zuordnung einer Wahrscheinlichkeit zu einem Einzelereignis sind in diesem Konzept nur zwei Werte vorgesehen. Ein einzelnes Konfidenzintervall kann den wahren Populationsparameter entweder beinhalten oder nicht, womit die Wahrscheinlichkeit dieses Einzelereignisses nur 0 oder 1 betragen kann.

Liberaler dagegen die Auslegung der Subjektiven Wahrscheinlichkeitstheorie³. Wenn sich in 95% aller Ziehungen der wahre Populationsparameter innerhalb der berechneten Konfidenzintervalle befindet, ist die Wahrscheinlichkeit p , daß sich der wahre Populationsparameter in dem 95% Konfidenzintervall der eben gezogenen Stichprobe befindet, gleich .95; lediglich mit einer Wahrscheinlichkeit $p = .05$ befindet sich der wahre Populationsparameter außerhalb des Konfidenzintervalls: Die Wahrscheinlichkeit für ein Einzelereignis kann somit alle Werte zwischen 0 bis 1 annehmen.

Hier ist nicht der Platz die Gegensätze beider Schulen zu überwinden (für eine ausführliche Diskussion siehe Stegmüller, 1973). Man sollte sich lediglich bewußt sein, daß beispielsweise eine Interpretation wie „Der wahre Populationsparameter befindet sich mit einer Wahrscheinlichkeit von $p = .95$ in diesem (einen) Konfidenzintervall“ nur einer Schule entspricht (Subjektive Wahrscheinlichkeitstheorie), und daß daneben auch noch eine andere Interpretationsmöglichkeit (Klassische Wahrscheinlichkeitstheorie) besteht.

Zwischen Konfidenzintervallen und Signifikanztests besteht jedoch ein wichtiger Unterschied: Bei Konfidenzintervallen erlaubt der Schluß auf eine einzelne Stichprobenziehung Aussagen über den interessierenden Populationsparameter

³ Da sowohl die Klassische wie die Subjektive Wahrscheinlichkeitstheorie das Bayes Theorem anerkennen, verzichte ich auf eine Gleichstellung von Subjektiver Wahrscheinlichkeitstheorie mit Bayes-Statistik (siehe auch Reichardt & Gollob, 1997).

(Genauigkeit der Schätzung des Effekts); Signifikanztests hingegen erlauben diesen Schluß auf den interessierenden Populationsparameter nicht. Man weiß nur, daß die Wahrscheinlichkeit, dieses oder ein extremeres Ergebnis unter Gültigkeit der H_0 erzielt zu haben, dem alpha-Fehler entspricht. Insofern liefert ein Schluß entsprechend der Subjektiven Wahrscheinlichkeitstheorie für Konfidenzintervalle ungleich mehr Informationen als für Signifikanztests.

Die eben gemachten Ausführungen gelten für den am häufigsten vorkommenden Fall, daß man keine exakte Vorstellung über die Priorverteilung des Populationsparameter hat (Reichardt & Gollob, 1997) – gleichzeitig eine Gleichverteilung aber ausgeschlossen werden kann (*no usable prior distribution*). Liegt hingegen ein Vorwissen (*nonuniform prior distribution*) über die Verteilung des Populationsparameters vor, kann das Bayes-Theorem angewendet werden (siehe Edwards, Lindmann, Savage, 1963; Kleiter, 1980; Winkler, 1972); dies gilt sowohl für Schlüsse die aus einem Konfidenzintervall, als auch für Schlüsse die aus einem Signifikanztest gezogen werden können.

Nach der Klärung möglicher Interpretationen von Konfidenzintervallen wenden wir uns nun der Kritik und den Mißverständnissen zu, welche Konfidenzintervalle hervorgerufen haben.

Kritik und Mißverständnisse

Wie oben bereits angesprochen, erfuhren auch Konfidenzintervalle einschlägige Kritik. Demnach wäre es unlogisch, Signifikanztests abzulehnen und gleichzeitig Konfidenzintervalle als Alternative zu empfehlen:

„... a confidence interval can function to indicate which values could not be rejected by a two-tailed test with alpha at .05. In this function, the confidence interval could replace the report of null hypothesis for just one value, instead of communicating the outcome of the tests of all values as null hypotheses ... Cohen (1994) was illogical when he criticized the logic of null hypothesis testing and then advocated using the confidence interval because it reported the results of all statistical tests“ (Frick, 1996, S. 383).

Ähnlich äußert sich Hagen (1997, S. 22): „We cannot escape the logic of NHST [null hypothesis statistical testing] by turning to point estimates and confidence intervals“.

Diese Kritik erscheint jedoch falsch, verwirrend und unbegründet. Sie erscheint deswegen falsch, weil Konfidenzintervalle zwar als Signifikanztests interpretiert

werden können, aber nicht müssen, wie Schmidt und Hunter (1997) ausführen: „The assumption underlying this objection is that because confidence intervals can be interpreted as significance tests, they *must* be so interpreted. But this is a false assumption“ (S. 50). Konfidenzintervalle beinhalten jedoch spezifische Informationen, welche Signifikanztests verschleiern.

Konfidenzintervalle zeigen anschaulich, wie genau die Schätzung eines unbekanntes Populationsparameters ausfällt, wobei kleine Konfidenzintervalle genauere Schätzungen zulassen als große. Effektstärken geben darüber hinaus wertvolle Hinweise über die Größe des interessierenden Effekts. Genau diese Informationen, wie die Genauigkeit der Parameterschätzung und die Größe des interessierenden Effekts, verbergen jedoch p -Werte.

Da Konfidenzintervalle leichter zu verstehen sind als Signifikanztests, besitzen sie gegenüber Signifikanztests einen entscheidenden didaktischen Vorteil. Wer schon einmal Statistik gelehrt hat weiß, um wieviel schneller Studierende Konfidenzintervalle im Vergleich zu Signifikanztests verstehen. Beinhaltet ein Konfidenzintervall den Wert 0, läßt sich nichts mit hoher Sicherheit über die Richtung eines Effekts aussagen: Der Effekt kann positiv, negativ, und, theoretisch zumindest, auch null sein; befindet sich der Wert 0 außerhalb des 95% Konfidenzintervalls, kennt man das Vorzeichen der wahrscheinlichsten (95% Vertrauen) der Populationsparameter. Jeder Studierende sieht dies sofort anschaulich ein. Um wieviel verdrehter dagegen die Logik eines Signifikanztests: Unter Annahme der H_0 ist die Wahrscheinlichkeit dieses oder ein extremeres Ergebnis zu erhalten gleich p . Und jetzt? Auch wenn Konfidenzintervalle als Signifikanztests interpretiert werden können, spricht wenig dafür, dies zu tun (Schmidt & Hunter, 1997). Zusammengefaßt, das obige Argument, daß Konfidenzintervalle als Signifikanztests interpretiert werden müssen ist unrichtig und geht am Kern der Sache vorbei.

Im Gegensatz zu Signifikanztests bringen Konfidenzintervalle die Genauigkeit einer Parameterschätzung ans Licht. Daß ein Populationsparameter mit nahezu völliger Sicherheit ungleich null ist, wird als gegeben vorausgesetzt, und steht daher nicht im Mittelpunkt des Interesses.

Wie erwähnt beinhalten Konfidenzintervalle mehr Informationen als Signifikanztests. Genau dieses Argument aber, daß Signifikanztests wegen ihres geringeren Informationsgehaltes in manchen Situationen ökonomischer als Konfidenzintervalle wären, wird von manchen Autoren als Vorteil von Signifikanztests gewertet (siehe dazu Schmidt & Hunter, 1997). Oft wäre die genaue Information einer Punktschätzung samt Konfidenzintervall gar nicht notwendig und es genü-

ge – gleichsam für den ersten Überblick – ein rasches Abtasten signifikanter Ergebnisse: Wer kennt nicht das rasche Abtasten signifikanter Werte, um sich schnell und bequem in einer großen Menge von Ergebnissen zurechtzufinden?

Schmidt und Hunter (1997) widersprechen dieser Strategie und plädieren statt dessen, Effektstärken (*eta*, Cohen's *d*, *r*) anstatt von *p*-Werten zur schnellen Orientierung zu verwenden. Zugegeben, bei gleichen Stichprobenumfängen besteht eine perfekte Beziehung zwischen *p*-Werten und Effektstärken. Meistens jedoch sind Stichprobenumfänge innerhalb einer Studie und erst recht zwischen Studien verschieden. In diesen Fällen liefern *p*-Werte trivialere Informationen als Effektstärken und häufige Fehlinterpretationen sind die Folge: Etwa, daß ein signifikanter Korrelationskoeffizient $r = ,1$ ($N = 1.000$) in einer anderen Studie $r = ,4$ ($N = 30$) nicht signifikant „bestätigt“ werden konnte, womit widersprüchliche Ergebnisse vorlägen.

Die Konzentration auf Effektstärken hingegen vermeidet Trugschlüsse dieser Art: Beide Studien erbrachten einen positiven Zusammenhang und sprechen somit zu(un)gunsten der Hypothese X. Die zweite Studie ($r = ,4$) ist keine „Widerlegung“, sondern eine Bestätigung der ersten. Für einen ersten Überblick reichen diese Informationen aus, und Konfidenzintervalle geben dann zusätzliche Aufschlüsse, welche auch die einzelnen Stichprobenumfänge berücksichtigen. Zusammengefaßt, auch für den Zweck der schnellen Orientierung sind Effektstärken besser als *p*-Werte geeignet (Schmidt & Hunter, 1997).

Betrachtet man die vielen Vorteile von Konfidenzintervallen gegenüber Signifikanztests stellt sich die Frage, warum Konfidenzintervalle noch immer eine eher untergeordnete Rolle in der Darstellung statistischer Ergebnisse spielen. Der nächste Abschnitt versucht Antworten darauf zu geben.

Widerstände gegen Konfidenzintervalle

Zur Erklärung für die immer noch relativ geringe Verbreitung von Konfidenzintervallen lassen sich mehrere Vermutungen aufstellen (siehe dazu Reichardt & Gollob, 1997; Steiger & Fouladi, 1997): Mangelnde Verfügbarkeit von Konfidenzintervallen in Software-Paketen; die Notwendigkeit Signifikanztests durchzuführen, um in wissenschaftlichen Journalen publizieren zu können; die oftmals geringen Effektstärken, welche durch die Betonung eines „signifikanten Ergebnisses“ verschleiert werden; der Umstand, daß Konfidenzintervalle oft sehr breit sind und dadurch nur ungenaue Schätzungen zulassen – um einige der wichtigsten Gründe zu nennen.

Als weitere Ursache kommt die Heuristik des „Sozialen Beweises“ in Frage, wonach sich viele Personen nicht irren können; deswegen müsse „schon etwas Wahres daran sein“. Heuristiken, wie die des Sozialen Beweises, zeichnen sich aber gerade dadurch aus, daß sie oft gute, schnelle Schätzungen abgeben, manchmal aber in die Irre führen können (Tversky & Kahneman, 1974). Es bleibt zu hoffen, daß ein bewußtes Erkennen dieser Widerstände zu deren Überwindung führt.

Sollen Konfidenzintervalle Signifikanztests ersetzen?

Nach anfänglichen Meinungsverschiedenheiten bezüglich der Sinnhaftigkeit und Interpretierbarkeit von Konfidenzintervallen (siehe oben) zeichnet sich mittlerweile ein Konsens ab. So befürworten in dem kürzlich erschienenen Sammelband „What if there were no significance tests?“ (Harlow et al., 1997) alle elf Autoren – einschließlich der Befürworter von Signifikanztests – die Verwendung von Konfidenzintervallen (Harlow, 1997). Die zentrale Frage die sich somit stellt betrifft nicht mehr die Sinnhaftigkeit der Berichterstattung von Konfidenzintervallen, sondern ob Konfidenzintervalle und Signifikanztests nebeneinander Platz haben, oder ob Konfidenzintervalle Signifikanztests generell ersetzen sollten. Letzteres ist der extremere Standpunkt, da er eine Abschaffung von Signifikanztests impliziert, während der erste Standpunkt nach den Bedingungen fragt, unter welchen Konfidenzintervalle beziehungsweise Signifikanztests angemessener sind. Als Befürworter der Abschaffung von Signifikanztests und somit des zweiten Standpunkts fragen Hunter und Schmidt (1997): „Can you [defenders of significance tests] articulate even one legitimate contribution that significance testing has made ... to the development of cumulative scientific knowledge?“ I believe you will not be able to do so“ (S. 116). Prompt antwortet Abelson (1997) und nennt zwei Anwendungen: Erstens Entscheidungsexperimente zwischen zwei rivalisierenden Theorien, und zweitens das Testen der Kongruenz eines Modells mit empirischen Daten („Goodness-of-Fit“ Testen). Beide Anwendungen können jedoch – wie ich im folgenden argumentieren werde – ohne Signifikanztests besser bewerkstelligt werden.

Wenden wir uns zuerst der Anwendung von Signifikanztests als Entscheidungshilfe zwischen zwei rivalisierenden Theorien zu. Zum Beispiel soll Theorie A eine positive, Theorie B eine negative Korrelation für ein Experiment vorhersagen. In diesem Fall genüge lediglich eine statistisch signifikante Korrelation egal welcher Richtung, welche dann zugunsten einer der beiden Theorien sprechen würde: Eine statistisch signifikante positive Korrelation würde für Theorie A, eine

statistisch signifikante negative Korrelation für Theorie B sprechen. Die Größe des Effekts sei weitgehend uninteressant, da lediglich eine Entscheidung zwischen den beiden Theorien gesucht werde.

Nach wie vor treffen jedoch die eingangs erwähnten Kritikpunkte, wie die Unmöglichkeit, etwas über die Gültigkeit einer Hypothese zu erfahren, oder die Trivialität der Nullhypothese auf Signifikanztests zu. Das Konfidenzintervall aber ermöglicht ohne diese Schwächen eine Entscheidung zwischen beiden rivalisierenden Theorien: Eine Entscheidung zugunsten der Theorie A ($r > 0$) oder der Theorie B ($r < 0$) ist möglich, solange das Konfidenzintervall den Wert null ausschließt. Somit gibt es keinen Grund, Entscheidungen zwischen zwei Theorien durch Signifikanztests, und nicht durch Konfidenzintervalle zu lösen. Die Richtung der entsprechenden Effektstärke (r , d , etc.) samt dessen Konfidenzintervall entscheidet zugunsten einer Theorie.

Die zweite Anwendung bezieht sich auf die Übereinstimmung eines theoretisch postulierten Modells mit empirischen Daten. Ein bekanntes Beispiel dafür wäre die Prüfung der Modellkonformität von Strukturgleichungsmodellen (siehe Jöreskog & Sörbom; 1989). Man spezifiziert ein Modell im vorhinein und überprüft, ob die Daten signifikant von dem theoretisch postulierten Modell abweichen. Damit können auch verschiedene Modelle am selben Datensatz (bei gleichem Stichprobenumfang) verglichen werden. Das Modell mit dem höchsten p -Wert wird dann üblicherweise gewählt. Die Forscherin ist somit nicht an der Alternativ-, sondern an der Nullhypothese interessiert. Im Grunde basieren derartige Modelltestungen auf der gleichen Logik wie einfache t -Tests. Lautet im Falle des t -Test die Nullhypothese, daß zwei Mittelwerte der selben Population entstammen, entspricht nun die Nullhypothese einem Strukturgleichungsmodell. Somit sind auch die Probleme die gleichen. Große Stichprobenumfänge führen zu maximalen Teststärken und zeigen dadurch auch zufällige Abweichungen vom theoretisch postulierten Modell an; andererseits führen geringe Stichprobenumfänge zu geringen Teststärken und können dadurch nicht zwischen verschiedenen Modellen differenzieren.

Steiger und Fouladi (1997) bieten einen ausgezeichneten Überblick über alternative Kriterien der Anpassungsgüte von Daten an ein theoretisch postuliertes Modell, welche nicht auf Signifikanztests beruhen und somit die mit Signifikanztests verbundenen Probleme vermeiden (z.B. Goodness-of-Fit-Index (GFI) oder Root-Mean-Square-Error-of-Aproximation (RMSEA) für Strukturgleichungsmodelle). Somit sind alternative, besser geeignete Indizes, welche nicht auf der Be-

rechnung von p -Werten beruhen, in der Lage, Signifikanztests zum Zwecke der Überprüfung der Modellkonformität zu ersetzen. Zusammengefaßt, beide von Abelson (1997) angeführten Anwendungen von Signifikanztests können durch bessere Methoden ersetzt werden.

Gefahr möglicher Fehlinterpretationen von Effektstärken

In den bisherigen Ausführungen habe ich auf die Vorteile von Effektstärken und Konfidenzintervallen gegenüber Signifikanztests hingewiesen. Trotz dieser Vorteile sind auch Effektstärken mit Vorsicht zu interpretieren – eine Vorsicht die vor allem für experimentellen Designs angebracht erscheint: Effektstärken hängen deutlich von der Manipulation der unabhängigen Variablen ab. Für die Stärke dieser experimentellen Variation gibt es jedoch kein Maß (Ronis, 1981). Zum Beispiel legt ein Versuchsleiter für sein Experiment den Versuchspersonen hypothetische Szenarien (Vignetten) vor, auf welche die Versuchspersonen antworten sollen. Eingebettet in ein varianzanalytischen 2×2 Design manipuliert er den ersten Faktor „emotionale Beziehung zwischen zwei Personen“, welche in den Vignetten interagieren (positive versus negative Beziehung). Es würde nun einen bedeutenden Unterschied machen, ob er die Faktorstufe „positive emotionale Beziehung“ als „allerbeste Freunde“ oder als „gute Bekannte“ operationalisiert. Je nach Stärke der Manipulation fallen die Effektstärken für den ersten Faktor „emotionale Beziehung“ unterschiedlich aus. Bei einer starken Manipulation des Beziehungsfaktors wird wahrscheinlich der Haupteffekt für den zweiten Faktor, als auch der Interaktionseffekt, gering ausfallen, da vermutlich der Beziehungsfaktor viel Varianz der abhängigen Variable erklären wird. Umgekehrt, eine schwache Manipulation des Beziehungsfaktors („gute Bekannte“) könnte den zweiten Haupt-, als auch den Interaktionseffekt vergrößern (analoges gilt für die Manipulation der Faktorstufe „negative emotionale Beziehung“). Somit sind Effektstärken immer nur relativ, zur Stärke der Variation der unabhängigen Variablen zu interpretieren. Dieses Problem ist geringer bei unabhängigen Variablen wie Geschlecht, Alter oder Persönlichkeitsmerkmalen, welche in natürlicher Variation vorliegen. Zusammengefaßt, die Darstellung von Effektstärken samt Konfidenzintervallen stellt einen deutlichen Fortschritt gegenüber Signifikanztests dar, deren Interpretation jedoch der Vorsicht bedarf, da Effektstärken vom Ausmaß der experimentellen Variation abhängen. Daraus folgt, daß experimentelle Designs

häufig nur psychologische *Prinzipien* oder *Mechanismen* aufzeigen können, und einer Aussage wie „Effekt X ist zwar vorhanden – aber gering“ wenig Bedeutung zukommen kann. Diese Probleme treffen selbstverständlich auch auf Metaanalysen zu.

Schlußfolgerung

Konfidenzintervalle vermeiden die Probleme klassischer Signifikanztests. Sie benötigen weder a-priori Hypothesen und prüfen keine trivialen Hypothesen. Konfidenzintervalle beinhalten die Informationen eines Signifikanztests und sind wesentlich leichter zu verstehen als diese, woraus sich deren didaktische Überlegenheit herleitet.

Im Sinne der Subjektiven Wahrscheinlichkeitstheorie interpretiert, gewährleisten Konfidenzintervalle ein Wahrscheinlichkeitsurteil über das Vorzeichen eines Effekts. Liegt der Wert null außerhalb des 95% Konfidenzintervalls, kennt man das Vorzeichen der wahrscheinlichsten (95% Vertrauen) Populationsparameter. Befindet sich der Wert null innerhalb des 95% Konfidenzintervalls läßt sich nichts mit großer Sicherheit über das Vorzeichen eines Effekts aussagen. Der Effekt kann positiv, negativ, und, zumindest theoretisch, auch null sein – wenn auch letzteres extrem unwahrscheinlich ist. Weiters läßt sich annehmen, daß der Effekt auch sehr klein (nahe null) sein könnte.

Die Frage, ob Signifikanztests durch Konfidenzintervalle zu ersetzen sind oder nicht, kann mit einem vorsichtigen „Ja“ beantwortet werden. Da Konfidenzintervalle die Informationen eines Signifikanztests enthalten, entsteht durch den Ersatz von Signifikanztests durch Konfidenzintervalle kein Informationsverlust und somit kein Risiko.

Insgesamt stellen Konfidenzintervalle neben Replikationen, graphischen Darstellungen und Metaanalysen einen methodisch überlegenen Ersatz für Signifikanztests dar, womit Konfidenzintervalle langfristig eine aussichtsreichere Forschung als bisher garantieren dürften.

Literatur

- [1] Abelson, R. P. (1997). A retrospective on the significance test ban of 1999 (If there were no significance tests, they would be invented). In L. L. Harlow, S. A. Mulaik & J. H. Steiger (Eds.), *What if there were no significance tests?* (S. 117-141). Hillsdale: Lawrence Erlbaum.
- [2] Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin*, *66*, 423-437.
- [3] Baril, G. L. & Cannon, J. T. (1995). What is the probability that null hypothesis testing is meaningless? *American Psychologist*, *50*, 1098-1099.
- [4] Bley Müller, J., Gehlert, G. & Gülicher, H. (1988). *Statistik für Wirtschaftswissenschaften* (5. Aufl.). München: Vahlen.
- [5] Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*, 997-1003).
- [6] Cohen, J. (1995). The earth is round ($p < .05$): Rejoinder. *American Psychologist*, *50*, 1103.
- [7] Cortina, J. M. & Dunlap, W. P. (1997). On the logic and the purpose of significance testing. *Psychological Methods*, *2*, 161-172.
- [8] DeFinetti, B. (1971). *Theory of probability: A critical introductory treatment* (Vol. 1). New York: Wiley.
- [9] Edwards, W., Lindman, H. & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, *70*, 193-242.
- [10] Frick, R. W. (1995). A problem with confidence intervals. *American Psychologist*, *50*, 1102-1103.
- [11] Frick, R. W. (1996). The appropriate use of null hypothesis testing. *Psychological Methods*, *1* 379-390).
- [12] Gigerenzer, G. (1993). The superego, the ego, and the id in statistical reasoning. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences. Methodological issues* (pp. 311-339). Hillsdale: Lawrence Erlbaum.
- [13] Gigerenzer, G.; Swijtink, Z.; Porter, T.; Daston, L.; Beatty, J. & Krüger, L. (1989). *The empire of chance*. Cambridge: Cambridge University Press.

- [14] Hagen, R. L. (1997). In praise of the null hypothesis statistical test. *American Psychologist*, 52, 15-24.
- [15] Harlow, L. L. (1997). Significance Testing Introduction and Overview. In L. L. Harlow, S. A. Mulaik & J. H. Steiger (Eds.), *What if there were no significance tests?* (S. 1-17). Hillsdale: Lawrence Erlbaum.
- [16] Harlow, L. L.; Mulaik, S. A. & Steiger, J. H. (Eds.), *What if there were no significance tests?* Hillsdale: Lawrence Erlbaum.
- [17] Hubbard, M. (1995). The earth is highly significantly round ($p < .0001$). *American Psychologist*, 50, 1098.
- [18] Jöreskog, K. G. & Sörbom, D. (1989). LISREL 7. *A guide to the program and applications* (2nd ed.). Chicago, IL: SPSS.
- [19] Kendall, M. & Stewart, A. (1979). *The advanced theory of statistics. Vol. 2. Inference and relationship*. London: Charles Griffin & Co.
- [20] Kleiter, G. D. (1969). Krise der Signifikanztests in der Psychologie. *Jahrbuch für Psychologie, Psychotherapie und medizinische Anthropologie*, 17, 144-163.
- [21] Kleiter, G. D. (1980). *Bayes Statistik: Grundlagen und Anwendungen*. Berlin: DeGruyter.
- [22] McCraw, K. O. (1995). Determining false alarm rates in null hypothesis testing research. *American Psychologist*, 50, 1099-1100.
- [23] Menges, G. (1968). *Grundriß der Statistik. Teil 1: Theorie*. Köln: Westdeutscher Verlag.
- [24] Mulaik, S. A., Raju, N. S. & Harshman, R. A. (1997). There is a time and place for significance testing. In L. L. Harlow, S. A. Mulaik & J. H. Steiger (Eds.), *What if there were no significance tests?* (S. 65-115). Hillsdale: Lawrence Erlbaum.
- [25] Parker, S. (1995). The "difference of means" may not be the "effect size". *American Psychologist*, 50, 1101-1102.
- [26] Reichardt, C. S. & Gollob, H. F. (1997). When confidence intervals should be used instead of statistical significance tests, and vice versa. In L. L. Harlow, S. A. Mulaik & J. H. Steiger (Eds.), *What if there were no significance tests?* (S. 259-284). Hillsdale: Lawrence Erlbaum.
- [27] Ronis, D. L. (1981). Comparing the magnitude of effects in ANOVA designs. *Educational and Psychological Measurement*, 41, 993-1000.

- [28] Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology. Implications for training of researchers. *Psychological Methods, 1*, 115-129.
- [29] Schmidt, F. L. & Hunter, J. E. (1997). Eight common but false objections to the discontinuation of significance testing in the analysis of research data. In L. L. Harlow, S. A. Mulaik & J. H. Steiger (Eds.), *What if there were no significance tests?* (S. 37-64). Hillsdale: Lawrence Erlbaum.
- [30] Sedlmeier, P. (1996). Jenseits des Signifikanztest-Rituals: Ergänzungen und Alternativen. *Methods of Psychological Research Online, 1*, 41-63.
- [31] Steiger, J. H. & Fouladi, R. T. (1997). Noncentrality interval estimation and the evaluation of statistical models. In L. L. Harlow, S. A. Mulaik & J. H. Steiger (Eds.), *What if there were no significance tests?* (S. 221-257). Hillsdale: Lawrence Erlbaum.
- [32] Stegmüller, W. (1973). *Probleme und Resultate der Wissenschaftstheorie und Analytischen Philosophie IV: Personelle und statistische Wahrscheinlichkeit*. Berlin: Springer.
- [33] Svyantek, D. J. & Ekeberg, S. E. (1995). The earth is round (So we can probably get there from here). *American Psychologist, 50*, 1101.
- [34] Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.
- [35] Tukey, J. W. (1991). The philosophy of multiple comparisons. *Statistical Science, 6*, 100-116.
- [36] Tversky, A & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*, 1124-1131.
- [37] Winkler, R. L. (1972). *Introduction to Bayesian Inference and Decision*. New York: Holt, Rinehart & Winston.
- [38] Witte, E. H. (1980). *Signifikanztest und statistische Inferenz*. Stuttgart: Enke.
- [39] Wright, A. & Ayton, P. (1994). *Subjective probability*. Chichester: Wiley.