



Signifikanz

Ingrid Scharlau, Tobias Jenert

Der Signifikanztest hat eine ausgesprochene Erfolgsgeschichte vorzuweisen, obwohl er immer wieder als problematische Praktik ausgewiesen wird. Vor diesem Hintergrund stellt sich die Frage, ob der Signifikanztest weitergehende Funktionen jenseits der begrenzten statistischen Aussagekraft erfüllt. Der Signifikanzstern () entpuppt sich dabei als symbolischer Repräsentant einer Praktik, die hilft, Zugang und Zugehörigkeit zur disziplinären Gemeinschaft zu regeln. Er verweist auf Praktiken des Forschens und Publizierens, er schafft Ordnung, reduziert Komplexität und ermöglicht Orientierung.*

In vielen empirischen Wissenschaften steht der Stern (*) für das signifikante Ergebnis eines statistischen Tests. Er kürzt also die Formulierung *statistisch signifikant* ab. Diese ist selbst wieder eine Verkürzung der genauen Testergebnisse auf eine Standardformulierung. Dabei werden auch die Voraussetzungen des Tests und der methodologischen Erwägungen, die ihm vorausgehen, aus der Darstellung gewissermaßen herausgekürzt. Zentral für die folgende Argumentation ist, dass damit eine Abkürzung im Denken genommen wird und wissenschaftliches Urteilen durch eine quasi-automatische und vermeintlich objektive Schlussform ersetzt wird. Unser Beitrag wird von der Vermutung geleitet, dass die Karriere des Signifikanzsterns nicht allein und vielleicht nicht einmal vorwiegend auf der Leistungsfähigkeit des zugrundeliegenden Testverfahrens beruht. Wir fragen deswegen nach der weiterführenden Symbolkraft des Sterns im Kontext wissenschaftlichen Handelns und akademischer Kulturen, mit besonderer Aufmerksamkeit für eine wissenssoziologische und wissenschaftskritische Perspektive. In einem ersten Abschnitt schildern und kritisieren wir knapp das Verfahren des Signifikanztests, seine Geschichte und seine Verwendung im Schließen aus Daten, die – darauf liegt unser Fokus – durch seine Aussagekraft nicht gedeckt ist. Im zweiten Abschnitt widmen wir uns der weitergehenden Frage nach der Semantik des Begriffs *signifikant* und seiner Bedeutung für die Etablierung wissenschaftlicher Gemeinschaften.

Das statistische Verfahren, seine Praxis und seine Kritik

Signifikanztests haben sich in vielen Wissenschaften überraschend schnell durchgesetzt und sind auch heute noch weit verbreitet, obwohl es viele Einwände gegen sie gibt und alternative statistische Verfahren, etwa die Bayes-Inferenz, existieren. Wir diskutieren den Signifikanzstern exemplarisch an einem besonders charakteristischen Mikrokosmos, der

Psychologie. Viele der Beobachtungen dürften aber ähnlich für andere Wissenschaften gelten. Als charakteristisch kann die Psychologie hier gelten, weil sie gleichzeitig mit der und durch die Verbreitung von Signifikanztests eine schnelle sowie wirksame Stärkung ihrer angenommenen Wissenschaftlichkeit und gesellschaftlichen Rolle erfuhr. Sie hat Signifikanztests so stark adaptiert, dass andere Verfahren des Schließens aus Daten praktisch verdrängt wurden. Ein geeigneter Fall ist die Psychologie auch deshalb, weil sie in den vergangenen Jahrzehnten gerade in Bezug auf ihre quantitativen Methoden für andere Wissenschaften zum Vorbild geworden ist – und auch, weil die Folgen dieser Praxis in ihr inzwischen besonders deutlich geworden sind.

Signifikanztests wurden in der ersten Hälfte des 20. Jahrhunderts zunächst für die Agrarwissenschaft entwickelt und in der Psychologie ab Mitte der 1930er Jahre erstmals verwendet. Über die amerikanische Graduiertenausbildung wurden sie rasch zum Standard und ersetzen dabei andere statistische Verfahren („inference revolution“, Gigerenzer/Swijtink/Porter/Daston/Beatty/Krüger 1989, Kapitel 6; Rucci/Tweney 1980). Seit 1952 werden sie im *Publication Manual der American Psychological Association* (APA) empfohlen:

„Convention permits the single asterisk (*) to indicate statistics which are ‘significant at the 5% level,’ and the double asterisk (**) to mean ‘significant at the 1% level’.“ (APA Council of Editors 1952: 33)¹

Bereits Mitte der 1950er Jahre enthielt die überwiegende Mehrheit der Artikel einflussreicher Zeitschriften in der Psychologie Signifikanztests (und beinahe ausschließlich positive Ergebnisse; Sterling 1959); heute liegt der Anteil nahe bei 100%. Im Jahr 1962 legte der Herausgeber für das sehr renommierte *Journal of Experimental Psychology* sogar fest, dass Artikel ohne signifikantes Ergebnis nicht und solche, die nur ein Signifikanzniveau von .05 erreichen, nur ausnahmsweise publiziert werden können (Melton 1962). Eine ähnliche Entwicklung lässt sich in der Erziehungswissenschaft beobachten, die Signifikanztests ebenfalls früh einsetzte, da sie auf diese Weise in die Praxis, hier vor allem die Bildungspolitik, wirken konnte (Danziger 1987). Auch weitere Fächer wie Soziologie, Archäologie, Politikwissenschaft und Ökonomie sowie später noch die Medizin und Neurowissenschaften übernahmen die Verfahren mit Nachdruck.

Wir fahren mit einer sehr kurzen Skizze der Verfahren fort (weiterführend und sehr lesenswert etwa Gigerenzer 1987, 2004; Gigerenzer/Swijtink/Porter/Daston/

Beatty/Krüger 1989, Kapitel 3 und 6; Meehl 1978). Typischerweise in Tabellen oder Grafiken gebraucht, während in Texten die Darstellung $p < .05$ üblicher (und von der APA vorgeschrieben) ist, steht der Stern in der Regel als Abkürzung für einen signifikanten statistischen Nullhypothesentest. Dieser Test² prüft die Geltung einer Nullhypothese, beispielsweise *Die Lernstrategie X hat keinen Effekt auf Kompetenzen*. Die Nullhypothese ist in der Regel für die Forschung wenig interessant (man will ja schließlich wissen, welche Lernstrategien einen Effekt haben) und sehr spezifisch (*kein Effekt*, also exakt 0). Wenn unter Geltung dieser Hypothese ein vorher festgelegtes Signifikanzniveau unterschritten wird, wird das Testergebnis als *signifikant*, genauer als *statistisch signifikant* bezeichnet. Dies bedeutet, dass die erhobenen (oder extremere) Daten unter Annahme der Nullhypothese (und anderer Annahmen über die Verteilungsfunktion der Daten) eine Wahrscheinlichkeit unterhalb des Signifikanzniveaus haben. Ist dies der Fall, wird die Nullhypothese abgelehnt. Ein statistisch gerechtfertigter Schluss über alternative Hypothesen oder Erklärungen, etwa, dass die Lernstrategie wirksam ist, kann aus der Ablehnung nicht gezogen werden – auch wenn diese Praxis weit verbreitet ist.

Das Signifikanzniveau beziehungsweise der akzeptierte Alpha-Fehler steht dafür, dass die Nullhypothese tatsächlich³ zutrifft, aber aufgrund der Ergebnisse des statistischen Tests abgelehnt wird. Das Fehlerniveau kann gewählt werden und sollte dem zu tolerierenden Schlussfehler angepasst werden; praktisch haben sich fünf Prozent durchgesetzt, ohne dass dafür spezielle Gründe ersichtlich wären. Bemerkenswert ist, dass die Bedeutung eines signifikanten Ergebnisses selbst von Expert_innen oft falsch verstanden wird (Haller/Krauss 2002; für Mutige zur Selbstüberprüfung Dienes 2008: 212), etwa dahingehend, dass die Wahrscheinlichkeit, in einer zukünftigen Untersuchung dieselben Ergebnisse zu erhalten, 95 % beträgt, oder dass die Wahrscheinlichkeit, dass die Lernstrategie nicht wirkt, sehr gering ist. Das zentrale Missverständnis ist hier die Annahme, dass Signifikanz etwas über die Wahrscheinlichkeit von Hypothesen aussagt. Tatsächlich beschreibt sie jedoch nur, wie erwähnt, die Wahrscheinlichkeit von Daten unter der Nullhypothese. Carver (1978: 383) nennt diese sehr verbreiteten Missverständnisse „fanciful inferences“, um zu betonen, dass der Signifikanztest ohne solche Unterstellungen kaum seine dominante Stellung hätte erhalten können.

Fisher, einer der Erfinder von Signifikanztests, war überzeugt, sie seien eine vollkommen objektive Form des Schließens aus Daten. Auch wenn dies sicherlich nicht von allen geteilt wurde, sickerte es doch in die Darstellung von Tests ein. Wie Gigerenzer, Swijtink,

Porter, Daston, Beatty und Krüger (1989) berichten, wurde der Signifikanztest von Beginn an als kanonisch dargestellt, als feststehende Tatsache und objektives, neutrales Werkzeug, das keiner Semantik zu bedürfen scheint. Auch in Zeitschriftenpublikationen werden Signifikanztests auf die erwähnten Formeln und den Stern reduziert. Vor diesem Hintergrund entwickelt sich, was Gigerenzer und Kollegen das „null ritual“ nennen und was im Signifikanzstern symbolisiert ist: die immer gleiche mechanische Anwendung desselben Verfahrens auf beliebige Daten:

„(1) Set up a statistical null hypothesis of ‚no mean difference‘ or ‚zero correlation.‘ Don’t specify the predictions of your research hypothesis or of any alternative substantive hypotheses. (2) Use 5% as a convention for rejecting the null. If significant, accept your research hypothesis. (3) Always perform this procedure.“ (Gigerenzer/Krauss/Vitouch 2004: 392)

In der Folge dieser Entwicklung lassen sich einige Probleme beobachten, die in den vergangenen Jahren zu einer stürmischen Methodendiskussion in verschiedenen Wissenschaften geführt haben. Eines davon ist der hohe Anteil statistisch signifikanter Befunde; nicht signifikante Befunde wurden, ganz im Sinne Meltons, jahrzehntelang kaum publiziert (Farelli 2011), was vermutlich zu einer erheblichen Verzerrung des Gesamtbildes geführt hat (Ioannidis 2005). Infolge des durch Zeitschriften ausgelösten Drucks, signifikante Ergebnisse zu erzielen, entwickelten und verbreiteten sich fragwürdige Forschungspraktiken, etwa Hypothesen aus einem Datensatz zu entwickeln und danach an diesem statistisch zu prüfen (Kerr 1998), Daten durch verschiedene Tests zu durchforsten, von denen nur die signifikanten berichtet werden, (John/Loewenstein/Prelec 2012), die Erhebung abubrechen, nachdem Signifikanz erreicht wurde (Simmons/Nelson/Simonsohn 2011), oder wunderliche Datensätze aus der Analyse auszuschließen. All diese weit verbreiteten Praktiken verletzen Grundvoraussetzungen des Signifikanztests.

Selbst wenn sie korrekt angewendet würden, sind Signifikanztests nicht einfache Werkzeuge, sondern implizieren spezifische wissenschaftstheoretische und methodologische Vorstellungen. Dies ist im Falle von empirischen Wissenschaften deswegen besonders bedeutsam, weil viele von ihnen wissenschaftstheoretische und -methodologische Überlegungen aus der Forschung selbst auslagern, sodass diese als Gegenstand der Reflexion unsichtbar werden. Fishers Vorstellung eines quasi-mechanischen Schließens haben wir bereits erwähnt. Auch wenn viele heuti-

ge Wissenschaftler_innen diese Vorstellung explizit vermutlich ablehnen würden, schimmert sie im konkreten Einsatz von Signifikanztests zumindest durch. So spielen die Daten selbst, ihre Sichtung und Beschreibung in der konkreten Forschungspraxis eine marginale Rolle. Mit der breiten Verwendung der Varianzanalyse, deren Fokus auf den Effekten einzelner Bedingungen liegt, setzen sich ein einzelner Typus des Experiments (Danziger 1985, 1987) und eine bestimmte Vorstellung von Kausalität durch, die auf den Prinzipien der Isolation von Faktoren und der Kontrolle beruht. Theoriegeleitete wird zunehmend durch induktive Forschung ersetzt. Insgesamt verschreiben sich Wissenschaften, in denen Signifikanztests zum Standard werden, einer Art „incremental encyclopedism“ (Bazerman 1987: 273), in dem einzelne Forscher_innen innerhalb eines inhaltlich und methodisch weitgehend standardisierten Rahmens kleine Teilstücke zum Wissen beitragen.

Diese Entwicklung lässt sich auch im Zeitschriftenartikel nachzeichnen, der zeitgleich mit der Verbreitung von Signifikanztests erhebliche Änderungen durchmacht, die sich bis heute halten und in den vergangenen Jahrzehnten vielleicht sogar noch einmal verstärkt haben. Belege hierfür sind eine zunehmende Uniformität der Methoden, flachere Ableitungen von Forschungsproblemen aus der Literatur wie auch Schlussfolgerungen, die sich nicht mehr weit von den Daten entfernen, und die Verschiebung des rhetorischen Fokus von einer überzeugenden Argumentation auf den Nachweis, dass man keine Fehler gemacht hat und den Standards Genüge tut (Danziger/Dzinas 1997). Ein Teil dieser Änderungen ist im *Publication Manual* der APA kodifiziert. Dieses Manual ist mehr als eine Sammlung stilistischer Konventionen, nämlich Ausdruck wie Ursache der disziplinspezifischen Epistemologie und Werte und eine wichtige Sozialisationserfahrung von Wissenschaftler_innen (Madigan/Johnson/Linton 1995: 428). Seit einigen Jahrzehnten werden die APA-Konventionen auch in anderen Wissenschaften zunehmend zum Standard, etwa in Teilen der Pädagogik/Erziehungswissenschaft, der Soziologie, der Psycholinguistik und didaktischer Diskurse. Wir behaupten nicht, dass die Einführung des Signifikanztests all diese Änderungen verursacht hat. Auch wenn aufgrund der Chronologie Einflüsse naheliegen, handelt es sich hier eher um ein dichtes Gewebe miteinander verflochtener Prozesse, in denen sich wissenschaftliche Praktiken eher langsam verschieben als plötzlich ändern. Der Signifikanzstern wirkt in diesem Prozess wie ein stabilisierendes Moment.

Interessanterweise wurde diese verkürzte Praxis seit Beginn von scharfer Kritik begleitet, sodass die Schwächen des Signifikanztests seit seinem Entste-

hen bekannt waren. Die Kritik konnte sich jedoch nur wenig Gehör verschaffen und insbesondere dem praktischen Erfolg wenig entgegensetzen (insbesondere, aber nicht nur Cohen 1994; Gigerenzer 2004; Meehl 1978) und so haben sich weder bescheidene Forderungen nach zusätzlicher Information wie Konfidenzintervallen (Loftus 1993) noch die Forderung nach substantiellen quantitativen Modellen (Meehl 2006) noch der radikale Vorschlag einiger Mitglieder der *APA Task Force on Statistical Inference*, auf den Signifikanztest zu verzichten (Fidler 2002), durchgesetzt.

Schon den frühen Kritiken merkt man einen gereizten Ton an, der sich in den Diskussionen bis heute hält. Zur Illustration Loftus (1991: 103):

„Despite the stranglehold that hypothesis testing has on experimental psychology, I find it difficult to imagine a less insightful means of transiting from data to conclusions.“

Insgesamt ergibt sich der Eindruck starrer Praxis auf der einen und starrer Gegenwehr auf der anderen Seite.⁴ Hartnäckig und erstarrt scheint diese Debatte aber auch deswegen zu sein, weil sich in ihr zwei Weisen ausdrücken, Wissenschaft zu tun. Die eine Weise charakterisiert die Mitwirkung an der Produktion von Erkenntnis für praktische und pragmatische Zwecke einschließlich der Sicherung der gesellschaftlichen Relevanz der eigenen Wissenschaft und der Sicherung der eigenen Karriere. Die andere orientiert sich stärker an der Suche nach Wahrheit oder zumindest reinen Formen der Erkenntnisgenerierung. In ihr scheint noch der Habitus weltabgewandter Gelehrter durch, die übersehen, wie eng die Verwendung von Signifikanztests mit vielen anderen Aspekten von Wissenschaft verflochten ist. Während die einen weitermachen wollen wie bisher (und man ihnen übrigens zugestehen muss, dass die Psychologie genau in der Zeit, in der sich die fragwürdige Statistik durchsetzte, zu großer gesellschaftlicher Relevanz fand), möchten sich die anderen am liebsten neben die Praxis stellen und bei Null anfangen. Es stecken aber alle im gleichen Geflecht der historischen Entwicklung, und der weitere Weg wird wohl ein komplexes und erst im Rückblick durchschaubares Bewegungsmuster zwischen *weiter so* und *ganz anders* sein.

Sinnstiftung in der Wissenschaft

Bis hierhin haben wir gezeigt, dass der Signifikanztest eine ausgesprochene Erfolgsgeschichte vorzuweisen hat, obwohl er offenkundig und wiederkehrend als problematische wissenschaftliche Praktik

ausgewiesen wird. Vor diesem Hintergrund stellt sich uns die Frage, ob der Signifikanztest weitergehende Funktionen jenseits der begrenzten statistischen Aussagekraft erfüllt. Wir möchten uns dieser Frage nähern, indem wir uns die Struktur der Wissenschaft als System kultureller Gemeinschaften vergegenwärtigen und fragen, was der Signifikanzstern als Symbol für diese Gemeinschaften bedeutet.

Das Wissenschaftssystem ist in disziplinäre Gemeinschaften gegliedert, die sich als Kulturen mit jeweils eigenen und teils sehr distinkten Merkmalen beschreiben lassen (Becher/Trowler 2001). Kulturen verfügen über jeweils eigene Bedeutungssysteme: Bestimmte Symbole sind Ausdruck tieferliegender kultureller Sinnstrukturen, die sich Außenstehenden nicht ohne Weiteres erschließen. Der Signifikanzstern kann so als ein Symbol verstanden werden, das auf weiterführende Sinnzusammenhänge in der Wissenschaft verweist. Anders formuliert: Welche Botschaft sendet die einzelne Person mit dem Gebrauch des Signifikanzsterns? Inwiefern signalisiert und legitimiert das Erzielen signifikanter Ergebnisse die Zugehörigkeit zu ihrer Scientific Community?

Der Duden (o.D.) unterscheidet neben der statistischen zwei weitere Bedeutungen des Adjektivs *signifikant*. Alltags- oder bildungssprachlich bedeutet es „in deutlicher Weise als wesentlich, wichtig, erheblich erkennbar“ und „in deutlicher Weise als kennzeichnend, bezeichnend, charakteristisch, typisch erkennbar“. Speziell für den wissenschaftlichen Kontext sprechen Onwuegbuzie und Leech (2004) von „multiple significances“ und weisen neben der statistischen eine praktische, klinische und ökonomische Signifikanz aus. Praktische Signifikanz bezieht sich auf die Stärke eines Zusammenhangs. Zusammenhänge könnten zwar statistische Signifikanz erreichen, in ihrer Wirkung aber zu klein sein, um für die Beantwortung der Fragestellung oder das Treffen einer konkreten Entscheidung bedeutsam zu sein. Klinische Signifikanz ermisst, inwiefern eine Intervention für die Betroffenen eine tatsächlich wahrnehmbare Verbesserung ihrer Lebensumstände bewirkt. Ökonomische Signifikanz bezeichnet die Ressourceneffizienz einer Intervention und fragt danach, ob sich mit einer Intervention unter realisiertem Ressourceneinsatz ein relevantes Ergebnis (im Sinne der klinischen Signifikanz) erzielen lässt. Während der Stern die statistische Signifikanz repräsentiert, umschreiben die anderen Signifikanzen das, was alltagssprachlich unter *signifikant* verstanden wird, nämlich verschiedene Ausprägungen von lebensweltlicher Bedeutsamkeit.

Im Kern verweist der Stern also nur auf einen engen Teilaspekt des semantischen Umfangs von Signifikanz. Statistisch signifikante Ergebnisse sind keines-

wegs immer und zwingend auch besonders bedeutsam oder wichtig. Allerdings, so unsere These, wird dem Stern gerade als Abkürzung in der wissenschaftlichen Praxis häufig eine Bedeutung zugeschrieben, die auch die alltagssprachliche Semantik umfasst: Signifikante Ergebnisse sind wichtige Ergebnisse und nur wenn statistisch Signifikanz erreicht wird, ist Forschung wichtig genug, um publiziert zu werden.

Der Stern hat in der Wissenschaft, zumindest innerhalb bestimmter disziplinärer Gemeinschaften, also einen Bedeutungsüberschuss erworben. Wenn wir an dieser Stelle kein intentionales Handeln einzelner Forscher_innen oder eine planvolle Strategie unterstellen, so müssen wir fragen, warum sich Forschungs-, Publikations- und Rezeptionspraktiken etabliert haben, die diesen Bedeutungsüberschuss produzieren. Betrachten wir wissenschaftliche Disziplinen als kulturelle Gemeinschaften – mit Lave und Wenger (1991) könnte man auch von Communities of Practice sprechen –, gehen wir davon aus, dass der Stern mit der Zeit zu einem geteilten Symbol geworden ist, dessen Gebrauch eine Funktion für die kulturelle Gemeinschaft erfüllt. Kulturelle Symbole dienen häufig dazu, Zugehörigkeit auszudrücken, sich gegenüber anderen Kulturen abzugrenzen und einzelnen Personen den Zugang zur Gemeinschaft zu ermöglichen oder zu verwehren. Lave und Wenger (1991) sprechen dabei von einer „Legitimate Peripheral Participation“: Neue Mitglieder müssen sich gegenüber der Gemeinschaft bewähren, indem sie bestimmte Praktiken adaptieren. Erst wenn diese Praktiken verinnerlicht sind und (aus der Perspektive etablierter Mitglieder der Gemeinschaft) zufriedenstellend ausgeführt werden, werden die Neuen in den Kern der Gemeinschaft aufgenommen und es wird ihnen erlaubt, höherwertige und gegebenenfalls auch von der Norm abweichende Praktiken auszuüben. Betrachten wir den Stern als kulturelles Symbol, so können wir den konstatierten Bedeutungsüberschuss als symbolische Repräsentation einer Praktik interpretieren, die dabei hilft, den Zugang und die Zugehörigkeit zu einer disziplinären Gemeinschaft zu regeln. Er erfüllt demnach eine wichtige Funktion für die Gemeinschaft, weil er auf eine Fülle dahinterliegender Praktiken (zu forschen und zu publizieren) verweist, damit Ordnung schafft, Komplexität reduziert und Orientierung ermöglicht.

Uns sind keine Daten dazu bekannt, ob Forscher_innen, die keine Signifikanztests verwenden, tatsächlich der Zugang zur Scientific Community verwehrt wurde. Als Narrativ sind die entsprechenden Zusammenhänge im Diskurs allerdings präsent. Oben haben wir bereits Loftus' Polemik gegen den Signifikanztest erwähnt. Er fährt fort:

„Somewhere along the line, however, we all internalized one lesson that is entirely correct: The more you reject the null hypothesis, the more likely it is that you'll get tenure.“ (Loftus 1991: 103; ähnlich Gigerenzer/Krauss/Vitouch 2004; Nosek/Spies/Motyl 2012)

Handlungsspielräume und Alternativen sind jedenfalls erst kürzlich, nach der Replikationskrise, offensichtlich geworden (etwa Chambers 2017; Dienes 2008).

Neben dieser Ordnungs- und Orientierungsfunktion innerhalb des kulturellen Gefüges disziplinärer Gemeinschaften lässt sich auch nach Funktionen des Sterns im Verhältnis zu anderen gesellschaftlichen Subsystemen und Akteuren fragen. Was bedeutet es, wenn die Wissenschaft gegenüber einer nichtwissenschaftlichen Öffentlichkeit betont, man habe (statistisch) signifikante Ergebnisse vorzuweisen? Zunächst ist anzumerken, dass es seit Jahrzehnten einen Trend zur „Verwissenschaftlichung der Sprache des öffentlichen Lebens“ (Polenz 2000: 495) gibt. Prinzipiell ist es also nicht außergewöhnlich, dass ein Begriff aus der wissenschaftlichen Fachsprache zunehmend auch außerhalb der Wissenschaft gebraucht wird. Allerdings kann man fragen, warum Wissenschaftler_innen in der Außendarstellung von Ergebnissen gerade auf die statistische Signifikanz verweisen. Ein Grund könnte darin liegen, die eigene Aussage über den Verweis auf ein technisches Verfahren zu rechtfertigen, anstatt ein Phänomen oder eine empirische Beobachtung durch eine argumentativ haltbare Theorie zu erklären. Damit wird eine Distanz zur eigenen Interpretationsleistung, dem In-Beziehung-Setzen von lebensweltlichem Phänomen, empirischen Beobachtungen und theoretischen Modellierungen hergestellt: Das Verfahren trifft die Aussage, nicht die Wissenschaftler_in. So werden sichere Ergebnisse versprochen, die von der individuellen Interpretations- und Argumentationsleistung der Forscher_innen unabhängig sind. Wie erwähnt, zeigt sich diese Distanzierung auch innerwissenschaftlich im Forschungsartikel.

Schluss

Forschung verspricht immer einen Überschuss und muss dies auch wohl, um ihre gesellschaftliche Bedeutung zu sichern. Der Signifikanzstern trägt hierzu bei, indem er einerseits, wie wir hoffen gezeigt zu haben, die Mehrdeutigkeit des Begriffs *signifikant* ausnutzt, andererseits zentrale Schritte wissenschaftlichen Schließens und Begründens abkürzt und wissenschaftliches Schließen auf eine bestimmte, von Forscher_in, Lebenswelt und Öffentlichkeit gleichermaßen distanzierte Weise interpretiert.

Anmerkungen

- Wir gehen hier nicht auf die Frage des Signifikanzniveaus ein. Das Ergebnis eines Tests ist nicht mehr oder minder signifikant, was Meehl (1978: 78) in die – je nach Lesart zynische oder humorvolle – Formulierung „a difference in the expected direction at the .05 (one asterisk), .01 (two asterisks!), or .001 (three asterisks!) level of significance“ gebracht hat. Das Signifikanzniveau kann gewählt werden, wobei sich .05, .01 und .001 als typische Werte durchgesetzt haben, aber ein Test ist nicht hochsignifikant, wenn er nicht nur das gesetzte Niveau von .05, sondern auch gleich .01 erreicht. Stetig ist der p -Wert, der die Wahrscheinlichkeit angibt, unter Geltung der Nullhypothese dieses oder ein extremeres Ergebnis zu erhalten.
- Es handelt sich bei der heute verbreiteten Form des Signifikanztests um eine inkongruente Mischform aus zwei älteren Verfahren (Gigerenzer 2004).
- In der vereinfachten Ontologie, die dem Signifikanztest zugrunde liegt, gibt es Einflüsse oder es gibt sie nicht. Auch über diese Abkürzung wäre noch einiges zu sagen.
- Vielleicht ist diese Kombination ein Grund, warum Chambers seinem 2017 veröffentlichten Buch *The Seven Deadly Sins of Psychology* das folgende Zitat von F. Scott Fitzgerald voranstellt: „One should be able to see that things are hopeless and yet be determined to make them otherwise.“

Referenzen

- APA Council of Editors (1952). Publication Manual of the American Psychological Society. *Psychological Bulletin*, 49 (4:2), 389–449.
- Bazerman, Charles (1987). Codifying the Social Scientific Style: In John S. Nelson/Allan McGill/Donald N. McCloskey (Hg.), *The Rhetoric of the Human Sciences*, Madison: University of Wisconsin Press, 125–144.
- Becher, Tony/Trowler, Paul (2001). *Academic Tribes and Territories*. 2. Aufl. Buckingham: Open University Press.
- Carver, Ronald P. (1978). The Case against Statistical Significance Testing. *Harvard Educational Review*, 48, 378–399.
- Chambers, Chris (2017). *The Seven Deadly Sins of Psychology*. Princeton: University Press.
- Cohen, Jacob (1994). The Earth is Round ($p < .05$). *American Psychologist*, 49, 997–1003.
- Danziger, Kurt (1985). The Origins of the Psychological Experiment as a Social Institution. *American Psychologist*, 40, 133–140.
- Danziger, Kurt (1987). Statistical Methods and the Historical Development of Research Practice in American Psychology. In Lorenz Krüger/Gerd Gigerenzer/Mary S. Morgan (Hg.), *Ideas in the Sciences, The Probabilistic Revolution*, Bd. 2, Cambridge: MIT Press, 35–47.
- Danziger, Kurt (1996). The Practice of Psychological Discourse. In Kenneth J. Gergen/Carl F. Graumann (Hg.), *Historical Dimensions of Psychological Discourse*, Cambridge: Cambridge University Press, 17–35.
- Danziger, Kurt/Dzinas, Katalin (1997). How Psychology Got Its Variables. *Canadian Psychology*, 38, 43–48.
- Dienes, Zoltan (2008). *Understanding Psychology as a Science*. Houndsmill: Palgrave MacMillan.
- Duden (o.D.). signifikant. *Duden online*. URL: duden.de/node/166971/revision/167007 [1.6.2021]
- Fanelli, Daniele (2011). Negative Results are Disappearing from Most Disciplines and Countries. *Scientometrics*, 90, 891–904.
- Fidler, Fiona (2002). The Fifth Edition of the APA Publication Manual. *Educational and Psychological Measurement*, 62, 749–770.
- Gigerenzer, Gerd (1987). Probabilistic Thinking and the Fight against Subjectivity. In Lorenz Krüger/Gerd Gigerenzer/Mary S. Morgan (Hg.), *The Probabilistic Revolution: Vol. II. Ideas in the Sciences*, Cambridge: MIT Press, 11–33.
- Gigerenzer, Gerd (2004). Mindless Statistics. *The Journal of Socio-Economics*, 33, 587–606.
- Gigerenzer, Gerd/Krauss, Stefan/Vitouch, Oliver (2004). The Null Ritual. In David Kaplan (Hg.), *The Sage Handbook of Quantitative Methodology for the Social Sciences*, Thousand Oaks: Sage, 391–408.
- Gigerenzer, Gerd/Swijtink, Zeno G./Porter, Theodore/Daston, Lorraine/Beatty, John/Krüger, Lorenz (Hg.) (1989). *The Empire of Chance: How Probability Changed Science and Everyday Life*. Cambridge: Cambridge University Press.
- Haller, Heiko/Krauss, Stefan (2002). Misinterpretations of Significance. *Methods of Psychological Research Online*, 7, 1–20.
- Ioannidis, John P. A. (2005). Why Most Published Research Findings Are False. *PLoS Medicine*, 2, e124.
- John, Leslie K./Loewenstein, George/Prelec, Drazen (2012). Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling. *Psychological Science*, 23, 524–532.
- Kerr, Norbert L. (1998). HARKing: Hypothesizing After the Results are Known. *Personality and Social Psychology Review*, 2, 196–217.
- Lave, Jean/Wenger, Etienne (1991). *Situated Learning: Legitimate Peripheral Participation*. Cambridge: Cambridge University Press.
- Loftus, Geoffrey R. (1991). On the Tyranny of Hypothesis Testing in the Social Sciences. *Contemporary Psychology*, 36, 102–105.
- Loftus, Geoffrey R. (1993). Editorial Comment. *Memory & Cognition*, 21, 1–3.
- Madigan, Robert/Johnson, Susan/Linton, Patricia (1995). The Language of Psychology. *American Psychologist*, 50, 428–436.
- Meehl, Paul E. (1978). Theoretical Risks and Tabular Asterisks. *Journal of Consulting and Clinical Psychology*, 46, 103–115.
- Meehl, Paul E. (2006). The Power of Quantitative Thinking. In Niels G. Waller/Leslie J. Yonce/William M. Grove/David Faust/Mark F. Lenzenweger (Hg.), *A Paul Meehl Reader*, Mahwah: Erlbaum, 433–444.
- Melton, Arthur W. (1962). Editorial. *Journal of Experimental Psychology*, 64, 553–557.
- Nosek, Brian A./Spies, Jeffrey R./Motyl, Matt (2012). Scientific Utopia: II. Perspectives on Psychological Science, 7, 615–631.
- Onwuegbuzie, Anthony J./Leech, Nancy L. (2004). Enhancing the Interpretation of “Significant” Findings. *The Qualitative Report*, 9, 770–792.
- Polenz, Peter von (2000). *Deutsche Sprachgeschichte vom Spätmittelalter bis zur Gegenwart. Bd. I: 14. bis 16. Jahrhundert*. Berlin: De Gruyter.
- Rucci, Anthony J./Tweney, Ryan D. (1980). Analysis of Variance and the “Second Discipline” of Scientific Psychology. *Psychological Bulletin*, 87, 166–184.
- Simmons, Joseph P./Nelson, Leif D./Simonsohn, Uri (2011). False-Positive Psychology. *Psychological Science*, 22, 1359–1366.
- Sterling, Theodore D. (1959). Publication Decisions and Their Possible Effects on Inferences Drawn from Tests of Significance – or vice versa. *Journal of the American Statistical Association*, 54, 30–34.

