Eine überlebenswichtige Warnung … interessiert?

Robert Schleusener

Zusammenfassung
Dieser Aufsatz möchte auf einen Alarmruf aufmerksam machen, der seit Jahrzehnten weltweit ungehört verhallt. Statistiker kritisieren die Anwendung des Signifikanztests und des p-Werts – in der heute gebräuchlichen Form – scharf. Diskussionen der letzten Jahre gipfelten 2016 in der Veröffentlichung eines Statements der American Statistical Association (ASA) und Anfang 2019 in der Herausgabe einer umfangreichen Sonderausgabe des „The American Statistician". Der Leitartikel dieser Sonderausgabe „Moving to a World Beyond p <0.05" sowie einige andere in diesem Kontext erschienene Artikel sind Grundlage dieses Aufsatzes, der die gewünschte breite Diskussion unterstützen und sie in die deutschsprachige osteopathische Szene tragen möchte. Dafür wurden einige Artikel ins Deutsche übersetzt, die der Autor Interessierten bei Anfrage an die unten angegebene Mailadresse gerne zusendet.

Schlüsselwörter
ASA-Statement, statistische Signifikanz, p-Wert, Reproduzierbarkeitskrise, Replizierbarkeitskrise

Abstract
This essay aims to draw attention to an alarm call that has been echoing unheard around the world for decades. Statisticians sharply criticize the use of the significance test and the p-value – in today's common form. Discussions in recent years culminated in 2016 with the publication of a statement by the American Statistical Association (ASA) and in early 2019 with the publication of an extensive special edition of "The American Statistician". The editorial of this special edition „Moving to a World Beyond p <0.05", as well as some other articles published in this context, are the basis of this essay, which aims to support the desired broad discussion and bring it into the German-speaking osteopathic scene. Some articles have been translated into German, which the author would be happy to send to interested parties on request to the e-mail address provided.

Keywords
statistical significance, p-value fallacy, replicability crisis, reproducibility crisis

Wir sinken … na und?

Während Sie gerade darüber nachdenken, warum in den Suppen Ihrer Tischnachbarn, neben all den Gemüsestreifen, sechs Muscheln schwimmen, obwohl auf Ihrem Teller nur vier sind, kehrt Thomas Andrews zurück. Der Schiffbauingenieur hat nach der leichten Erschütterung, die alle vor ein paar Minuten gespürt haben, nach dem Rechten gesehen. Er sagt: „Wir haben einen erheblichen Schaden erlitten, die Titanic *wird sinken*." Wären Sie nicht überrascht, sich antworten zu hören: „Ach wissen Sie, bei meinen Schiffsreisen bin ich an Details nicht interessiert. Hauptsache wir erreichen pünktlich New York!"?
Die Gemeinschaft der Statistiker warnt seit Jahrzehnten vor einem Problem, das durchaus dem Alarmruf „Wir sinken" entspricht, und sie beklagt die Reaktionen der Adressaten, die durch die Warnung in keiner Weise beunruhigt scheinen.

Wie kann das sein?

Das eigene Handeln zu hinterfragen und gegebenenfalls zu ändern fällt menschlichen Gehirnen naturgemäß schwer. Dieser Umstand ist Ziel wissenschaftlicher Forschung. Amos Tversky und Daniel Kahneman sind (bzw. waren; Tversky verstarb 1996) herausragende Wissenschaftler auf diesem Gebiet. Bereits 1971 beschrieben sie in dem Artikel *„Belief in the law of small numbers"* den erstaunlichen Tatbestand, dass die Intuitionen, die Menschen bezüglich der Gesetze des Zufalls haben, fehlerhaft sind. Das drücke sich beispielhaft darin aus, dass Menschen „… eine Stichprobe, die nach dem Zufallsprinzip aus einer Population gezogen wurde, als *hochgradig repräsentativ*, das heißt, in allen wesentlichen Charakteristika ähnlich mit der Ursprungspopulation [betrachten] …". Für Stichproben gilt das statistische „Gesetz der *großen* Zahlen". Es besagt, dass eine Stichprobe mit zunehmender Größe in ihren wesentlichen Eigenschaften der Ursprungspopulation – bzw. Grundgesamtheit – immer ähnlicher wird [1]. Das ist eine vernünftige Vorstellung.
Wenn ich mit einer Kelle eine Stichprobe aus einem Suppentopf nehme, wird die Verteilung der verschiedenen Gemüse, von Fleisch und Muscheln wahrscheinlich nicht der Zusammensetzung im Topf entsprechen. Aber je größer ich die Kelle wähle, umso mehr entspricht die Zusammensetzung derjenigen im Topf (umrühren nicht vergessen!). Am Ende könnte die Kelle so groß sein, dass sie den gesamten Inhalt des Topfes fasst. Dann ist es aber keine Stichprobe mehr. Im Sinne des Gesetzes ist das die größte Zahl, die man bekommen kann. Wer *alles* untersuchen könnte, würde keine Stichprobe brauchen … *aber:* Die Welt ist schon ein ganz schön *großer* Suppentopf.
Das psychologisch begründbare „Gesetz der *kleinen* Zahlen", das Tversky und Kahneman entdeckt haben, besagt nun, dass Menschen diese Abbildung tatsächlicher Verhältnisse durch große Stichproben ganz selbstverständlich auch von kleinen erwarten. Sie sind also sicher, dass auch Stichproben geringer Größe die Charakteristika ihre Ursprungspopulation sehr gut abbilden.
Das ist eine unvernünftige Vorstellung! Jeder, der erst mit einem Teelöffel und

danach mit einer großen Kelle eine „Stichprobe“ aus einer Grundgesamtheit Suppentopf nimmt, kann das erkennen.

Ein mächtiges Dilemma

Kahneman und Tversky entwickelten gemeinsam die „Prospect Theory“, die 2002 mit dem Nobelpreis für Wirtschaft gewürdigt wurde. Die Essenz ihrer Zusammenarbeit beschreibt Kahneman in dem Buch „Schnelles Denken, langsames Denken“. Eine überragende Bedeutung spielen dabei die sogenannten kognitiven Verzerrungen („cognitive bias“). Darum zu wissen ist essenziell, da bewiesen werden konnte, dass menschliche Gehirne von gefühlten Wahrheiten oder gefühlten Zusammenhängen so überzeugt sind, dass nur mit sauberer statistischer Aufarbeitung die Chance besteht, einen Blick auf die tatsächlich bestehenden Zusammenhänge zu erhaschen. Dummerweise, so Kahneman, sind *unsere Gehirne* nicht dafür gemacht *statistisch zu denken*. Dies ist eine mühsame und energiezehrende Funktion des „langsamen Denkens“.
Wir müssen z.B. gegen das „Gesetz der kleinen Zahlen“, das sich für *uns* doch so richtig anfühlt, kontraintuitiv andenken. Man kann vermuten, dass hierin ein Grund dafür zu finden ist, warum nur wenig Interesse besteht an der „*Lehre über Methoden zum Umgang mit quantitativen Informationen und der Möglichkeit systematische Verbindung zwischen Erfahrung und Theorie herzustellen*“ [2] – oder kurz gesagt an „Statistik“.
Seien wir ehrlich, die Anstrengung, die Statistik hinter einer x-beliebigen Veröffentlichung zu verstehen, erlahmt schnell. Man gibt sich damit zufrieden, dass das schon irgendwie stimmen wird, und beruhigt sich damit, dass die Fallzahl ja ordentlich hoch ist und/oder dass der p-Wert ein „signifikantes“, wenn nicht sogar „hoch signifikantes“ Ergebnis belegt.
p ist kleiner als 0,01. Das muss doch so etwas bedeuten wie zu 99% sicher und reproduzierbar … oder nicht?!?

Wir stecken – ohne es angemessen wahrzunehmen – in einem *mächtigen* Dilemma

Zum einen können unsere Gehirne nur unter großen Anstrengungen „statistisch denken“. Damit ist das Entscheiden auf der Grundlage fundierter, durchdachter und verstandener Daten immer ein anstrengender Prozess. Nach Gefühl zu entscheiden fällt wesentlich leichter. Was nicht bedeutet, dass das immer falsch ist … aber wenn doch, fühlt es sich dummerweise nicht falsch an. Kahneman bezeichnet dies als „schnelles Denken“. „Schnelles Denken“ findet in unseren Gehirnen ununterbrochen statt, da wir ja zu jeder Zeit die Situation um uns herum bewerten müssen: Droht Gefahr? Kann man das essen? Ist das eine interessante Chance für mich?
Zum anderen wenden wir statistische Werkzeuge an, die in dem Ruf stehen zuverlässige Daten zu identifizieren: die Signifikanz und den p-Wert. Diese Werkzeuge werden aber oft nicht richtig eingesetzt. Daher sind die Informationen trotz aller Bemühungen verzerrt.
Ioannidis verdeutlichte das 2005 mit einem Beispiel: 100.000 Gene sollen auf einen möglichen Beitrag zu einer Krankheit getestet werden. *Geschätzt* wird, dass höchstens 10 Gene wirklich beteiligt sind (10/100.000=0,0001).
Wenn ein Gen jetzt mit dem p-Wert von 0,05 als mitverantwortlich identifiziert wird, stellt sich natürlich die Frage, wie groß die Wahrscheinlichkeit ist, dass das wirklich stimmt?
Unter der Annahme, dass $p<0,05$ eine 95%ige Wahrscheinlichkeit beschreibt, wird oft in 19 von 20 Fällen geantwortet (19/20=0,95). Dabei wird nicht bedacht, dass ja auch jedes 20. unbeteiligte Gen ein (falsch-positives) „signifikantes“ Ergebnis erbringt.
Bei 100.000 Genen findet man also nicht nur die 10 wirklich an der Krankheitsentstehung beteiligten, sondern auch 5.000 „falsch-positive“.
Die Wahrscheinlichkeit, dass eines der gefundenen Gene zu den Verursachern gehört, beträgt also nur 10 von 5.000, oder dezimal ausgedrückt: 0,002 statt der angenommenen 0,95 [3]. Aber das sagt natürlich keiner, wenn er ein signifikantes Ergebnis verkündet.
Fatalerweise hat sich das Signifikanzniveau trotzdem zu einem „Alles-oder-Nichts-Kriterium“ entwickelt, das darüber entscheidet, ob wissenschaftliche Ergebnisse der Öffentlichkeit zugänglich gemacht werden.
Das ist ein ganz schönes Loch in der Flanke unseres Kahns! Ist jemand bis hierhin beunruhigt?

Signifikanz wird signifikant falsch verstanden
Das Signifikanzniveau ist die Irrtumswahrscheinlichkeit; die Wahrscheinlichkeit irrtümlich die Hypothese „es gibt einen Effekt“ bestätigt zu sehen, obwohl der untersuchte Effekt nicht existiert.

Das Signifikanzniveau, das vor dem Test festgelegt wird, indem unter anderem die Wahrscheinlichkeit des zu untersuchenden Effektes geschätzt (!) wird, besagt nicht, mit welcher Wahrscheinlichkeit eine Hypothese richtig ist. Es ist nicht identisch mit dem p-Wert (der nach dem Test berechnet wird), der ein Hinweis dafür ist, wie extrem das Ergebnis ist: Je kleiner der p-Wert, desto kleiner ist die Wahrscheinlichkeit, dass es keinen Effekt gibt [4, 5].
Sie verstehen mich nicht? Das kann ich gut verstehen! Geben Sie nicht auf und lassen Sie sich von Regina Nuzzo trösten. Die Statistikerin veröffentlichte im Februar 2014 in „Nature“ einen Artikel, den die Zeitschrift „Spektrum der Wissenschaft“ unter dem Titel „Umstrittene Statistik: Wenn Forscher durch den Signifikanztest fallen“ noch im selben Monat dem deutschsprachigen Raum im Internet frei zugänglich machte.

Nuzzo beginnt ihren Artikel mit folgendem markantem Beispiel, das zeigt, wie leicht der Signifikanztest seine Nutzer in die Irre führen kann [6]: 2010 entdeckte der Psychologiedoktorrand M. Motyl, dass im Vergleich mit politisch moderaten Personen sowohl rechts- wie linksextrem eingestellte Menschen bestimmte Grauschattierungen schlechter unterscheiden können. Die Stichprobe des Experiments war mit 2000 Probanden beeindruckend und der p-Wert adelte das Ergebnis mit einem Wert von 0,01 als „hoch signifikant".
Motyl und sein Betreuer Brian Nosek waren jedoch „an den Details interessiert". Die Überprüfung ihrer Analyse, inklusive der Berücksichtigung zusätzlicher Daten, kostete sie die spektakuläre Veröffentlichung. Der p-Wert lag nämlich nun bei 0,59 und damit nahe der „Fifty-fifty-Schwelle", an der scheinbare Korrelationen so gut gedeihen, ohne replizierbar zu sein. So wie sich nach einem Münzwurf niemand wundert, dass auf Zahl Wappen folgt, oder doch Zahl … wie der Zufall so spielt [7].

„Es muss was geschehen …!"

2014 nahm der Vorstand der „American Statistical Association" (ASA) umfangreiche Vorarbeiten für eine Grundsatzerklärung zur statistischen Signifikanz und dem Gebrauch des p-Wertes in Angriff. Erklärtes Ziel war, die immer und immer wieder geäußerte Kritik am p-Wert durch ein Statement, das aus einer breiten Fachdiskussion hervorgegangen sein sollte, zu bereichern.
Zu den jüngeren Diskussionsbeiträgen, auf welche die ASA reagierte, gehörte unter vielen anderen der oben vorgestellte Artikel „Statistical Errors" von Regina Nuzzo, der inzwischen zu den meistgelesenen „Nature"-Artikeln gehört [8]. In der Woche des Erscheinens von „Statistical Errors" postete der Statistiker Jeff Leek:

> „Das Problem ist nicht, dass die Menschen die p-Werte schlecht verwenden, es ist, dass die überwiegende Mehrheit der Datenanalyse nicht von Personen durchgeführt wird, die richtig geschult sind, um Datenanalysen durchzuführen."
> *([9]; zit. nach [11])*

Aber auch ältere Beiträge „beflügelten" den ASA-Vorstand, wie T. Siegfried, der 2010 schrieb:

> „Es ist das schmutzigste Geheimnis der Wissenschaft: Die ‚wissenschaftliche Methode' zum Testen von Hypothesen durch statistische Analysen steht auf einem schwachen Fundament."
> *([10]; zit. nach [11])*

Der ASA-Vorstand sah sich in der Pflicht, weil „viel Verwirrung und sogar Zweifel an der Gültigkeit der Wissenschaft zunehmen" [11].
Professor Nuzzo hilft uns dies zu verstehen. Sie führt in ihrem Aufsatz aus: Ronald Fisher, der den Signifikanztest in den 1920er-Jahren ersann, habe nie eine Signifikanz, wie wir sie heute verstehen, im Sinn gehabt. Er habe das Verfahren als Teil eines nichtmathematischen Prozesses gesehen, in dem Forscher unter anderem ein Werkzeug nutzen, mit dem sie einschätzen können, ob ihre Daten durch Zufall erklärbar sind *oder eben nicht*. Denn erst wenn Daten wahrscheinlich nicht zufällig sind, sind sie bedeutsam genug für eine genauere Untersuchung. Das bedeutet, Fisher sah den p-Wert als ein relativ grobes Sieb am Anfang einer Untersuchung und nicht als krönenden Abschluss an ihrem Ende.
Für die Bewertung der Bedeutsamkeit einer Untersuchung waren für Fisher zudem andere Komponenten unverzichtbar. Nicht zuletzt das Hintergrundwissen der Forscher, die dadurch abschätzen können, wie plausibel oder wahrscheinlich das Phänomen angenommen werden kann, das sich durch aufregende erste Ergebnisse in den Fokus drängt. Denn je unplausibler die Hypothese, umso wahrscheinlicher wird ein Fehlalarm – *auch* bei niedrigem p-Wert.
Außerdem, so Nuzzo, ist die Entwicklung, die zu den heutigen Missverständnissen führte, nur durch die Rivalität zwischen Fisher und dem Mathematiker J. Neyman sowie dem Statistiker E. Pearson zu verstehen. Beide waren prominente Vertreter einer Bewegung, die nach Werkzeugen für evidenzbasierte Entscheidungen suchte.
Der Schlagabtausch zwischen diesen hervorragenden Denkern war so ausdauernd, das pragmatischere Wissenschaftler in der Zwischenzeit Statistikhandbücher für ihre Kollegen schrieben – gleichwohl ohne immer die Feinheiten der unterschiedlichen Methoden durchdrungen zu haben. Dadurch entstand eine Mischung aus Fishers p-Wert und dem regelbasierten System von Neyman und Pearsson – *und* der p-Wert von 0,05 als entscheidende Signifikanzschwelle, bekam seine alles überragende Bedeutung, die er *so* nie hätte haben sollen [6]. Denn der p-Wert ist keineswegs so verlässlich oder objektiv, wie es viele gerne hätten. Nach Ansicht der Kritiker des p-Wertes könnte sogar die Mehrzahl der veröffentlichten Ergebnisse falsch sein, eben, weil das zugrundeliegende Konzept traditionell falsch verstanden wird [12].
Nach der Devise „schlimmer geht immer" hat sich zudem eine Praxis entwickelt, die der Psychologe Uri Simonsohn unter dem Begriff „p-Hacking" bekannt gemacht hat [13]. Dies bedeutet: ausprobieren, bis das gewünschte Ergebnis erhalten wird – ein veröffentlichungswürdiges Signifikanzniveau.
Hensel et al. stellten mit *PROMOTE* 2015 eine randomisierte kontrollierte Studie mit 400 Frauen im letzten Schwangerschaftsdrittel vor. Primärer Parameter war der Einfluss osteopathischer Behandlung bei Rückenschmerzen und -funktion. Durch unerwartet viele Studienabbrecherinnen wurde die Probandenzahl unterschritten, die zu Beginn als Mindestzahl festgelegt worden war. Die aufwendige Studie wurde trotzdem beendet und die Daten inten-

siv statistisch bearbeitet. Hensel et al. kommen zu dem Schluss, dass die Befunde darauf hindeuten, dass signifikante Behandlungseffekte für Rückenschmerzen und -funktion bestehen ($p<0{,}001$).

Aber nicht nur die zusätzlich zur Standardvorsorge osteopathisch behandelte Gruppe war im Vergleich zur reinen Standardvorsorge signifikant verbessert, sondern auch die Placebogruppe (Standardversorgung und Scheinbehandlung mit einem nicht aktiven Ultraschallkopf). In der Diskussion wird erwogen, dass die Scheinbehandlung vielleicht auch einen therapeutischen Effekt hatte [14] … was bleibt, ist $p<0{,}001$.

Da schon gezeigt wurde, dass Hypothesen wie „Fallschirme haben beim Springen aus einem Flugzeug keinen nachweisbaren Nutzen" [15, 16] oder „Schokolade hilft beim Abnehmen" [17] durchaus mit einem Signifikanzniveau belegbar sind, das heute notwendig ist, um in einem renommierten Journal veröffentlicht zu werden, scheint die Missbilligung des Konzeptes nicht ganz unbegründet.

In dieser Situation (Sie erinnern sich? Wir sinken!) auf die Warnungen der Fachleute zu hören, kann radikale Entscheidungen zur Folge haben, wie sie die Herausgeber von „Basic and Applied Social Psychology" trafen. Sie entschieden sich, p-Werte zu verbieten … ([18]; zit. nach [11]).

Erst vor wenigen Monaten veröffentlichte „Nature" den Aufsatz „Retire statistical significance" [12]. Amrhein, Greenland und McShane fordern hier dazu auf, den p-Wert als Signifikanzkriterium aufzugeben – gemeinsam mit mehr als 800 weiteren Fachleuten, die sich per Unterschrift zu der Kernaussagen dieses Aufsatzes bekennen. Auch diesen Beitrag machte „Spektrum der Wissenschaft" dem deutschsprachigen Raum zeitnah zugänglich [19].

Die Autoren sagen, dass sie die Verwendung des p-Wertes in speziellen Situationen, wie z.B. der Überwachung von Qualitätsstandards in einem Produktionsprozess, durchaus für sinnvoll halten. Sie wollen auch nicht so verstanden werden, dass das, was bis jetzt nicht glaubwürdig war, nun automatisch glaubwürdig sein soll. Vielmehr rufen sie dazu auf p-Werte nicht länger auf die traditionelle „Entweder-oder"-Art zu verwenden, also um eine Hypothese entweder zu bestätigen oder zu widerlegen [18]. Die deutschen Herausgeber weisen treffend darauf hin, dass „die statistische Signifikanz … so tief in der wissenschaftlichen Praxis verankert [ist], dass ein Verzicht darauf schmerzhaft wäre …" und die Redaktion von „Nature" daher aktuell auch nicht ihre Bewertung statistischer Analysen von Artikeln zu verändern plane. Aber das Publikum wird aufgefordert, seine Meinung zu äußern. Die Diskussion scheint durch das ASA-Statement also wirklich neue Power bekommen zu haben.

Eine wichtige Rolle spielen in der Zukunft natürlich die Höheren Schulen und Universitäten, die durch ihre statistischen Lehrpläne die Realität der praktischen Arbeit von Wissenschaftlern, aber z.B. auch von Journalisten oder Politikern maßgeblich beeinflussen.

Die Grundsatzerklärung der ASA wird daher nicht ohne Grund mit einer „lustigen kleinen Geschichte" eingeleitet. Auf einem Diskussionsforum wurde gefragt, warum $p=0{,}05$ immer noch unterrichtet werde? Die Antwort war: Weil das immer noch das sei, was verwendet werde. „Und warum verwenden so viele Menschen immer noch $p=0{,}05$?" „Weil es das ist, was ihnen beigebracht wurde" [11].

So ist die ASA zu der Überzeugung gekommen, dass es nicht mehr genügt immer und immer wieder zu rufen:

„Tut es nicht!"
… begründet Schlussfolgerungen *nicht ausschließlich* damit, dass ein Schwellenwert wie $p<0{,}05$ unterschritten wurde!
… glaubt *nicht*, dass ein Effekt existiert, nur weil er statistisch signifikant ist.
… glaubt *nicht*, dass ein Effekt nicht existiert, nur weil er statistisch nicht signifikant war.
… glaubt *nicht*, dass der p-Wert die Wahrscheinlichkeit angibt, dass der beobachtete Effekt allein durch Zufall hervorgerufen wurde, oder die Wahrscheinlichkeit, dass die Testhypothese wahr ist.
… und *zieht keine* Schlussfolgerungen zur wissenschaftlichen oder praktischen Bedeutung auf der Grundlage statistischer Signifikanz (oder deren Fehlen).
~~[20]~~

Und darum bezieht die ASA jetzt sehr deutlich Stellung.

Diese Diskussion müssen wir als Osteopathen verstehen, und wir müssen uns an ihr beteiligen.

Was einige Fragen aufwirft:

- Ist es angesichts der hier vorgestellten Warnung (global wird von der Reproduzierbarkeitskrise und Replizierbarkeitskrise gesprochen) nicht sinnvoller über die Plausibilität unserer Konzepte zu diskutieren, bevor wir versuchen „signifikante" Forschung zu betreiben – besonders, da wir mit unseren Konzepten ja nicht selten Plausibilitätsvorstellungen anderer Fachbereiche berühren?
- Haben wir bis auf Weiteres mehr in die Waagschale zu werfen als unsere subjektive Behandlungserfahrung? Und wäre das zum jetzigen Zeitpunkt so schlimm, *wenn* wir diese subjektive Expertise so aufarbeiten, dass sie wissenschaftlich verwertbar wird? Sollten wir nicht, wie es in der Wissenschaft selbstverständlich ist, Arbeiten von Kolleginnen und Kollegen dadurch würdigen, dass wir ihre Ergebnisse überprüfen, anstatt immer neue Forschungsfragen zu kreieren?
- Wie können wir statistische Werkzeuge so einsetzen, dass sie auch mit den begrenzten Möglichkeiten funktionieren, die uns in der Regel zur Verfügung stehen? Wäre z.B. die saubere Aufarbeitung von Daten mit den Mitteln der beschreibenden Statistik (also Tabellen, Diagrammen, Mittelwert, Median, Modus etc.) vorerst nicht besser, als mit Wahrscheinlichkeitsrechnungen fragwürdige Signifikanzen zu produzieren?

Osteopathie ist (noch?!) nicht universitär. Den allermeisten von uns fehlen die Ressourcen, um wissenschaftlich zu arbeiten, und es ist auch nicht verwerflich, wenn man so eine Arbeit nicht machen möchte. Dass es Einige gebetsmühlenartig immer weiter fordern, weil sie glauben, dass es der Osteopathie auf ihrem Weg zu Akzeptanz, Anerkennung und Berufsbild nützt, wird der aktuellen Situation nicht gerecht. Hier entlasten uns ein Stück weit das ASA-Statement – und John P.A. Ioannidis. Als wenn er die Osteopathie ansprechen würde, resümierte Ioannidis 2005 in seinem oft zitierten Artikel „Why most published research findings are false“:

> „In diesem Rahmen ist ein Forschungsergebnis weniger wahrscheinlich, wenn die in einem Bereich durchgeführten Studien kleiner sind; wenn die Effektgrößen kleiner sind; wenn es eine größere Anzahl und geringere Vorauswahl an getesteten Beziehungen gibt; wenn es eine größere Flexibilität bei Designs, Definitionen, Ergebnissen und Analysemodi gibt …“
> *[3]*

- Und ist es vor diesem Hintergrund nicht schon fast lächerlich, erfahrene Therapeuten – die keine wissenschaftliche Arbeit verfasst haben – aus diesem wichtigen Diskurs zu verdrängen, weil man ihnen versagt ihre Erfahrungen im Austausch mit anderen Dozentinnen und Dozenten im Unterrichten weiterzugeben?

Ende oder Anfang?

Signifikanz ist seit Jahrzehnten die Allzweckwaffe im Kampf um wissenschaftliche, politische und gesellschaftliche Anerkennung sowie im Kampf um gesellschaftliche Ressourcen. Signifikanz konnte nur dazu werden, weil die Testung des Signifikanzniveaus und die daraus abgeleiteten Schlussfolgerungen oft falsch eingesetzt wurden und werden – und das gegen den Rat von Fachleuten. Trotzdem wurden und werden auf der Grundlage sogenannter „hoch signifikanter“ Ergebnisse wissenschaftliche, politische, wirtschaftliche und gesellschaftliche Entscheidungen gefällt. Diese Entscheidungen müssen nicht falsch sein, doch das hat dann nicht unbedingt etwas mit dem Signifikanzniveau zu tun … aber das wissen die Entscheidungsträger oft nicht.

Also lassen Sie uns einfach die Anweisungen des Personals beachten und die Rettungsboote besetzen, bevor das Schiff untergeht. Rudern wir in die unbekannte, dunkle Nacht hinaus, mal sehen, was wir dort finden können.
Natürlich werden wir in der Ferne das Schiff noch lange sehen. Einige werden daher zu Recht fragen, ob es nicht gemütlicher wäre, dort die Arbeit fortzusetzen, bis es nicht mehr geht. Das ist eine selbstbestimmte Entscheidung. Aber der lädierte Kahn *wird* untergehen und dann reißt er die mit, die noch an Bord sind … sagen die Experten.

Korrespondenzadresse:

Robert Schleusener
Tibusstraße 1a
48143 Münster
info@praxis-schleusener.de

Weitere Leseempfehlung

Peng, R. (2015) The reproducibility crisis in science: a statistical counterattack. Significance 12(3): 30–32. Im Internet: https://rss.onlinelibrary.wiley.com/doi/full/10.1111/j.1740-9713.2015.00827.x

Literatur

[1] Tversky, A., Kahneman, D. (1971) Belief in the law of small numbers. Hebrew University of Jerusalem, Psychological Bulletin, 1971, Vol. 76, No. 2. 105–110
[2] Rinne, H. (2008) Taschenbuch der Statistik. 4., vollständig überarb. und erw. Aufl. Deutsch, Frankfurt, M., ISBN 978–3-8171–1827–4, S. 1
[3] Ioannidis JPA (2005) Why most published research findings are false. PLoS Med 2(8): e124
[4] https://www.umm.uni-heidelberg.de/inst/biom/lexikon/data/p006.html
[5] https://www.umm. uni-heidelberg.de/inst/biom/lexikon/data/i018.html
[6] Nuzzo, Regina (2014) Wenn Forscher durch den Signifkanztest fallen [der Artikel wurde unter dem Titel „Statistical errors“ in Nature 506 erstveröffentlicht (S. 150–152, 2014) – Veröffentlichung in Deutsch in: Spektrum – Die Woche, 8. KW 2014
[7] Nosek, B.A., Spies J.R., Motyl, M. (2012) Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. Perspect Psychol Sci 7: 615–631
[8] http://www.altmetric.com/details/2115792#score
[9] Leek, J. (2014), „On the Scalability of Statistical Procedures: Why the p-Value Bashers Just Don't Get It, ”Simply Statistics Blog, Available at http://simplystatistics.org/2014/02/14/on-the-scalability-of-statistical-procedures-why-the-p-value-bashers-just-dont-get-it/
[10] Siegfried, T. (2010), „Odds Are, It's Wrong: Science Fails to Face the Shortcomings of Statistics,” Science News, 177, 26. Available at https://www.sciencenews.org/article/odds-are-its-wrong
[11] Wasserstein, Ronald L.; Lazar, Nicole A. (2016) The ASA's Statement on p-Values: Context, Process, and Purpose, The American Statistician, 70:2, 129–133. doi: 10.1080/00031305.2016.1154108 To link to this article: https://doi.org/10.1080/00031305.2016.1154108
[12] Amrhein, V. et al. (2019) Retire statistical significance. Nature 567, unter Verwendung der Daten von Schatz, P. et al. (2005) Archives of Clinical Neuropsychology 20, Fidler, F. et al. (2006) Conservation Biology 20, Hoekstra, R. et al. (2006) Psychonomic Bulletin & Review 13, Bernardi, F., Chakhai, L., Leopold, L. (2017) Sing me a song with social significance': the (mis)use of statistical significance testing in European sociological research. European Sociological Review 33, 1: 1–15, https://doi.org/10.1093/esr/jcw047
[13] Simmons, J., Nelson, L., Simonsohn, U. (2011) False-positive psychology: Undisclosed flexibility in data collection and analysis allow presenting anything as significant. Psychological Science 22 (11): 1359–66. doi: 10.1177/0956797611417632
[14] Hensel, K.L., Buchanan, St., Brown, S.K., Rodriguez, M., Cruser, A. (2015) Pregnancy Research on Osteopathic Manipulation Optimizing Treatment Effects: The PROMOTE Study A Randomized Controlled Trial. Am J Obstet Gynecol 212 (1): 108.e1–9 doi: 10.1016/j.ajog.2014.07.043
[15] Smith, G.C., Smith, S.; Pell, J.P. (2003), Parachute use to prevent death and major trauma related to gravitational challenge: systematic review of randomised controlled trials, BMJ 327: 1459–1461
[16] Yeh, R.W., Valsdottir, L.R. et al. (2018) Parachute use to prevent death and major trauma when jumping from aircraft: randomized controlled trial. BMJ 2018;363: k5094. doi: 10.1136/bmj.k5094
[17] https://upload.wikimedia.org/wikipedia/commons/3/39/Chocolate_with_high_Cocoa_content_as_a_weight-loss_accelerator.pdf
[18] Trafimow, D., Marks, M. (2015) Editorial. Basic and Applied Social Psychology 37, 1–2
[19] Amrhein, V., Greenland, S, McShane, B. (2019) Schickt die statistische Signifikanz in den Ruhestand! Veröffentlichung in Deutsch. Spektrum der Wissenschaft 5/2019
[20] Wasserstein, R.L., Schirm, A.L., Lazar, N.A. (2019) Moving to a World Beyond p<0.05. The American Statistician; 73, S1: 1–19, Editorial: https://doi.org/10.1080/00031305.2019.1583913