Spektrum.de

| 19.02.2014 UMSTRITTENE STATISTIK

Wenn Forscher durch den Signifikanztest fallen

Grobe Fehler in Statistik: Der "p-Wert" gilt als Goldstandard, doch er führt in die Irre. Er schadet damit seit Jahren der Wissenschaft.

Regina Nuzzo

EXKLUSIVE ÜBERSETZUNG AUS nature Für

einen kurzen Augenblick stand Matt Motyl an der Schwelle zum wissenschaftlichen Ruhm. 2010 entdeckte er bei einem Experiment, dass Extremisten die Welt in Schwarz und Weiß sehen – und das ganz buchstäblich.

Seine Ergebnisse waren "vollkommen eindeutig", erinnert sich der Psychologiedoktorand an der University of Virginia in Charlottesville. Seine Studie an 2000 Probanden ergab, dass die Links- oder Rechtsextremen unter den Teilnehmern diverse Grauschattierungen schlechter unterscheiden können als politisch moderater eingestellte Menschen.

"Das war nicht nur eine sexy Hypothese", sagt Motyl, "sie war auch durch die Befunde eindeutig gestützt." Das zeigte sich, als er den entsprechenden p-Wert errechnete – eine gängige Art, um die Güte

eines statistischen Belegs zu beziffern. Er lag bei 0,01 – das gilt im Allgemeinen als "hochsignifikant". Die Veröffentlichung in einem Topjournal war zum Greifen nah.

Dann aber kam ihm die Realität in die Quere. Motyl und sein Betreuer Brian Nosek wussten um den Streit über die mangelnde Reproduzierbarkeit wissenschaftlicher Studien und waren vorsichtig geworden. Besser noch einmal überprüfen, dachten sie sich, und bezogen zusätzliche Daten in die Analyse ein.

Prompt sprang der p-Wert auf 0,59 – nicht einmal in die Nähe der Schwelle, ab der ein Ergebnis als signifikant gilt: Unter 0,05 hätte er liegen müssen. So aber verschwand der Effekt. Und mit ihm Motyls Traum vom frühen Ruhm [1].

Der p-Wert ist keineswegs so sicher oder objektiv wie gedacht

Wie sich später herausstellte, lag es nicht an Motyls Daten oder einem Rechenfehler. Schuld war die trügerische Natur des p-Werts selbst. Der ist nämlich keineswegs so verlässlich oder objektiv, wie es viele Wissenschaftler gerne hätten. "P-Werte leisten nicht, was sie sollen. Ganz einfach, weil sie das nicht können", sagt der Wirtschaftswissenschaftler Stephen Ziliak von der Roosevelt University in Chicago. Er zählt zu den bekanntesten Kritikern der Art, wie Wissenschaftler heutzutage Statistik treiben.

© FOTOLIA / WRANGLER (AUSSCHNITT)

Was sagt die Statistik? | Manchmal nicht das, was man meint. Der p-Wert ist heutzutage allgegenwärtig, sagt aber gar nicht das, was viele meinen. Das hat zur Publikation vieler falscher Ergebnisse geführt.

Für Forscher sind Fälle wie Motyls deswegen brisant, weil sie nicht allein dastehen. In der Wissenschaft ist vor einiger Zeit eine Diskussion um die Reproduzierbarkeit von Studienergebnissen entbrannt. Der Stanford-Epidemiologe John Ioannidis brachte im Jahr 2005 den Stein ins Rollen: Die Mehrzahl der veröffentlichten Ergebnisse sei falsch, behauptete er [2].

Seitdem entdecken Wissenschaftler ein schwer wiegendes Problem nach dem anderen. Ist die Art und Weise, wie Studienergebnisse bewertet werden, überhaupt brauchbar? Und was wären die Alternativen? Es gilt, Verfahren zu finden, mit denen sich möglichst alle Fehlalarme aussortieren lassen, ohne dabei ein richtiges Ergebnis zu übersehen.

> "Dass so viele Studienergebnisse schlichtweg falsch sind, hat als Weckruf gewirkt"
>
> (Steven Goodman)

"Ändere deine statistische Philosophie, und plötzlich werden ganz andere Dinge wichtig", sagt Steven Goodman, Mediziner und Statistiker, ebenfalls an der Stanford University. "Dann sieht man, dass solche Regeln nicht einfach vom Himmel gefallen sind, sondern dass sie von den Methoden abhängen, die wir uns selbst gewählt haben."

Signifikanz im ganz altmodischen Sinn

Kritik an den allgegenwärtigen p-Werten ist nichts Neues. In den neun Jahrzehnten ihrer Existenz hat man sie wahlweise als "Schmeißfliegen" bezeichnet – also nervend und nicht kaputt zu kriegen –, als "des Kaisers neue Kleider" – das heißt, mit offensichtlichen Problemen behaftet, die jeder ignoriert – oder als die "sterile intellektuelle Harke", die die Wissenschaft zu einer fruchtlosen Umtriebigkeit verführt [3]. Einmal wurde vorgeschlagen, das Verfahren in "Statistischer Hypothesen-Interferenz-Test" umzubennen [3], vermutlich weil man damit ein so passendes Akronym bilden kann.

Das Bemerkenswerte ist, dass Ronald Fisher, als er das Verfahren in den 1920er Jahren entwickelte, niemals einen alles entscheidenden Test im Sinn hatte. Der englischen Statistiker wollte Forschern lediglich eine Möglichkeit an die Hand geben,

informell die Signifikanz ihrer Daten einzuschätzen – und zwar "Signifikanz" in einem ganz altmodischen Sinn: bedeutsam genug für einen zweiten, genaueren Blick.

Seine Idee war folgende: Man mache ein Experiment und überprüfe dann, ob die Daten nicht auch mit dem vereinbar sind, was der reine Zufall liefern würde.

Dazu sollte ein Forscher zunächst einen Dummy aufbauen, der das Gegenteil dessen ausdrückt, was es zu belegen gilt – etwa dass es keinen statistischen Zusammenhang zwischen zwei Beobachtungen gibt oder dass sich zwei Gruppen nicht unterscheiden. Das ist die so genannte Nullhypothese.

Als Nächstes würde er oder sie die Rolle des Advocatus Diaboli spielen und – unter der Annahme, dass die Nullhypothese zutrifft – berechnen, mit welcher Wahrscheinlichkeit man Ergebnisse erhält, die mindestens genauso extrem ausfallen wie die, die man tatsächlich gemessen hat. Diese Wahrscheinlichkeit beschreibt der p-Wert.

Je kleiner er ist, desto unwahrscheinlicher ist es, dass die Nullhypothese zutrifft. Im Umkehrschluss, überlegte Fisher, bedeutet dies, dass an der Ausgangshypothese etwas dran sein müsste.

Baustein für einen formlosen Prozess

Damit mag das Verfahren zwar sehr präzise wirken. Doch für Fisher war es bloß ein einzelner Baustein in einem formlosen, nicht mathematischen Prozess, mit dem Wissenschaftler zu einer wissenschaftlichen Erkenntnis gelangen. Neben neuen Beobachtungen würden beispielsweise auch das Hintergrundwissen des Forschers und andere Erwägungen mit einfließen.

Doch unversehens wurde seine Idee von einer Bewegung erfasst, deren Anhänger einen ganz anderen Ansatz verfolgten: Sie suchten nach Möglichkeiten, evidenzbasierte Entscheidungen auf eine möglichst solide und möglichst objektive Grundlage zu stellen.

Vorkämpfer dieser Bewegung, die sich Ende der 1920er Jahre formierte, waren Fishers ärgste Rivalen: der polnische Mathematiker Jerzy Neyman und der britische Statistiker Egon Pearson, die eine alternative Methode zur Datenanalyse vertraten. Das System umfasste beispielsweise Konzepte wie die

"Teststärke" eines Experiments oder auch den Begriff der falsch positiven und falsch negativen Ergebnisse – inzwischen allesamt Bestandteile jedes Einführungsseminars in die Statistik. Den p-Wert allerdings ließen Neyman und Pearson nicht ganz zufällig links liegen.

Über der Fehde der Rivalen verloren Wissenschaftler die Geduld

Denn "schlimmer als nutzlos" seien Teile von Fishers Arbeit, schrieb Neyman. Der hielt dagegen: Neyman verfolge einen "kindischen" Ansatz, der "ein Schrecken für die intellektuelle Freiheit des Westens" sei.

Während sich die Kontrahenten dergestalt befehdeten, verloren andere Forscher die Geduld und verfassten einfach selbst Statistikhandbücher für ihre Kollegen. Leider waren viele dieser Autoren in der Statistik nicht bewandert genug, um die philosophischen Feinheiten beider Ansätze zu durchdringen. Also schufen sie einen Hybriden, in den sie sowohl Fishers einfach zu berechnenden p-Wert packten als auch Neymans und Pearssons wohltuend rigoroses, regelbasiertes System. Und hoben bei dieser Gelegenheit gleichzeitig den p-Wert von 0,05 als entscheidende Signifikanzschwelle auf das Podest, auf dem er bis heute ruht. "So sollte der p-Wert nie genutzt werden", sagt Goodman.

Im Endergebnis führte dies beispielsweise zu der heutigen Verwirrung darüber, was der p-Wert eigentlich ausdrückt [4]. Motyls Studie über politische Extremisten liefert passendes Anschauungsmaterial. Die meisten Wissenschaftler würden seinen ursprünglichen p-Wert von 0,01 so interpretieren, dass die Wahrscheinlichkeit eines Fehlalarms bei einem Prozent liegt.

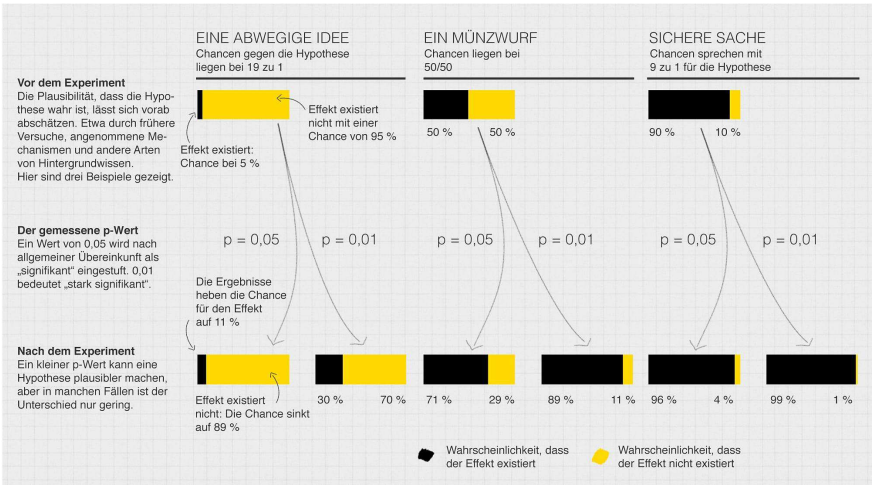
Das ist jedoch falsch. Denn eine solche Aussage kann der p-Wert überhaupt nicht treffen, da er die Daten lediglich unter Berücksichtigung einer spezifischen Nullhypothese zusammenfasst. Es ist unmöglich, mit ihm rückwärts zu rechnen und Aussagen über die zu Grunde liegende Realität zu treffen.

Eine entscheidende Information fehlt

Dazu fehlt nämlich eine entscheidende Information: die Wahrscheinlichkeit, mit der der gesuchte Effekt selbst auftritt. Wer das außer Acht lässt, macht den gleichen Fehler, wie jemand, der morgens mit Kopfschmerzen aufwacht und vermutet, dass er einen

seltenen Hirntumor hat. Das ist zwar durchaus möglich, aber so unwahrscheinlich, dass deutlich mehr Belege nötig sind, um Alltagserklärungen wie beispielsweise eine allergische Reaktion auszuschließen.

Je unplausibler die Hypothese – denken Sie an Telepathie, Aliens, Homöopathie –, desto höher ist die Wahrscheinlichkeit, dass sich ein aufregender Befund als Fehlalarm entpuppt. Und zwar völlig unabhängig vom p-Wert.



© REGINA NUZZO, NACH: SELLKE, T. ET AL., AM. STAT. 55, S. 62–71, 2001; DT. BEARBEITUNG: SPEKTRUM.DE (AUSSCHNITT)

Wie man den p-Wert deutet |
Der p-Wert gibt an, ob ein gemessenes Resultat auch durch Zufall erklärt werden kann. Die eigentlich interessante Frage beantwortet er dagegen nicht: Was sagt das über die Korrektheit der Hypothese aus?

Diese hängt nicht nur von den Ergebnissen ab, sondern auch davon, wie plausibel die Hypothese vor dem Experiment war.

Das sind schwer zu fassende Unterscheidungen. Deshalb haben Statistiker eine Hand voll Regeln aufgestellt, mit denen sich die Wahrscheinlichkeit von falsch positiven oder negativen Ergebnissen abschätzen lässt. Einer verbreiteten Rechnung [5] zufolge entspricht ein p-Wert von 0,01 einer Fehlalarmwahrscheinlichkeit von 11 Prozent. Dieser Wert wiederum hängt stark von der zu Grunde liegenden Wahrscheinlichkeit ab, dass es den zu belegenden Effekt wirklich gibt.

Liegt in einem solchen Fall der p-Wert bei 0,05, steigt die Wahrscheinlichkeit, einem Fehlalarm aufzusitzen, auf mindestens 29 Prozent.

Was bedeutet der p-Wert für die Aussagekraft?

Übertragen auf Motyls Experiment – und seinen p-Wert von 0,01 – heißt das: In einem von zehn Fällen dürfte sein Experiment ihm einen Effekt vorgaukeln, der gar nicht existiert. Die Wahrscheinlichkeit, dass Kollegen seine Studie replizieren können, liegt überdies auch nicht bei 99 Prozent, wie oft gehört, sondern eher bei 73 Prozent.

Fordert man vor der Replikation ein ebenfalls hoch signifikantes Resultat, schrumpft sie gar auf 50 Prozent [6, 7]. Anders gesagt: Sein Unvermögen, die Studie zu replizieren, ist nicht weniger überraschend, als wenn er bei einem Münzwurf auf "Kopf" getippt hätte und "Zahl" gekommen wäre.

Viele Kritiker stoßen sich auch am "unsauberen Denken", das der p-Wert ihrer Meinung nach fördert. Bestes Beispiel ist die Tatsache, dass er die Aufmerksamkeit von der Effektgröße weglenkt.

> "Was wir uns fragen sollten, ist: 'Wie groß ist der Effekt, mit dem wir es zu tun haben?' und nicht: 'Gibt es überhaupt einen Effekt?'"
>
> (Geoff Cumming)

Letztes Jahr ergab eine Studie an 19 000 Personen [8], dass sich Eheleute, die sich online kennen gelernt haben, mit geringerer Wahrscheinlichkeit wieder scheiden lassen ($p < 0{,}002$). Außerdem war die Chance, dass sie mit ihrer Ehe zufrieden waren, höher als bei jenen, die sich offline kennen lernten ($p < 0{,}001$).

p-Werte sagen nichts über die Relevanz des Ergebnisses

Das klingt nach einem eindrucksvollen Ergebnis. Doch die Auswirkungen des Onlinedatings waren bei näherem Hinsehen minimal: Dating über Partnerbörsen drückte die Scheidungsrate von 7,67 Prozent auf 5,96 Prozent, die Zufriedenheit – gemessen auf einer Sieben-Punkte-Skala – stieg von 5,48 auf 5,64.

Die Signifikanz eines Ergebnisses sage eben nichts über die praktische Bedeutung aus, erklärt Geoff Cumming, emeritierter Psychologe von der La Trobe University in Melbourne. Er spricht

von der "Verführungskraft" der p-Werte: Sie täuschen Sicherheit vor, doch die eigentlich interessanten Fragen werden außer Acht gelassen. "Was wir uns fragen sollten, ist: 'Wie groß ist der Effekt, mit dem wir es zu tun haben?' und nicht: 'Gibt es überhaupt einen Effekt?'"

Der vermutlich schlimmste Fehler von allen steckt jedoch hinter dem, was der Psychologe Uri Simonsohn von der University of Pennsylvania unter dem Begriff "P-Hacking" bekannt gemacht hat. Eine gezielte Datenmassage, die noch unter zahllosen anderen Bezeichnungen läuft wie beispielsweise "data dredging", also "Datenbaggern", oder "significance chasing", das Jagen nach Signifikanz. "P-Hacking bedeutet, verschiedene Dinge auszuprobieren, bis man am Ende das gewünschte Ergebnis erhält", sagt Simonsohn. Oft geschehe das sogar unbewusst.

"P-Hacking" hilft bei der Datenmassage

Inzwischen kann man den Ausdruck sogar, als vermutlich ersten Begriff aus der Statistik, im Onlinewörterbuch "Urban Dictionary" finden. Die Verwendungsbeispiele sagen fast schon alles: "Dieses Ergebnis scheint durch P-Hacking zu Stande gekommen zu sein", heißt es da, "die Autoren ließen eine Bedingung weg, damit der Gesamt-p-Wert auf unter 0,05 fiel". Oder: "Sie ist ein P-Hacker, sie schaut sich die Daten immer schon an, wenn das Experiment noch läuft."

Solche Methoden führen dazu, dass die Ergebnisse einer explorativen Studie – bei der Forscher ihr Gebiet erst einmal auf einen interessanten Befund hin abklopfen – im Nachhinein zur Bestätigung einer Hypothese umgedeutet werden. Dafür hätte es aber eigentlich eines viel rigoroseren Aufbaus bedurft.

Die Erkundungsstudien lassen den Experimentatoren hingegen deutlich mehr Freiheit, weshalb ihre Ergebnisse grundsätzlich immer mit Vorsicht genossen werden. Dass jemand in einer Veröffentlichung eine solche Studie kurzerhand zu einem Hypothesentest umdeklarierte, fällt meistens erst dann auf, wenn ein anderer das Experiment wiederholt.

Mit Hilfe von Simulationen hat Simonsohn gezeigt [9], dass es bereits genügt, sein Vorgehen bei der Datenanalyse an einigen wenigen Stellen zu ändern. Allein das kann die Falschpositivrate der Studie auf 60 Prozent hochtreiben. In einem solchen Fall lässt sich praktisch alles Beliebige "nachweisen".

Der Grund muss nicht zwangsläufig böser Wille sein. P-Hacking komme heutzutage gehäuft vor, sagt Simonsohn, weil oft nach sehr geringen Effekten in verrauschten Daten gesucht werde. Wie weit verbreitet das Problem ist, sei schwer zu beziffern, die Lage aber vermutlich ernst, sagt er.

Bei einer Untersuchung psychologischer Studien [10] entdeckte er beispielsweise Hinweise darauf, dass die publizierten p-Werte merkwürdig oft in der Nähe von 0,05 lagen – nichts anderes würde man erwarten, wenn Forscher so lange nach signifikanten p-Werte fischen, bis ihnen ein brauchbarer ins Netz geht.

Wenig ändert sich – trotz Kritik

Trotz aller Kritik hat sich bisher nur wenig getan. "Das Prozedere bei der statistischen Auswertung hat sich seit den Zeiten von Fisher, Neyman und Pearson kaum geändert", sagt Goodman.

John Campbell, der heute als Psychologe an der University of Minnesota in Minneapolis forscht, beklagte die Zustände schon 1982, als er Editor des "Journals of Applied Psychology" war: "Es ist praktisch unmöglich, die Autoren von ihren p-Werten abzubringen. Und je mehr Nullen nach dem Komma stehen, umso unerbittlicher hängen sie daran." [11]

Auch Kenneth Rothman von der Boston University in Massachusetts setzte alles daran, den Gebrauch von p-Werten durch Autoren einzudämmen, als er im Jahr 1989 das Journal "Epidemiology" auflegte. Ohne nachhaltigen Erfolg: Mit seinem Ausscheiden im Jahr 2001 begann sofort die Rückkehr der p-Werte.

Ioannidis durchforstet derzeit die Datenbank PubMed auf der Suche danach, wie Forscher unterschiedlicher Disziplinen mit dem p-Wert und anderen Statistiken umgehen. "Schon ein oberflächlicher Blick auf eine Auswahl aktueller Veröffentlichungen zeigt, dass p-Werte nach wie vor äußerst beliebt sind."

Eine Reform müsste es schaffen, eingefahrene Sitten und Gebräuche auszumerzen: Wie Statistik an den Unis gelehrt wird, wie Studienergebnisse ausgewertet und interpretiert werden und wie sie in den Fachjournals weitergegeben werden.

Immerhin würden Wissenschaftler inzwischen einräumen, dass es ein Problem gibt, sagt Goodman. "Dass so viele Studienergebnisse schlichtweg falsch sind, hat als Weckruf gewirkt." Forschern wie

John Ioannidis sei es zu verdanken, dass die Sorgen der Statistiker nicht mehr länger als reine Theorie wahrgenommen werden. "Die Probleme, die die Statistiker vorhergesagt haben, sind genau die, die wir jetzt beobachten", sagt Goodman, "nur haben wir noch keine brauchbare Lösung parat."

Welche Maßnahmen helfen?

Laut den Experten gäbe es Maßnahmen, die helfen könnten. Cumming etwa schlägt vor, dass Forscher immer auch die Effektgröße und das Konfidenzintervall veröffentlichen. Diese Werte drückten das aus, was der reine Signifikanzwert nicht könne: das Ausmaß und die relative Bedeutung des Effekts.

Autoren sollten ihre Arbeit mit einem Siegel "garantiert ohne P-Hacking" versehen.

Einige seiner Kollegen halten es auch für viel versprechend, den p-Wert durch Methoden abzulösen, die auf den Satz von Bayes bauen. Dabei werden Wahrscheinlichkeiten als Plausibilität des Eintretens eines Ergebnisses beschrieben und nicht durch dessen potenzielle Häufigkeit.

Dadurch hält zwar eine gewisse Subjektivität Einzug in die Statistik – jenes Moment also, das die Pioniere vom Anfang des 20. Jahrhunderts um jeden Preis heraushalten wollten. Doch die bayessche Statistik macht es vergleichsweise einfach, Hintergrundwissen über die Welt in Schlussfolgerungen einfließen zu lassen und zu berechnen, wie sich Wahrscheinlichkeiten verändern, wenn neue Belege und Hinweise hinzukommen.

Andere vertreten einen eher pluralistischen Ansatz: Wissenschaftler sollten verschiedenartige Methoden auf ein und denselben Datensatz anwenden. Stephen Senn, ein Statistiker vom Centre for Public Health Research in Luxembourg City, vergleicht diesen Ansatz mit dem eines Staubsaugerroboters, der nicht allein aus einer Zimmerecke herausfindet. Jeder Weg, die Daten zu analysieren, endet unweigerlich an einer Wand, und es braucht ein gewisses Maß an Alltagsverstand, um wieder in die Gänge zu kommen.

Wenn die verschiedenen Verfahren alle unterschiedliche Antworten liefern, sagt er, "ist das ein Hinweis darauf, dass wir kreativer werden müssen und herausfinden, woran das liegt." Das könne nur zu einem besseren Verständnis der zu Grunde liegenden Realität führen.

Eine Antwort: Alles offenlegen

Für Simonsohn liegt die beste Absicherung für einen Forscher darin, alles offenzulegen. Autoren sollten ihre Arbeit mit einem Siegel "garantiert ohne P-Hacking" versehen, indem sie folgenden Satz hinzufügen: "Wir berichten im Folgenden, wie wir unsere Stichprobengröße ausgewählt haben, welche Daten wir ausgelassen haben (wenn das der Fall war) sowie welche Manipulationen und Messgrößen zur Anwendung kamen."

Damit, hofft jedenfalls Simonsohn, würden sie dem P-Hacking einen Riegel vorschieben – oder zumindest den Leser auf Schwindeleien aufmerksam machen. Dann könne sich jeder eine entsprechende Meinung bilden.

Alternative: Ein zweistufiger Prozess

Eine verwandte Idee, die in letzter Zeit Zulauf bekommt, ist die zweistufige Analyse, auch *preregistered replication* ("vorab registrierte Replikation") genannt, erklärt der Politikwissenschaftler und Statistiker Andrew Gelman von der New Yorker Columbia University. Bei diesem Ansatz werden explorative und hypothesentestende Studien klar als solche gekennzeichnet. Anstatt beispielsweise vier kleine Experimente auszuführen und die Ergebnisse zu einem Paper zu vereinen, würden Forscher erst mit zwei kleinen explorativen Studien das Gebiet auf interessante Beobachtungen durchforsten – ohne sich dabei allzu viel Gedanken um mögliche Fehlalarme zu machen.

Erst dann würde das Team auf Basis der Ergebnisse eine Studie entwerfen, von der sie sich die Bestätigung ihrer neu gewonnenen Erkenntnisse erhoffen, und ihre Absicht in einer Datenbank wie dem Open Science Framework kundtun. Sobald sie die entsprechende Studie durchgeführt haben, können sie sie zusammen mit den Erkundungsstudien in einem Paper der Öffentlichkeit vorstellen.

Ein solches Vorgehen lasse sowohl Raum für viele Freiheiten, findet Gelman, sei aber trotzdem rigoros genug, um die Zahl falscher Befunde zu drücken.

Wenn die Zahlen auf dem Tisch liegen

Grundsätzlich sei es jedoch an der Zeit, sagt Goodman, dass sich Wissenschaftler die Grenzen der konventionellen Statistik bewusst machten. Vor allem sollte die eigene Analyse der Ergebnisse um eine seriöse wissenschaftliche Einschätzung ihrer Plausibilität ergänzt werden – alles das, was üblicherweise erst später unter dem Abschnitt "Diskussion" abgehandelt werde: Welche Ergebnisse erbrachten ähnliche Untersuchungen? Gibt es einen Mechanismus, der die Beobachtung erklären könnte? Deckt sich der Befund mit klinischen Erfahrungen? Und dergleichen mehr.

Der Statistiker Richard Royall von der Johns Hopkins Bloomberg School of Public Health in Maryland drückt es so aus: Es gebe drei Dinge, die sich jeder Wissenschaftler nach einem Experiment fragen sollte – "Welche Belege habe ich?", "Was soll ich selbst glauben?" und "Was soll ich tun?". Kein Verfahren könne auf alle drei Fragen gleichzeitig antworten, sagt Goodman: "Wenn die Zahlen auf dem Tisch liegen, sollte die wissenschaftliche Diskussion losgehen und nicht aufhören."

Dieser Artikel erschien unter dem Titel "Statistical errors" in Nature 506, S. 150-152, 2014.

Regina Nuzzo
Regina Nuzzo ist freiberufliche Journalistin und Professorin für Statistik an der Gallaudet University in Washington, D.C.

QUELLEN