Forschung & Lehre | Ausgabe 6/20 | S. 510-513

Paradigmenwechsel

Warum statistische Signifikanztests abgeschafft werden sollten

| NORBERT HIRSCHAUER | CLAUDIA BECKER | **Seit Mitte des 20. Jahrhunderts hat sich das Konzept der statistischen Signifikanz als quasi universeller Standard zur Beurteilung der Glaubwürdigkeit von Studienergebnissen etabliert. Fast ebenso alt ist die Kritik an fehlerhaften Praktiken und Schlussfolgerungen, die mit statistischen Signifikanzaussagen eng verbunden sind. Manche Kritiker sprechen von einem „Kult" des Signifikanztestens, den sie als maßgebliche Ursache der Replikations- und Vertrauenskrise der Wissenschaft ansehen.**

Empirische Studien müssen ihren Lesern drei zentrale Informationen vermitteln: Erstens muss kommuniziert werden, welche Daten mit welchen Methoden analysiert wurden. Zweitens muss der in den Daten gefundene Sachverhalt (die *empirische Evidenz*) beschrieben werden. Drittens ist die zentrale Frage nach der *Validität* zu beantworten: Welche Schlussfolgerungen können vernünftigerweise aus den Studienergebnissen gezogen werden und wie hoch ist die verbleibende Unsicherheit? Dies wird in der Wissenschaftstheorie als *Induktion* oder *Inferenz* bezeichnet. Ein adäquater Induktionsschluss muss alles berücksichtigen, was an Wissen vorliegt. Er darf nicht mit der Beschreibung der Ergebnisse einer einzelnen Studie gleichgesetzt werden. Induktion ist vielmehr ein nachgelagerter Schritt, der danach fragt, was man aus dem *Besonderen* (den konkret analysierten Daten) auf das *Allgemeine* (den interessierenden realen Sachverhalt) ableiten kann.

Für den Induktionsschluss – z.B. bei der Frage, ob ein bestimmter Stoff krebserregend ist oder nicht – gibt es keinen statistischen Automatismus, der wissenschaftliches Argumentieren ersetzen könnte. Ein datenbasierter Automatismus wäre aber so „praktisch", dass Studienergebnisse oft so interpretiert werden, als ob es ihn gäbe. Die daraus resultierenden Schnellschüsse tragen zur aktuell beklagten *Replikationskrise* bei; d.h. die Aussagen vieler Studien lassen sich in Folgestudien nicht bestätigen. Neben Wunschdenken hat die statistische Terminologie maßgeblich zu diesem Missstand beigetragen, da sie mit Begriffen wie Signifikanz, Irrtumswahrscheinlichkeit und Hypothesentest fast zwangsläufig sprachliche Fehlassoziationen hervorruft.

Seit Jahrzehnten sind statistische Signifikanztests sowie der „Nachweis" von Ergebnissen mit Neuigkeitswert durch Signifikanzsternchen quasi die Voraussetzung für wissenschaftliches Publizieren. Die meisten empirischen Studien (z.B. Regressionsanalysen) berechnen deshalb nach der datenbasierten Schätzung von Zusammenhängen routinemäßig p-Werte und sprechen bei $p \leq 0{,}05$ von „*statistischer Signifikanz*". Häufig wird der p-Wert auch als „*Irrtumswahrscheinlichkeit*" bezeichnet. Beide Begriffe sind hochproblematisch, da sie nicht nur bei Laien, sondern nachweislich auch bei vielen Forschern zu Fehlinterpretationen führen:

1. Entgegen jedem umgangssprachlichen Verständnis ist es falsch, „statistisch signifikant" mit „groß" oder „wichtig" gleichzusetzen. Bei hinreichend großen Datensätzen wird jeder noch so kleine Effekt statistisch signifikant.
2. Ein in den Daten identifizierter, aber „nicht signifikanter" Effekt kann nicht als Indiz dafür gewertet werden, dass kein (bedeutender) Effekt vorliegt. In kleinen Stichproben findet man oft „nicht signifikante" Ergebnisse. Diese sind zwar wenig belastbar, aber es wäre unsinnig, einen positiven Befund als Indiz für die Nicht-Existenz eines Effekts zu interpretieren. Ein Abzählen „signifikanter" versus „nicht signifikanter" Studien ist deshalb ungeeignet, um sich einen Überblick über den Erkenntnisstand in einem Gebiet zu verschaffen.
3. Trotz der Bezeichnung „Irrtumswahrscheinlichkeit" ist der p-Wert nicht die Wahrscheinlichkeit, einen Irrtum zu begehen, wenn man im Lichte der Studienergebnisse folgert, dass ein Effekt da ist (siehe Erklärung unten).

Das Problem der traditionellen Veröffentlichungspraxis

Oft wird der p-Wert nicht als das dargestellt, was er ist, nämlich eine statisti-

AUTOREN

Norbert Hirschauer ist Professor für Unternehmensführung im Agribusiness an der Universität Halle-Wittenberg. Zu seinen Arbeitsschwerpunkten gehören u.a. Risikomanagement und fehlerhafte Inferenzschlüsse.

Claudia Becker ist Professorin für Statistik an der Universität Halle-Wittenberg. Zu ihren Arbeitsschwerpunkten gehören die statistische Methodenentwicklung sowie die Konzeption von Befragungen.

sche Größe, die bei Vorliegen einer Zufallsauswahl eine kleine Hilfestellung für den Induktionsschluss geben kann. Insbesondere in Verbindung mit vordergründig selbsterklärenden Signifikanzaussagen wird vielmehr das Ross zum Reiter gemacht; d.h. ein spezifisches statistisches Hilfsmittel wird mit dem Inferenzschluss als solchem gleichgesetzt. Das ist bequem, aber falsch und führt dazu, dass die willkürliche *Dichotomisierung* in „statistisch signifikante“ und „nicht statistisch signifikante“ Ergebnisse als Nachweis für die Existenz bzw. Nichtexistenz eines Effektes interpretiert wird. Das Problem kommt in sich scheinbar selbstgenügenden Statements zum Ausdruck, dass statistische Tests eben gezeigt hätten, dass ein Zusammenhang signifikant sei oder nicht.

Die verbreitete Praxis, adäquate kontextbezogene Induktionsschlüsse durch einen statistischen Ja/Nein-Automatismus zu ersetzen, haben die beiden amerikanischen Wissenschaftler Ziliak und McCloskey schon im Jahr 2008 in ihrem gleichnamigen Buch als „*Cult of Statistical Significance*“ kritisiert. Bereits zehn Jahre davor hatten sie nachgewiesen, dass die Mehrzahl aller Beiträge mit Signifikanztests im renommierten *American Economic Review* der Jahre 1980 bis 1989 schwere Fehlinterpretationen wie z.B. die Gleichsetzung von „signifikant“ und „wichtig“ aufwies. Auch in vielen Statistiklehrbüchern fanden McCloskey und Ziliak schwere Fehler.

Seitdem hat sich die Kritik weiter intensiviert, in der wissenschaftlichen Praxis hat sich aber wenig getan. Vor diesem Hintergrund trat die *American Statistical Association* (ASA) im Jahr 2016 in einem noch nie dagewesenen Schritt mit einer *p*-Wert bezogenen Methodenwarnung an die Öffentlichkeit. Darin wies sie explizit darauf hin, dass der *p*-Wert nicht geeignet ist, um zu bestimmen, ob eine Hypothese richtig oder ein Ergebnis wichtig ist. Drei Jahre darauf erschien im *The American Statistician* eine Sonderausgabe, in der bereits im Editorial empfohlen wurde, Signifikanztests aufzugeben („*Don't say statistically significant*“). Nahezu zeitgleich wurde Anfang 2019 in *Nature* ein Aufruf zur Abschaffung von Signifikanztests veröffentlicht („*Retire statistical significance*“), der von über 800 Wissenschaftlern unterzeichnet wurde.

Inzwischen berichtet auch die allgemeine Presse (DIE ZEIT, Handelsblatt, Süddeutsche Zeitung) über die statistische Methodendebatte. Im Gegensatz zu den *National Academies of Sciences* der USA, die Ende 2019 vor dem Hintergrund der Signifikanzdebatte einen Konsensbericht veröffentlichten, haben sich die Wissenschaftsinstitutionen im deutschsprachigen Raum des Themas bisher kaum angenommen. Auch in den Wirtschaftswissenschaften scheinen die Rezeption der Debatte und institutionelle Reformbemühungen schwach ausgeprägt zu sein. Das ist bedauernswert – und unverständlich. Immerhin haben einige der renommiertesten Ökonomiejournale erste Reformen umgesetzt. So untersagen bspw. das *American Economic Review*, *Econometrica* und die vier *American Economic Journals* die Verwendung von Hervorhebungen wie Sternchen, die dichotome Fehlinterpretationen hervorrufen könnten. Außerdem ist es dort Standard, in den Ergebnistabellen Standardfehler statt *p*-Werten auszuweisen.

Was sagen *p*-Werte aus?

Sieht man von (hier nicht behandelten) randomisierten Experimenten ab, geht es bei *p*-Werten um die Ungenauigkeit stichprobenbasierter Größen. Ausgehend von einer *Zufallsstichprobe* will man zu einer definierten *Grundgesamtheit* hin generalisieren. Stellen wir uns vor, man wolle die Einkommensunterschiede zwischen Männern und Frauen in der Grundgesamtheit der erwerbstätigen Berliner Bevölkerung anhand einer Zufallsstichprobe von 50 Personen untersuchen. In der Stichprobe sei das durchschnittliche Einkommen der Männer mit 22 Euro/Stunde um 4 Euro (22 Prozent) höher als das der Frauen. Natürlich fände man die Einkommensdifferenz vertrauenswürdiger, wenn wir statt 50 Personen eine Zufallsauswahl von 50 000 Personen getroffen hätten. Und am vertrauenswürdigsten wäre das Ergebnis, wenn wir die Einkommen aller Berliner Erwerbstätigen erfasst hätten. In diesem Fall wäre statistische Inferenz weder nötig noch möglich. Die Einkommensunterschiede in der interessierenden Grundgesamtheit wären bekannt.

Was aber ist, wenn mehrere kleine Studien, jeweils mit einer Zufallsstichprobe von 50 Personen, gemacht werden? Da jede Stichprobe zufallsbedingt etwas anders ausfällt, würden wir in jeder Studie eine andere Differenz und damit einen anderen Schätzwert für die Einkommensdifferenz in Berlin finden. Die Verteilung der sich über viele Studien hinweg ergebenden Schätzwerte bezeichnet man als *Stichprobenverteilung*. Der sog. Stichproben- oder *Standardfehler* bezeichnet die Streuung der Stichprobenverteilung. Der *p*-Wert basiert auf dem Stichprobenfehler und damit dem Gedankenexperiment der wiederholten Ziehung einer Zufallsstichprobe gleicher Größe („statistische Replikation“). Bezogen auf unser Beispiel gäbe er an, wie wahrscheinlich es ist, die in der Stichprobe gefundene Differenz von 4 Euro oder mehr in sehr häufig wiederholten Zufallsziehungen von 50 Personen zu finden, *wenn* man unterstellen würde, dass es in der Grundgesamtheit keinen Einkommensunterschied gäbe.

Worin liegen die Missverständnisse bei Signifikanztests?

Trotz des irreführenden Begriffs sind Signifikanztests kein geeignetes Hilfsmittel, um zu entscheiden, ob ein relevanter Effekt vorhanden ist oder nicht. Im Gegenteil: Wiederholte Zufallsstichproben liefern zwangsläufig unterschiedliche Schätzwerte und das Unterschreiten der *p*-Wert Grenze von 0,05 zeigt nicht an, dass ein Effekt wahr ist. Bei adäquater (erwartungstreuer) Schätzung nähern vielmehr die über viele Zufallsstichproben hinweg gewonnenen Schätzwerte zusammengenommen den wahren Sachverhalt an. Für den Erkenntnisgewinn (z.B. bzgl. des Einkommensunterschieds zwischen Männern und Frauen) werden also die Schätzungen *aller* ordentlich gemachten Studien benötigt, unabhängig von der Ausprägung des jeweiligen *p*-Werts.

Die dem Testen inhärente Dichotomisierung führt dagegen nicht nur zu einer Vernachlässigung „nichtsignifikanter“ Schätzwerte (z.B. durch Nichtpublikation), sondern auch zu einer Überbewertung „hochsignifikanter“ Schätzwerte. In unserem Beispiel würde man mit den Schätzwerten aus den Stichproben mit den sehr kleinen *p*-Werten den Einkommensunterschied zwangsläufig überschätzen – ganz entgegen einer landläufigen Assoziation, die „hoch signifikant“ mit „glaubwürdig“ gleichsetzt. Mit den Schätzwerten aus den Stichproben mit den sehr hohen *p*-Werten würde man den Einkommensunterschied dagegen unterschätzen. Ein zutreffendes Bild bekommen wir nur, wenn wir alle Studien berücksichtigen. Nimmt man die Dichotomie des Signifikanztestens ernst, kommt es also zu

einer Fehleinschätzung. Und wenn man sie nicht ernst nimmt und jede ordentlich gemachte Studie als Wissensbeitrag ansieht, braucht man Signifikanztests nicht.

Welchen Nutzen haben p-Werte?

Wie bereits von der ASA betont wurde, stellen *p*-Werte nicht die Wahrscheinlichkeit von Hypothesen dar. Allerdings kennzeichnen abnehmende *p*-Werte eine zunehmende Unvereinbarkeit der Daten einer Zufallsstichprobe mit der üblicherweise als „keine Differenz" oder „kein Zusammenhang" formulierten Nullhypothese. Ohne Verbindung mit einem willkürlichen Grenzwert sind *p*-Werte deshalb nicht per se zu verwerfen. Sie liefern aber im Vergleich zu anderen inferenzstatistischen Größen weniger Information bzgl. der Frage, was man aus dem Befund *einer* konkreten Zufallsstichprobe (*signal*) im Lichte des nicht zu vermeidenden Stichprobenfehlers (*noise*) schließen kann. Hilfreicher sind Größen wie z.B. *t*- oder *z*-Werte, die direkt das Verhältnis zwischen der Stärke des geschätzten Sachverhalts und dem Stichprobenfehler (*signal-to-noise ratio*) abbilden.

Wenn auf den Stichprobenfehler zurückgegriffen wird, muss klar kommuniziert werden, dass dieser sich lediglich auf die Unsicherheit bezieht, die durch die Ziehung einer Zufallsstichprobe entsteht. Die Voraussetzungen für seine Anwendung sind nicht gegeben, wenn keine Zufallsauswahl vorliegt, sondern leichter zugängliche, aber verzerrte *Convenience Samples* (z.B. freiwillige Befragungsteilnehmer) analysiert werden. Das ist häufig der Fall. Der Stichprobenfehler hilft auch nicht bei der Einschätzung anderer, oft weit bedeutsamerer Fehlerquellen. Man denke in unserem Fall z.B. an die Frage, ob das Konstrukt „Einkommen" richtig gemessen wurde: Hat man das gemessen, was man wissen wollte oder hätte man besser das Monatserwerbseinkommen sowie zusätzlich die Kapitaleinkünfte erfassen sollen? In anderen Forschungskontexten könnte es darum gehen, inwieweit die Ergebnisse aus artifiziellen Labor- oder Befragungssituationen auf die Realität übertragbar sind. Das sind alles Fragen, die für den Induktionsschluss und damit die Validität essentiell sind, die aber nichts mit Inferenzstatistik und dem Stichprobenfehler zu tun haben. Wissenschaftliche Inferenz ist weit mehr als statistische Inferenz.

Paradigmenwechsel institutionell unterstützen

Im Lichte der Beharrungskräfte des Wissenschaftssystems und nicht zuletzt der Perpetuierung von Fehlern über Lehrbücher und die akademische Lehre sollte man nicht nur darauf hoffen, dass irreführende Signifikanzaussagen durch die wissenschaftliche Praxis nach und nach aufgegeben werden. Vielmehr sollte ein solcher „Paradigmenwechsel" auch institutionell befördert werden. Dem Aufruf in *Nature* („*Retire statistical significance*") folgend sollten wissenschaftliche Fachgesellschaften und Journale vor allem klarstellen, dass die Dichotomie statistischer Signifikanztests mit Blick auf den Induktionsschluss nicht begründbar ist.

Insbesondere die Überarbeitung von Journalrichtlinien sollte schnell erfolgen. Ansonsten besteht die Gefahr, dass man als Forscher weiterhin von Gutachtern, die im herkömmlichen Prozedere verhaftet sind, zu Signifikanzdeklarationen und dem Ausweis von Sternchen „gezwungen" wird. Ohne explizite Reform von Journalrichtlinien sind auch Diskussionen mit Nachwuchswissenschaftlern vorprogrammiert, die mit besorgtem Blick auf ihre Veröffentlichungsliste fordern, man solle es „nicht so genau nehmen", da die Journale eben immer noch Signifikanztests und scheinbar eindeutige Aussagen mit Neuigkeitswert einfordern. Käme man dem nach, würde man wider besseres Wissen weitere Beiträge publizieren, die Irrtümer perpetuieren und die Replikationskrise verstärken. Allerdings ist die Besorgnis der Doktoranden ernst zu nehmen. Sie und ihre Betreuer wären ohne institutionelle Unterstützung für den überfälligen Wandel noch lange Zeit einem enormen und unguten Spannungsfeld ausgesetzt.

Der p-Wert ist nicht die Wahrscheinlichkeit einer Hypothese

Nehmen wir an, aus einer Box mit 99 idealen (nichtmanipulierten) Münzen [*P(Kopf)* = 0,5] und einer manipulierten Münze [*P(Kopf)* = 0,75] werde zufällig eine Münze gezogen. Diese zeige bei einem fünfmal wiederholten Testwurf 5 x Kopf. *Wenn* es eine ideale Münze wäre, wäre bei sehr vielen Wiederholungen des Tests „fünfmaliger Münzwurf" nur in 3,125 Prozent (= $0{,}5^5$) der Fälle 5 x Kopf zu erwarten. Diese bedingte Wahrscheinlichkeit entspricht dem *p*-Wert. Sie ist aber nicht die Wahrscheinlichkeit, bei Verwerfung der Nullhypothese „ideale Münze" (= keine Manipulation) einen Irrtum zu begehen. Hierfür muss man zusätzlich wissen, wie hoch bei der manipulierten Münze die Wahrscheinlichkeit für 5 x Kopf ist. Sie beträgt 23,73 Prozent (= $0{,}75^5$). Man muss zudem die *vor* dem Testwurf bekannten (*A-priori*-)Wahrscheinlichkeiten von 99 Prozent und einem Prozent berücksichtigen, dass man anfangs eine ideale bzw. eine manipulierte Münze gezogen hatte. Nach dem Satz von Bayes kommt man *nach* dem Testwurf auf eine (*A-posteriori*-) Wahrscheinlichkeit von 92,88 Prozent [= 0,03125·0,99/(0,03125·0,99+0,237·0,01)], dass man eine ideale Münze hat. Trotz des *p*-Werts von 3,125 Prozent wird man also die Nullhypothese „ideale Münze" nicht verwerfen. Der Informationsgewinn durch den Wurftest führt lediglich dazu, dass man ein Update der Wahrscheinlichkeit von 99 auf 92,88 Prozent vornimmt.

Literaturhinweise:

The ASA's statement on p-values: context, process, and purpose (2016). The American Statistician 70(2): 129–133.

Statistical Inference in the 21st Century: A World Beyond p < 0.05 (2019). The American Statistician 73(sup1).

Retire statistical significance (2019). Nature 567: 305-307.

Reproducibility and Replicability in Science. Consensus Study Report (2019). National Academies of Sciences, Engineering and Medicine.