DOI: 10.1007/s10273-017-2108-3

Norbert Hirschauer, Oliver Mußhoff, Sven Grüner

False Discoveries und Fehlinterpretationen wissenschaftlicher Ergebnisse

Implikationen für die Wissenschaftskommunikation

Das Vertrauen der Bevölkerung gegenüber der Wissenschaft hängt in starkem Maß von der Art der Vermittlung wissenschaftlicher Erkenntnisse ab. Dabei ist es entscheidend, durch eine adäquate Interpretation statistischer Analysen eine vernünftige Bewertung der Zuverlässigkeit wissenschaftlicher Aussagen vorzunehmen. Oft scheitert das bereits an der Interpretation der „statistischen Signifikanz". Die Autoren des vorliegenden Beitrags legen dar, dass die Berücksichtigung des gesamten Forschungsstands in einem bestimmten Gebiet sowie sachlich angemessene und für die Adressaten verständliche statistische Interpretationen zentrale Voraussetzungen für eine gelingende Wissenschaftskommunikation sind.

Das Wort „postfaktisch" ist in aller Munde. Die Gesellschaft für deutsche Sprache hat es zum Wort des Jahres 2016 gewählt: Es kennzeichne die Entwicklung, dass Fakten in der gesellschaftlichen Auseinandersetzung an Bedeutung verlieren und zunehmend größere Bevölkerungsschichten in ihrem Widerwillen gegen „die da oben" nur noch Bestätigungen für vorgefertigte Einstellungen suchen. Diese Entwicklung untergräbt auch die Akzeptanz wissenschaftlicher Aussagen, die bestenfalls als „alternative" Sichtweisen bezeichnet und schlimmstenfalls als interessensgeleitete Manipulationsversuche der „Eliten" wahrgenommen werden. Die damit grundsätzlich infrage gestellte Rolle der Forschung bei der Bereitstellung und Überprüfung von Sachaussagen spiegelt sich in der aktuellen Diskussion zur Vertrauenskrise der Wissenschaft wider. Laut Wissenschaftsbarometer 2016[1] ist ein Drittel der deutschen Bevölkerung der Ansicht, dass die Menschen zu sehr der Wissenschaft vertrauen und zu wenig ihren Gefühlen und dem Glauben. Peter Strohschneider, der Präsident der Deutschen Forschungsgemeinschaft (DFG), bezeichnete diese Entwicklung in seiner diesjährigen Neujahrsansprache sogar als wissenschafts- und demokratiegefährdend.[2]

Die Diskussion über das fehlende Vertrauen der Bevölkerung in die Wissenschaft wird von vielen Forschern hauptsächlich als Verständigungsproblem wahrgenommen und deshalb unter dem Blickwinkel der Wissenschaftskommunikation geführt.[3] Die dabei vertretenen Erklärungsansätze sind unterschiedlich. Häufig liegt der Fokus aber auf den Herausforderungen durch neue Medien und gesellschaftliche Veränderungsprozesse.[4] Dabei hat man das veränderte Verhalten der Bürger als „Informationsnachfrager" (Stichwort: neue soziale Netzwerke), das Auftreten neuer

1 Wissenschaft im Dialog/TNS emnid: Wissenschaftsbarometer 2016, https://www.wissenschaft-im-dialog.de/projekte/wissenschaftsbarometer/wissenschaftsbarometer-2016/.

2 Rede von Peter Strohschneider, dem Präsidenten der Deutschen Forschungsgemeinschaft (DFG), anlässlich des Neujahrsempfangs der DFG in Berlin, 13.1.2015, http://www.dfg.de/download/pdf/dfg_im_profil/reden_stellungnahmen/2015/150113_rede_strohschneider_neujahrsempfang.pdf.

3 Vgl. Wissenschaft im Dialog: Dokumentation zum 9. Forum Wissenschaftskommunikation vom 5.-7.12.2016 in Bielefeld, https://www.wissenschaft-im-dialog.de/forum-wissenschaftskommunikation/eindruecke-2016/.

4 Vgl. z.B. P. Weingart, L. Guenther: Science communication and the issue of trust, in: Journal of Science Communication, 15. Jg. (2016), H. 05, C01: S. 1-14.

Prof. Dr. Norbert Hirschauer ist Professor für Unternehmensführung im Agribusiness am Institut für Agrar- und Ernährungswissenschaften der Martin-Luther-Universität Halle-Wittenberg.

Prof. Dr. Oliver Mußhoff ist Professor für Landwirtschaftliche Betriebslehre am Department für Agrarökonomie und Rurale Entwicklung der Georg-August-Universität Göttingen.

Dr. Sven Grüner ist wissenschaftlicher Mitarbeiter am Lehrstuhl für Unternehmensführung im Agribusiness der Martin-Luther-Universität Halle-Wittenberg.

„Informationsanbieter“ (Stichwort: wissenschaftsfremde Akteure) und – sozusagen als unerwünschte Technikfolge – eine veränderte Informationsdynamik (Stichwort: postfaktisches Denken) im Blick.

Für die in der Wissenschaftskommunikation Tätigen sind neue Kommunikationstechnologien außerordentlich wichtig. Schließlich müssen sie immer wieder Strategien identifizieren, mit denen Kommunikation im Lichte veränderter „Informationsmuster“ gelingt. Beispielsweise wäre ein ausschließlicher Rückgriff auf Printmedien heutzutage in den meisten Fällen nicht mehr adäquat. Ein neben neuen Informationsmustern wichtiger Grund für die Erosion des Vertrauens in die Wissenschaft wird in der bisherigen Wissenschaftskommunikationsdebatte aber kaum angesprochen: die hohe Zahl von False Discoveries und die Tatsache, dass über die Disziplinen hinweg eine Vielzahl von Ergebnissen, die als „statistisch signifikant“ ausgewiesen wurden, in Folgestudien nicht reproduziert werden können (*Reproducibility Crisis*).

Eine wichtige Ursache für die Reproduktionskrise wird im „Cult of Statistical Significance“ gesehen, d.h. der sich historisch entwickelten „Fixierung“ von Forschern, Gutachtern und Zeitschriftenherausgebern auf statistisch signifikante Ergebnisse.[5] Als Kriterium der statistischen Signifikanz wird der p-Wert genutzt, der das üblicherweise akzeptierte Signifikanzniveau von 0,05 unterschreiten soll. Das Wissen, dass sowohl Gutachter als auch Herausgeber statistisch signifikante Ergebnisse bevorzugen, verlockt einerseits Forscher dazu, „p-Hacking“[6] zu betreiben, also bewusst oder unbewusst verschiedene Datensets und Analysemethoden auszuprobieren und dann selektiv das Modell darzustellen, mit dem sich „veröffentlichungsfähige“ p-Werte erzielen lassen. Andererseits werden solche Ergebnisse dann tatsächlich bevorzugt publiziert. Im Ergebnis ergibt sich ein *Publikationsbias*, d.h. man findet in den Publikationen eine Vielzahl aufsehenerregender Positivergebnisse. Die Negativergebnisse sind aber unterrepräsentiert, da Studien, die vorhergehende Studien nicht bestätigen (würden), entweder nicht veröffentlicht oder gar nicht durchgeführt werden. Wenn dann doch eine Überprüfung stattfindet, ergibt sich das angeführte Problem der fehlenden Reproduzierbarkeit.

Die diesen Sachverhalt kritisierenden Zeitschriftenbeiträge tragen einschlägige Titel wie z.B. „Why most published research findings are false“[7], „An investigation of the false discovery rate and the misinterpretation of p-values“[8], „Statistical errors“[9] oder „Corrupt research: the case for reconceptualizing empirical management and social science“[10]. Eine Veröffentlichung in der Zeitschrift Nature, die die Methodenwarnung der *American Statistical Association*[11] zu statistischen Signifikanztests und dem p-Wert von Anfang 2016 aufgreift, bringt den Stand der Dinge wie folgt auf den Punkt: „Misuse of the P value – a common test for judging the strength of scientific evidence – is contributing to the number of research findings that cannot be reproduced, the American Statistical Association (ASA) warns in a statement released today. The group has taken the unusual step of issuing principles to guide use of the P value, which it says cannot determine whether a hypothesis is true or whether results are important.“[12]

Vor diesem Hintergrund ist die Frage zu stellen, ob es überhaupt moralisch vertretbar wäre, über „gute“ Kommunikation Vertrauen in einen nicht vertrauenswürdigen Gegenstand herzustellen. Zudem ist zu vermuten, dass dies dauerhaft nicht gelingt. Angesichts der von der Wissenschaft selbst diagnostizierten Reproduktionskrise[13] ist es nachvollziehbar, wenn die Öffentlichkeit angeblich wissenschaftlich untermauerten „statistisch signifikanten“ Ergebnissen nicht traut. Ist das Problem der Wissenschaftskommunikation also gar nicht in erster Linie ein Problem der Kommunikation, sondern ein Problem der Wissenschaft – einer Wissenschaft, die immer wieder Aufsehen erregende „statistisch signifikante Studienergebnisse“ produziert, der Überprüfung (Reproduzierbarkeit) von Einzelstudien aber

5 S. T. Ziliak, D. N. McCloskey: The Cult of Statistical Significance. How the Standard Error Costs Us Jobs, Justice, and Lives, Michigan 2008.

6 Diese Begriffsbildung wird Simmons et al. zugeschrieben, die konstatieren: „it is unacceptably easy to publish ‚statistically significant‘ evidence consistent with any hypothesis. […] In the course of collecting and analyzing data, researchers have many decisions to make: Should more data be collected? Should some observations be excluded? […] it is common (and accepted practice) for researchers to explore various analytic alternatives, to search for a combination that yields ‚statistical significance‘, and to then report only what ‚worked‘. The problem, of course, is that the likelihood of at least one (of many) analyses producing a falsely positive finding at the 5% level is necessarily greater than 5%.“ Vgl. J. P. Simmons, L. D. Nelson, U. Simonsohn: False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant, in: Psychological Science, 22. Jg. (2011), Nr. 11, S. 1359.

7 J. P. A. Ioannidis: Why most published research findings are false, in: PLoS Medicine, 2. Jg. (2005), Nr. 8, e124, S. 0696-0701.

8 D. Colquhoun: An investigation of the false discovery rate and the misinterpretation of p-values, in: Royal Society Open Science, 1. Jg. (2014), Nr. 3, S. 140-216.

9 R. Nuzzo: Statistical errors. P-values, the ‘gold standard’ of statistical validity, are not as reliable as many scientists assume, in: Nature, 506. Jg. (2014), Nr. 7487, S. 150-152.

10 R. Hubbard: Corrupt research: the case for reconceptualizing empirical management and social science, Los Angeles 2016.

11 R. L. Wasserstein, N. A. Lazar: The ASA’s statement on p-values: context, process, and purpose, in: The American Statistician, 70. Jg. (2016), Nr. 2, S. 129-133.

12 M. Baker: Statisticians issue warning on P values, in: Nature, 531. Jg. (2016), Nr. 7593, S. 151. „This is the first time that the 177-year-old ASA has made explicit recommendations on such a foundational matter in statistics, says executive director Ron Wasserstein. The society’s members had become increasingly concerned that the P value was being misapplied in ways that cast doubt on statistics generally, he adds.“

13 Vgl. z.B. M. Baker: Is there a reproducibility crisis, in: Nature, 533. Jg. (2016), Nr. 7604, S. 452-454.

zu wenig Aufmerksamkeit schenkt und die Nicht-Bestätigung vorher gefeierter Ergebnisse oft „laut“ beschweigt?

Verbreitete Fehlinterpretationen statistischer Signifikanzaussagen

Im Kern geht es bei der diagnostizierten Vertrauenskrise um die vernünftige Bewertung der Zuverlässigkeit einer wissenschaftlichen Aussage. Bei diesem Bewertungsvorgang kommen auch die in der Wissenschaftskommunikation Tätigen nicht umhin, sich mit Fragen der Inferenz und Induktion auseinanderzusetzen und statistische Signifikanzaussagen kritisch einzuordnen. Bevor man überhaupt die Frage nach einer erfolgreichen Kommunikation mit der Öffentlichkeit stellen kann, ist ja die Frage zu beantworten, wie die Kommunikation zwischen Wissenschaftlern und den in der Wissenschaftskommunikation Tätigen gelingen kann. Hierbei entsteht häufig ein Kommunikationsproblem, wie es klassischer kaum sein könnte: Man glaubt sich zu verstehen, aber man versteht sich nicht. Das liegt daran, dass die Begriffe der statistischen Fachsprache vielfach „quer“ zu jeglichem umgangssprachlichen Verständnis liegen. Wie die Reproduktionskrise belegt, führt dies bereits zwischen den Vertretern der Fachwissenschaften (z.B. Mediziner, Ökonomen, Psychologen) und der Statistik zu einem Kommunikationsproblem. Anscheinend interpretieren bereits viele Fachwissenschaftler die statistischen Ergebnisse ihrer eigenen empirischen Studien falsch. Dies setzt sich dann zwangsläufig in der Kommunikation zwischen den Fachwissenschaftlern und den in der Wissenschaftskommunikation Tätigen fort.

Bei Fragen nach der Glaubwürdigkeit von Studienergebnissen kommt man nicht am Inferenzschluss vorbei, da regelmäßig nicht ganze Populationen, sondern lediglich Stichproben untersucht werden können. Beim Inferenzschluss geht es um die Frage, welche Schlussfolgerungen man aus den Befunden einer Zufallsstichprobe (bei der man hoffentlich alles richtig gemessen hat) für die Grundgesamtheit ziehen kann. Über Jahrzehnte hat sich innerhalb der vorherrschenden „frequentistischen Statistik“ die Konvention herausgebildet, als Hilfsmittel für den Inferenzschluss auf den p-Wert zurückzugreifen und bei p-Werten bis zu 0,05 von statistisch signifikanten Ergebnissen zu sprechen. Häufig wird der p-Wert auch als Irrtumswahrscheinlichkeit bezeichnet. Beide Begriffe sind hochproblematisch, da sie insbesondere drei Fehlinterpretationen Vorschub leisten:[14]

1. Entgegen dem umgangssprachlichen Verständnis (und dem mancher Fachwissenschaftler) ist es falsch, den Begriff „signifikant“ mit „groß/wichtig“ gleichzusetzen.

2. Es ist ein auch unter Fachwissenschaftlern verbreiteter Trugschluss, ein „nicht signifikantes“ Ergebnis als Indiz oder gar Nachweis dafür zu werten, dass kein (bedeutender) Effekt vorliegt.

3. Der Begriff „Irrtumswahrscheinlichkeit“ provoziert geradezu die Fehlinterpretation, der p-Wert entspreche der Wahrscheinlichkeit der Nullhypothese (= kein Effekt) und damit der Wahrscheinlichkeit, einen Irrtum zu begehen, wenn man die Nullhypothese ablehnt.

Gleichzeitig sind es gerade die vordergründig selbsterklärenden Formulierungen „statistisch signifikant/nicht statistisch signifikant“ und „Irrtumswahrscheinlichkeit“, die in der Wissenschaftskommunikation genutzt werden, um dem Laien zu vermitteln, wie glaubwürdig bestimmte Studienergebnisse sind. Viel zu oft wird dabei „statistisch signifikant“ mit „wissenschaftlich“ gleichgesetzt. Das wirft die Frage auf, ob die aus den Bedeutungsunterschieden zwischen Fach- und Alltagssprache resultierenden Missverständnisse in der Wissenschaftskommunikation gelegentlich billigend hingenommen werden. Oder werden sie aufgrund von Eigeninteressen sogar manchmal gezielt ausgenutzt? Man kann ja etwas sagen, das in der Fachsprache richtig ist (das ist eine bequeme Rückfallposition), und gleichzeitig darauf bauen, dass das Gesagte umgangssprachlich rezipiert und in einer bestimmten (und möglicherweise erwünschten) Art und Weise missverstanden wird.

Nachstehend ein kleines Beispiel als anekdotische Evidenz für ein wahrscheinlich nicht intendiertes Missverständnis: Vor einiger Zeit stand in der ZEIT eine Meldung, wie es viele ähnliche gibt. Diese lautete sinngemäß:[15] Eine wissenschaftliche Studie hat herausgefunden, dass ein nasses Tuch im Nacken nicht gegen Nasenbluten hilft. Im weiteren Text wurde als Begründung für diese Aussage darauf verwiesen, dass wissenschaftlich „kein statistisch signifikanter Effekt“ der Behandlung (= nasses Tuch) nachgewiesen werden konnte. Das war nur eine kurze Meldung und die Studie selbst lag dem Leser natürlich nicht vor. Man kann aber davon ausgehen, dass hier zumindest bei der öffentlichen Wahrnehmung der Studienergebnisse ein Denkfehler entstanden ist. Vermutlich war das tatsächliche Studienergebnis, dass man die Nullhypothese (= kein Effekt des nassen Tuches) nicht mit dem üblicherweise geforderten Signifikanzniveau von 0,05 verwerfen konnte. Damit hat

14 Ein ebenfalls sprachlich bedingtes Missverständnis entsteht bei einem weiteren gebräuchlichen Hilfsmittel der frequentistischen Statistik, dem sogenannten Konfidenzintervall. So gibt das üblicherweise genutzte 95%-Konfidenzintervall – entgegen jeglicher alltagssprachlichen Assoziation – *nicht* an, dass der Wert der untersuchten Größe mit 95%iger Wahrscheinlichkeit innerhalb der angegebenen Intervallgrenzen liegt.

15 Vgl. C. Drösser: Hilft ein nasser Lappen im Nacken gegen Nasenbluten?, ZEIT Online vom 14.7.2016, http://www.zeit.de/2016/28/hausmittel-nasenbluten-lappen-nass-epistaxis-stimmts.

man aber diese Nullhypothese nicht bestätigt; man hat noch nicht einmal ein Indiz für die Nullhypothese gefunden. Möglicherweise hatte der Wissenschaftler in der statistischen Fachsprache alles ohne Fehler formuliert, ist dann aber falsch interpretiert worden und hat nichts mehr gegen die Fehlinterpretation unternommen.

Vor dem Hintergrund der allgegenwärtigen Fehlinterpretationen statistischer Signifikanzaussagen seien nachstehend die virulentesten Missverständnisse kurz angesprochen.[16]

(1) Falsche Gleichsetzung von „signifikant" mit „groß/wichtig"

Die Gefahr, „statistisch signifikant" mit „groß/wichtig" gleichzusetzten, ist insbesondere dann hoch, wenn das Adjektiv „statistisch" weggelassen und nur von „signifikanten" und „nicht signifikanten" Effekten gesprochen wird. In der Folge findet man häufig Formulierungen, die signifikante im Vergleich zu nicht signifikanten Ergebnissen mit dem Adjektiv „stärker" oder „mehr" belegen. Das ist falsch. Wenn eine Variable X einen „signifikanten" Einfluss auf eine Variable Y hat, bedeutet das lediglich, dass die bedingte Wahrscheinlichkeit gering ist, dass der beobachtete (oder ein stärkerer) Effekt als Zufallsbefund bei einer häufig wiederholten Stichprobenziehung (daher der Name „frequentistische Statistik") auftauchen würde, wenn er in der Grundgesamtheit nicht da wäre.

Obwohl große Stichproben oft als vorteilhaft wahrgenommen werden, ist die Gleichsetzung von (statistisch) „signifikant" und „groß/wichtig" gerade bei hohen N ein Problem. Dies liegt daran, dass die p-Werte *ceteris paribus* mit steigendem N sinken. Jeder Effekt, egal wie bedeutungslos er ist, wird bei steigenden N irgendwann „statistisch signifikant". Zur Illustration folgendes Beispiel: Zwei Futtermittel werden verglichen, indem 100 Schweine (Gruppe A) das herkömmliche Futtermittel und 100 Schweine (Gruppe B) ein „verbessertes" Futtermittel bekommen. Gruppe A weist durchschnittliche Tageszunahmen von 700 g auf. In Gruppe B liegen diese bei 701 g. In beiden Gruppen beträgt die Standardabweichung der Tageszunahmen 10 g. Es ergibt sich ein p-Wert von 0,24 (einseitiger t-Test) und man würde den gefundenen Unterschied als „nicht statistisch signifikant" bezeichnen. Findet man bei einer Gruppengröße von jeweils 600 (1000) Schweinen genau denselben Unterschied von 1 g zwischen den beiden Gruppen, ergibt sich bereits ein p-Wert von 0,042 (0,013). Der Effekt wird also zunehmend „signifikanter", er bleibt aber von seiner Größe her ökonomisch bedeutungslos.

(2) Fehlschlüsse bei Überschreiten des konventionellen Signifikanzniveaus

Die präzise Formulierung für ein nicht statistisch signifikantes Ergebnis ($p > 0,05$) lautet: Die Nullhypothese, dass der Regressor X keinen Einfluss auf Y hat, kann nicht mit dem üblicherweise geforderten Signifikanzniveau von maximal 0,05 abgelehnt werden. Dies entspricht dem „Satz vom ausgeschlossenen Dritten", nach dem eine Aussage so zu formulieren ist, dass entweder sie selbst oder ihre Verneinung zutrifft. Die Aussage „Hans ist entweder blond oder nicht blond" ist richtig. Die Aussage „Wenn Hans nicht blond ist, ist er schwarzhaarig" ist dagegen eine Verletzung des Satzes vom ausgeschlossenen Dritten, durch die eine falsche Dichotomie entsteht. Ein analoger Trugschluss droht auch bei der Interpretation von p-Werten über 0,05, wenn laxe Formulierungen wie die folgenden genutzt werden:

- Der Einfluss von X auf Y ist nicht statistisch signifikant.
- Der Einfluss von X auf Y ist statistisch nicht signifikant.
- Der Einfluss von X auf Y ist nicht signifikant.

Von der letzten Formulierung aus, die bereits nahelegt, dass man gefunden habe, dass kein (wichtiger) Effekt da ist, ist es nur ein kurzer Weg zur eindeutig falschen Schlussfolgerung: „Die Studie zeigt, dass ein (relevanter) Einfluss von X auf Y nicht vorhanden ist". Schuld an diesem Trugschluss sind Formulierungen, die die falsche Dichotomie „*entweder* Ablehnung der Nullhypothese *oder* Bestätigung der Nullhypothese" nahelegen.[17] Der Fehler findet sich auch in Formulierungen, bei denen er nicht auf den ersten Blick offensichtlich ist. So werden nicht signifikante Ergebnisse oft dahingehend kommentiert, dass sie im Widerspruch zu (theoretischen) Erwartungen stehen, die die Existenz des Effekts nahelegen. Das kann man aber nicht sagen, da $p > 0,05$ kein Indiz dafür ist, dass der Effekt nicht vorliegt.

Bei der Kommunikation von Forschungsergebnissen an die Öffentlichkeit sind falsche Dichotomien möglicherweise das größte Problem. Fachjournalisten und Politikern, die im Kampf um die öffentliche Aufmerksamkeit und Meinung stehen, ist eine „passende" Meldung „X hat keinen Einfluss auf Y!" vielleicht oft lieber als eine wissenschaftlich zutreffende, aber „langweilige" oder „unpassende" Aussage, dass noch keine Aussage möglich ist.

16 Auf weitere Gründe für die Nicht-Reproduzierbarkeit von Studienergebnissen (z.B. Selection-Bias) sowie grundsätzliche inferenzstatistische Methodenfragen (einfache Signifikanztests versus Poweranalyse, Bayes'sche Statistik versus frequentistische Statistik etc.) wird an dieser Stelle nicht eingegangen.

17 Motulsky bringt das treffend auf den Punkt: „The absence of evidence is not evidence of absence." Vgl. H. J. Motulsky: Essential Biostatistics. A Nonmathematical Approach, Oxford 2016, S. 19.

(3) Fehlschlüsse bei Unterschreiten des konventionellen Signifikanzniveaus

(Geringe) p-Werte werden oft als (geringe) Wahrscheinlichkeit der Nullhypothese fehlinterpretiert. Eine maßgebliche Ursache für diesen *Inverse Probability Error*[18] ist die statistische Konvention, den p-Wert als „Irrtumswahrscheinlichkeit" zu bezeichnen. Trotz dieser Bezeichnung gibt der p-Wert aber nicht die Wahrscheinlichkeit an, dass die Nullhypothese (kein Effekt) zutrifft. Er bezeichnet damit auch nicht die (*A-posteriori*-)Wahrscheinlichkeit, bei Ablehnung der Nullhypothese einen Irrtum zu begehen. Der p-Wert ist lediglich die bedingte Wahrscheinlichkeit, dass bei häufig wiederholten Stichprobenziehungen der gefundene Effekt (oder ein stärkerer) wieder beobachtet werden würde, wenn man als Gedankenexperiment unterstellt, dass in der Grundgesamtheit kein Effekt da sei.[19]

Der Sachverhalt lässt sich an einem Münzwurfbeispiel zeigen, bei dem man vorab mit 1%iger Wahrscheinlichkeit eine manipulierte Münze [P(Kopf) = 0,75] und mit 99%iger Wahrscheinlichkeit eine nicht manipulierte (ideale) Münze [P(Kopf) = 0,5] zieht. Nun wirft man die gezogene Münze fünfmal und beobachtet 5x Kopf. Bei einer idealen Münze (= kein Effekt), wäre bei sehr vielen Wiederholungen des Experiments „fünfmaliger Münzwurf" nur in 3,125% (= $0{,}5^5$) der Fälle 5x Kopf zu erwarten. Diese bedingte Wahrscheinlichkeit entspricht dem p-Wert. Sie ist aber nicht die Wahrscheinlichkeit, bei Verwerfung der Nullhypothese „ideale Münze" einen Fehler zu machen. Hierfür muss man noch wissen, wie hoch bei der manipulierten Münze die Wahrscheinlichkeit für 5x Kopf ist. Sie beträgt 23,73% (= $0{,}75^5$). Man muss zudem die vor dem Münzwurfexperiment bekannten (*A-priori*-) Wahrscheinlichkeiten von 1% und 99% berücksichtigen, dass man eine manipulierte bzw. eine ideale Münze gezogen hat. Nach dem Satz von Bayes kommt man nach dem Wurfexperiment auf eine (*A-posteriori*-)Wahrscheinlichkeit von 92,88% [= 0,03125·0,99/(0,03125·0,99+0,237·0,01)], einen Irrtum zu begehen, wenn man die Nullhypothese „ideale Münze" verwirft (False Discovery Rate). Trotz des p-Werts von 0,03125 wird man also die Nullhypothese nicht verwerfen. Der Informationsgewinn durch das Experiment führt lediglich dazu, dass man die *A-priori*-Wahrscheinlichkeit von 99% revidiert und a posteriori (also nach dem Experiment) nur noch mit 92,88%iger Wahrscheinlichkeit davon ausgeht, dass man es mit einer idealen Münze zu tun hat.[20]

Das Beispiel zeigt, dass die Praxis, die Einhaltung eines bestimmten Signifikanzniveaus als Bedingung für die Ablehnung der Nullhypothese anzusehen, zwar eine verbreitete Konvention ist, aber nicht mit der Einhaltung einer einheitlichen und akzeptablen Obergrenze für die False Discoveries im Einklang steht. Wie das Münzwurfbeispiel zeigt, können auch geringe p-Werte zu einer inakzeptabel hohen Wahrscheinlichkeit einer False Discovery führen. Das Beispiel zeigt auch, wie unsinnig eine Fixierung auf die Ergebnisse einer einzelnen Studie ist. Eigentlich geht es darum, einzuschätzen, wie glaubwürdig eine bestimmte wissenschaftliche Aussage über einen bestimmten Aspekt der Welt ist. Bei ausschließlichem Rückgriff auf die „statistische Signifikanz" und damit den p-Wert, der alleine auf der Grundlage der jeweiligen Studienstichprobe berechnet wird, abstrahiert man von jeglichem Vorwissen. Wissenschaft und Erkenntnisfortschritt beruhen aber auf Vorarbeiten, die mit den eigenen Ergebnissen kontrastiert und zusammengeführt werden müssen.

Ethischer Imperativ der Wissenschaftskommunikation

Eine moralisch vertretbare Wissenschaftskommunikation muss darlegen, zu welchem Wissensgewinn – ausgehend vom vorhandenen Vorwissen – eine bestimmte Studie geführt hat und welche Korrekturen gegebenenfalls am bisher unterstellten Wissensbestand vorzunehmen sind. Hierzu muss man sich mit Meta-Analysen und Bayes'scher Statistik beschäftigen. Daraus kann man viel lernen, z.B. dass mehrere Studien mit von der Wirkungsrichtung gleichen, aber nicht statistisch signifikanten Ergebnissen in der Summe eine außerordentlich starke Evidenz für das Vorhandensein eines Effekts darstellen können. Borenstein et al. kommentieren den Sachverhalt unter der Überschrift „An Ethical Imperative" wie folgt: „rather than looking at any study in isolation, we need to look at the body of evidence."[21] Das bedeutet, dass alle relevanten Informationen zur Einschätzung der Glaubwürdigkeit wissenschaftlicher Ergebnisse zu berücksichtigen sind. Dies schließt Negativergebnisse, also

18 J. Cohen: The earth is round (p < 0.05), in: American Psychologist, 49. Jg. (1994), H. 12, S. 997-1003.

19 Das Interessante an Fehler (2) und (3) ist, dass sie nicht einmal „als Fehler konsistent" sind und trotzdem gelegentlich gemeinsam auftreten. Wenn man den p-Wert gemäß Fehler (2) als Wahrscheinlichkeit der Nullhypothese interpretiert, würde man ja beispielsweise aus p = 0,15 schließen, dass die Wahrscheinlichkeit der Null bei 15% und die Wahrscheinlichkeit der Alternativhypothese bei 85% liege. Nach dieser Fehlinterpretation dürfte man den Fehler (2), ein nicht signifikantes Ergebnis (p = 0,15) als Bestätigung der Nullhypothese anzusehen, eigentlich nicht mehr machen.

20 Motulsky kommentiert das Spannungsverhältnis zwischen dem, was der p-Wert aussagt, und dem, was man letztlich wissen will, wie folgt: „Statistical hypothesis testing [based on p-values], ‚does not tell us what we want to know, and we so much want to know what we want to know that, out of desperation, we nevertheless believe that it does' (J. Cohen, a.a.O.). […] The question we want to answer is: Given these data, how likely is the null hypothesis? The question that a P value answers is: Assuming the null hypothesis is true, how unlikely are these data? These two questions are distinct, and so have distinct answers." H. J. Motulsky: Common Misconceptions about Data Analysis and Statistics, in: The Journal of Pharmacology and Experimental Therapeutics, 351. Jg. (2014), H. 8, S. 204.

21 M. Borenstein, L. V. Hedges, J. P. T. Higgins, H. R. Rothstein: Introduction to Meta-Analysis, Chichester 2009, S. xxi.

Informationen zur Nicht-Reproduzierbarkeit von Studien ein. Nur so – und hier schließt sich der Kreis – wird eine erfolgreiche und sachlich angemessene Wissenschaftskommunikation ermöglicht, die nicht das Vertrauen der Öffentlichkeit durch die Meldung einer Vielzahl nicht reproduzierbarer „statistisch signifikanter" Ergebnisse verspielt.

Abschließende Bemerkungen

Für die in der Wissenschaftskommunikation Tätigen sowie alle, die am Thema „Einschätzung der Glaubwürdigkeit wissenschaftlicher Ergebnisse" interessiert sind, seien abschließend ein paar weiterführende Literaturhinweise gegeben. Hirschauer et al.[22] systematisieren in ihrem Überblicksbeitrag die wichtigsten p-Wert-bezogenen Fehler und geben Hinweise darauf, wie die Probleme behoben werden könnten. Gigerenzer[23] zeigt, dass viele Menschen (inklusive Wissenschaftler) Probleme haben, Wahrscheinlichkeiten und insbesondere bedingte Wahrscheinlichkeiten (wie z.B. den p-Wert) richtig zu interpretieren. Er gibt auch Hinweise, wie in der Kommunikation von Sachverhalten, bei denen es um Wahrscheinlichkeiten geht, die Gefahr von Missverständnissen verringert werden kann. Borenstein et al.[24] bieten eine Einführung in Meta-Analysen und damit in die Frage, wie man herausfindet, was schon an Informationen vorliegt. Das ist die zentrale Voraussetzung für eine vernünftige Einschätzung, was eine neue Studie an zusätzlichen Erkenntnissen bringt. Zyphur und Oswald[25] geben einen anschaulichen Einblick in die Bayes'sche Inferenz. Mit Hilfe der Bayes'schen Statistik kann man die Ergebnisse einer Studie mit dem Vorwissen formal verknüpfen (siehe obiges Münzwurfbeispiel) und so zu einer als Wahrscheinlichkeit ausgedrückten Glaubwürdigkeit einer Tatsachenbehauptung kommen. Baker[26] gibt einen Überblick über eine aktuelle Befragung der Zeitschrift Nature zur Einschätzung der Reproduktionskrise unter Wissenschaftlern. In diesem Beitrag verweist Baker unter anderem auf eine Studie von Begley und Ellis,[27] die im Bereich der Krebsbiologie lediglich eine Rate von 10% reproduzierbarer Ergebnisse gefunden haben. Für den Bereich der Ökonomie beschreiben Duvendack et al.[28] – ausgehend vom Problem der Verzerrung wissenschaftlicher Publikationen zugunsten von Positivergebnissen (*Publikationbias*) – die Replikationspolitik aller 333 ökonomischen Web-of-Science-Zeitschriften und kommen zum Schluss, dass Replikation immer noch einen geringen Stellenwert hat.[29] Interessant für alle in der Wissenschaftskommunikation Tätigen ist sicherlich auch das „ASA Symposium on Statistical Inference", das die American Statistical Association im Oktober 2017 unter dem Titel „Scientific Method for the 21st Century: A World Beyond $p < 0.05$" veranstaltet.[30]

Wenn das Vertrauen der Öffentlichkeit in die Wissenschaft erhalten bzw. wiederhergestellt werden soll, müssen durch die Wissenschaftskommunikation adäquate Informationen bereitgestellt werden, mit deren Hilfe die Glaubwürdigkeit wissenschaftlicher Aussagen vernünftig eingeschätzt werden kann. Wissenschaftskommunikation gelingt nur, wenn der Öffentlichkeit statt Aufsehen erregender (und dann häufig falscher) Einzelmeldungen ein zutreffendes Bild des Wissensstandes in einem bestimmten Gebiet vermittelt wird. Das schließt spannende „Neuigkeiten" durchaus ein, darf sich aber nicht auf diese beschränken.

22 N. Hirschauer, O. Mußhoff, S. Grüner, U. Frey, I. Theesfeld, P. Wagner: Die Interpretation des *p*-Wertes – Grundsätzliche Missverständnisse, in: Jahrbücher für Nationalökonomie und Statistik, 236. Jg. (2016), Nr. 5, S. 557-575.

23 G. Gigerenzer: Das Einmaleins der Skepsis. Über den richtigen Umgang mit Zahlen und Risiken, Berlin 2002.

24 M. Borenstein, L. V. Hedges, J. P. T. Higgins, H. R. Rothstein, a.a.O.

25 M. J. Zyphur, F. L. Oswald: Bayesian Estimation and Inference: A User's Guide, in: Journal of Management, 41. Jg. (2015), H. 2, S. 390-420.

26 M. Baker: Is there a reproducibility crisis ..., a.a.O.

27 C. G. Begley, L. M. Ellis: Drug development: Raise standards for preclinical cancer research, in: Nature, 483. Jg. (2012), Nr. 7391, S. 531-533.

28 M. Duvendack, R. W. Palmer-Jones, W. R. Reed: Replications in Economics: A Progress Report, in: Econ Journal Watch, 12. Jg. (2015), H. 2, S. 164-191.

29 M. Duvendack (University of East Anglia, Großbritannien) und R. Reed (University of Canterbury, Neuseeland) betreiben „The Replication Network" für Ökonomen. Außerhalb der Ökonomie scheinen institutionalisierte Bemühungen zur Förderung von Replikationen aber stärker ausgeprägt zu sein. Im Bereich der Medizin wurde 2013 die globale Initiative „All Trials Registered/All Results Reported" (http://www.alltrials.net/) gestartet, die sehr breite Unterstützung von Wissenschaftlern und Wissenschaftsorganisationen gefunden hat. In manchen Bereichen gibt es sogar eigenständige Zeitschriften, wie z.B. das Journal of Negative Results in BioMedicine und die in mehreren Bereichen zu findenden All Results Journals, deren explizite Politik es ist, positive *und* negative Ergebnisse zu veröffentlichen.

30 Nähere Informationen unter http://ww2.amstat.org/meetings/ssi/2017/?utm_source=informz&utm_medium=email&utm_campaign=asa&_zs=WXUXe1&_zl=SNWb3.

Title: *False Discoveries and Misinterpretations of Scientific Findings – Implications for Science Communication*

Abstract: *High-quality science communication to the public depends to a large extent on the way research findings are translated into comprehensible language and common speech. In this communicative process, a reasonable evaluation of the trustworthiness of empirical findings, based on an adequate interpretation of statistical analyses, is absolutely crucial. This paper's authors argue that the credibility of science is jeopardised by two compromising developments within science itself: on the one hand, an inflation of ostensible empirical evidence related to misuses and misinterpretations of the concept of statistical significance, and, on the other, a sensationalist overvaluation of the results of single studies instead of an adequate representation of the available body of evidence in a given scientific field.*

JEL Classification: *A20, C12, C18*