DE GRUYTER OLDENBOURG Journal of Economics and Statistics 2016; 236(5): 557–575

Discussion Paper

Norbert Hirschauer*, Oliver Mußhoff, Sven Grüner, Ulrich Frey, Insa Theesfeld und Peter Wagner

Die Interpretation des *p*-Wertes – Grundsätzliche Missverständnisse

Interpreting *p*-values – Common flaws and misconceptions

DOI 10.1515/jbnst-2015-1030
Received November 24, 2015; revised April 9, 2016; accepted July 7, 2016

Zusammenfassung: Der *p*-Wert wird vielfach als Goldstandard für Inferenzschlüsse angesehen. Zur Validierung statistischer Zusammenhänge hat sich die Konvention herausgebildet, möglichst geringe *p*-Werte einzufordern und bei Werten unterhalb gewisser Schwellen (z. B. 0,05) von *statistisch signifikanten* Ergebnissen zu sprechen. Häufig wird der *p*-Wert auch als *Irrtumswahrscheinlichkeit* bezeichnet. Beide Begriffe sind problematisch, da sie Missverständnissen Vorschub leisten. Hinzu kommt, dass das sog. *p-hacking*, d. h. die gezielte Suche nach Auswertungen, die zu statistisch signifikanten Ergebnissen führen, Verzerrungen hervorrufen und die Rate falscher Entdeckungen (*false discovery rate*) erhöhen kann. Fehlinterpretationen des *p*-Wertes und auswertungsbedingte Verzerrungen wurden über die Jahrzehnte hinweg immer wieder kritisch diskutiert. In der empirischen Forschung scheinen sie aber persistent zu sein und in den letzten Jahren wurde die *p*-Wert-Debatte wegen der Nicht-Reproduzierbarkeit vieler Studien zunehmend intensiv geführt. Angesichts der über die Disziplinen verstreuten und oft auf Einzelaspekte abzielenden Literatur zur *p*-Wert-Problematik beschreibt dieser Methodenkommentar systematisch die wichtigsten Probleme und diskutiert die entsprechenden Lösungsvorschläge.

***Corresponding author: Norbert Hirschauer,** Professur für Unternehmensführung im Agribusiness, Martin-Luther-Universität Halle-Wittenberg, 06099 Halle (Saale), E-mail: norbert.hirschauer@landw.uni-halle.de
Oliver Mußhoff, Arbeitsbereich Landwirtschaftliche Betriebslehre, Georg-August-Universität Göttingen, Platz der Göttinger Sieben 5, 37073 Göttingen
Sven Grüner, Professur für Unternehmensführung im Agribusiness, Martin-Luther-Universität Halle-Wittenberg, 06099 Halle (Saale)
Ulrich Frey, Insa Theesfeld, Professur für Agrar-, Umwelt- und Ernährungspolitik, Martin-Luther-Universität Halle-Wittenberg, 06099 Halle (Saale)
Peter Wagner, Professur für Landwirtschaftliche Betriebslehre, Martin-Luther-Universität Halle-Wittenberg, 06099 Halle (Saale)

Bereitgestellt von | De Gruyter / TCS
Angemeldet
Heruntergeladen am | 07.03.17 15:30

Schlüsselwörter: Satz von Bayes, Hypothesentest, Irrtumswahrscheinlichkeit, *p-hacking*, statistische Signifikanz

Abstract: The *p*-value is often considered as the gold standard in inferential statistics. The standard approach for evaluating empirical evidence is to equate low *p*-values with a high degree of credibility and to refer to findings with *p*-values below certain thresholds (e.g., 0.05) as *statistically significant*. The *p*-value is also referred to as *error probability*. Both terms are problematic as they invite serious misconceptions. In addition, researchers' fixation on obtaining statistically significant results may introduce biases and increase the *rate of false discoveries*. Misinterpretations of the *p*-value as well as the introduction of bias through arbitrary analytical choices (*p-hacking*) have been critically discussed in the literature for decades. Nonetheless, they seem to persist in empirical research and criticisms of inappropriate approaches have increased in the recent past – mainly due to the non-replicability of many studies. Unfortunately, the critical concerns that have been raised in the literature are not only scattered over many academic disciplines but often also linguistically confusing and differing in their main reasons for criticisms. Against this background, our methodological comment systematizes the most serious flaws and discusses suggestions of how best to prevent future misuses.

Keywords: Bayes theorem, hypothesis testing, error probability, *p-hacking*, statistical significance

1 Einleitung

Der *p*-Wert wird vielfach als Goldstandard für Inferenzschlüsse angesehen. *p*-Werte werden gemeinhin als Hilfsmittel verstanden, Fehler 1. Art zu vermeiden. Darunter versteht man die Schlussfolgerung, dass ein Effekt da sei, obwohl er nicht da ist. Bei statistischen Auswertungen hat sich die Konvention herausgebildet, möglichst geringe *p*-Werte einzufordern und bei Werten unterhalb gewisser Schwellen (z. B. 0,05) von *statistisch signifikanten* Ergebnissen zu sprechen. Häufig wird der *p*-Wert auch als *Irrtumswahrscheinlichkeit* bezeichnet. Beide Begriffe sind problematisch, da sie Missverständnissen Vorschub leisten.

Erstens entsteht ein semantisches Missverständnis, wenn der Begriff „signifikant" mit „groß/wichtig" gleichgesetzt wird. Zweitens kann es zu Trugschlüssen kommen, wenn bei der Interpretation von statistisch nicht-signifikanten Ergebnissen Formulierungen genutzt werden, die den Satz vom ausgeschlossenen Dritten missachten und eine Bestätigung der Nullhypothese (kein Effekt) nahelegen. Drittens besteht die Gefahr, dass Forscher bei der Ausgestaltung statistischer Analysen durch das sog. *p-hacking* Verzerrungen produzieren und nur das publizieren, was im Sinne

Bereitgestellt von | De Gruyter / TCS
Angemeldet
Heruntergeladen am | 07.03.17 15:30

der Produktion signifikanter Ergebnisse „funktioniert“ hat.[1] Viertens leistet der Begriff „Irrtumswahrscheinlichkeit“ der semantischen Fehlinterpretation Vorschub, der *p*-Wert bezeichne die als *false discovery rate* bezeichnete Wahrscheinlichkeit, einen Irrtum zu begehen, wenn man die Nullhypothese ablehnt.[2]

Die o.g. Probleme wurden seit Jahrzehnten immer wieder diskutiert. Dies gilt insbesondere für Mediziner und Psychologen.[3] Im letzten Jahrzehnt wurden die Probleme zunehmend stärker wahrgenommen.[4] Neben Fehlinterpretationen liegt dies unter anderem an dem mit dem *p-hacking* verbundenen Problem der Nicht-Reproduzierbarkeit (Replikationskrise). In einer drastischen Reaktion auf die *p*-Wert-Krise haben die Herausgeber des Journals *Basic and Applied Social Psychology* Anfang 2015 die Verwendung von *p*-Werten in Veröffentlichungen ganz untersagt (Trafimow/Marks 2015).[5] Dieses Verbot sowie eine Vielzahl *p*-Wert kritischer Beiträge in hochrangigen Journalen bis hin zu *Nature* (vgl. Nuzzo 2014) und *ScienceNews* (vgl. Siegfried 2014) haben das Problembewusstsein in den empirischen Wissenschaften allgemein erhöht. Anfang März 2016 hat die American Statistical Association (ASA) sogar eine offizielle Erklärung abgegeben, wie *p*-Wert bezogene Fehler zu vermeiden sind (Wasserstein/Lazar 2016). Interessanterweise scheint in den Wirtschaftswissenschaften die Rezeption und Beteiligung an der Debatte eher schwach zu sein. Bedeutsame Ausnahmen sind die Beiträge von Ziliak und McCloskey (2008) und Krämer (2011), die eine große Verbreitung *p*-Wert bezogener Fehlinterpretationen in Beiträgen des *American*

1 Aufgrund seiner Prägnanz und Bekanntheit verwenden wir in diesem Beitrag den von Nuzzo (2014) popularisierten Begriff des *p-hacking* für signifikanzsuchendes Verhalten im Forschungsprozess.

2 Um wenig gebräuchliche deutsche Formulierungen wie z. B. „Rate falscher Entdeckungen“ zu vermeiden, verwenden wir in diesem Beitrag in Anlehnung an Colquhoun (2014) und Motulsky (2014) den Begriff der *false discovery rate* für (die sich nach einer Untersuchung ergebende Wahrscheinlichkeit der Nullhypothese und damit) die Wahrscheinlichkeit, einen Irrtum zu begehen, wenn man die Nullhypothese ablehnt. Für den begrifflichen Zusammenhang mit dem *multiple testing* Problem sei bspw. auf Storey (2011) verwiesen.

3 Ausgewählte Beispiele sind Sedlmeier und Gigerenzer (1989), Kirk (1996) sowie Sterne und Smith (2001). Aus systematischer Sicht besonders hervorzuheben ist der Reviewbeitrag von Nickerson (2000), der bereits vor eineinhalb Jahrzehnten einen sehr umfassenden Einblick in die Problematik gegeben hat.

4 Vgl. z. B. Ioannidis (2005), Goodman (2008), Simmons et al. (2011), Colquhoun (2014) und Motulsky (2014) sowie die Lehrbücher von Cumming (2012), Kline (2013) und Motulsky (2016).

5 Bereits in den 90er Jahren hat eine „Task Force on Statistical Inference“ der American Psychological Association sich mit zunehmenden Forderungen nach einem *p*-Wert Verbot in Publikationen auseinander gesetzt, dies dann aber trotz der berechtigten Kritik als zu extreme Maßnahme verworfen (vgl. Kline 2013: 21).

Bereitgestellt von | De Gruyter / TCS
Angemeldet
Heruntergeladen am | 07.03.17 15:30

Economic Review bzw. *German Economic Review* dokumentieren. Über die Gründe der geringen Rezeption der *p*-Wert Debatte in den Wirtschaftswissenschaften kann nur spekuliert werden. Angesichts der über die Disziplinen verstreuten und oft auf Einzelaspekte abzielenden Literatur fehlt möglicherweise vielfach ein systematischer Überblick über den Problemkomplex. Möglicherweise bestehen auch grundsätzliche Defizite in der Ausbildung.[6] Vor diesem Hintergrund wollen wir in diesem Methodenkommentar die wichtigsten der in der Literatur diskutierten Probleme systematisch und übersichtlich darstellen.

2 Probleme und Lösungsansätze

2.1 Fälschliche Gleichsetzung von „signifikant“ mit „groß“

Problembeschreibung: Geringe *p*-Werte werden üblicherweise mit dem Begriff „statistisch signifikant“ belegt. Eine semantische Fehlinterpretation entsteht, wenn dieser Begriff umgangssprachlich als Synonym von „groß/wichtig“ interpretiert wird. Die Gefahr ist insbesondere hoch, wenn das Adjektiv „statistisch“ weggelassen und nur noch von „signifikanten“ und „nicht-signifikanten“ Effekten gesprochen wird. In der Folge findet man dann häufig Formulierungen, die ein signifikantes Ergebnis im Vergleich zu einem nicht-signifikanten Ergebnis mit dem Adjektiv „stärker“ oder „mehr“ belegen. Das ist falsch (Motulsky 2014).[7] Wenn eine Variable *X* einen statistisch signifikanten Einfluss (Effekt) auf eine Variable *Y* hat, heißt das nicht, dass es sich um einen

6 Kline (2013: 10) spricht von einem „major educational failure“ und bereits vor eineinhalb Jahrzehnten konstatieren Sellke et al. (2001: 71) im *The American Statistician*: „The standard approach in teaching–of stressing the formal definition of a *p*-value while warning against its misinterpretation–has simply been an abysmal failure.“

7 Dieses Missverständnis scheint auch in den Wirtschaftswissenschaften hartnäckig zu sein. In einer Auswertung aller 182 ökonometrischen Beiträge der 80er Jahre in der *American Economic Review* finden McCloskey und Ziliak (1996: 106–107): „70 percent of the empirical papers in the American Economic Review papers did not distinguish statistical significance from economic, policy, or scientific significance. [...] 59 percent used the word ‘significance’ in ambiguous ways, at one point meaning ‘statistically significantly different from the null,’ at another ‘practically important’ or ‘greatly changing our scientific opinions,’ with no distinction.“ Krämer (2011: 462) findet in einer analogen Untersuchung aller 110 Beiträge mit statistischen Signifikanztests im *German Economic Review* von 2000 bis 2010, dass für über 56 % der Beiträge gilt: „Confusion of economic and statistical significance of estimated coefficients or effects.“ Auch in der ASA-Erklärung (Wasserstein/Lazar 2016: 10) wird das Problem angesprochen: „Smaller *p*-values do not necessarily imply the presence of larger or more important effects, and larger *p*-values do not imply a lack of importance or even a lack of effect.“

Bereitgestellt von | De Gruyter / TCS
Angemeldet
Heruntergeladen am | 07.03.17 15:30

großen oder wichtigen Einfluss handelt. Es bedeutet vielmehr, dass die Wahrscheinlichkeit gering ist, dass das beobachtete Ergebnis als Zufallsbefund auftauchen würde, wenn kein Effekt da wäre. „Signifikant“ bedeutet also, dass man eine geringe Wahrscheinlichkeit hat, den Effekt in den Daten zu finden, wenn er gar nicht vorliegt.

Obwohl große Stichproben oft ohne weiteres Nachdenken als vorteilhaft wahrgenommen werden, ist die Gleichsetzung von (statistisch) „signifikant“ und „wichtig“ gerade bei sehr großen Stichproben ein Problem. Dies liegt daran, dass die p-Werte c.p. mit steigendem N sinken. Das heißt, jeder Effekt, egal wie bedeutungslos (klein) er ist, wird bei hohen N irgendwann statistisch signifikant. Ein inhaltlich bedeutungsloser Mini-Effekt wird allerdings auch durch große Stichproben nie zu einem wichtigen Effekt (Wasserstein/Lazar 2016).

Lösungsansatz: Um sprachlich bedingte Fehlinterpretationen zu vermeiden, schlagen Armstrong (2007), Colquhoun (2014) und Motulsky (2014) vor, in wissenschaftlichen Veröffentlichungen ganz auf das Wort „signifikant“ zu verzichten. Angesichts der langen Tradition des Begriffs ist fraglich, ob dies durchsetzbar ist. Praktikabler ist möglicherweise, dieses Problem in der Lehre intensiv anzusprechen und Nachwuchswissenschaftler systematisch aufzufordern, (i) das Wort „signifikant“ nicht als Synonym für „groß/wichtig“ zu gebrauchen und (ii) es immer mit dem Zusatz „statistisch“ zu verwenden, wenn die Gefahr eines Missverständnisses besteht (Mittag/Thompson 2000). Auf Journalebene könnten die Gutachter explizit aufgefordert werden, auf diese Fehlinterpretation zu achten und problematische Formulierungen zu korrigieren. In Verbindung mit der Verpflichtung (z. B. in den Richtlinien der Journale), die Effektstärke zu diskutieren wäre dies ein Schritt vorwärts, der nach Goodman (2008) mit wenig Aufwand erreichbar wäre. In den Richtlinien könnte auch – wie bereits von der American Psychological Association empfohlen (APA 2010) – die Verwendung von Konfidenzintervallen gefordert werden, wenn man es mit aussagekräftigen Maßeinheiten zu tun hat. Durch die Angabe des Konfidenzintervalls wird die Bandbreite der Effektgröße in einer leicht nachvollziehbaren Form kommuniziert, ohne dass man auf die Signifikanzinformation verzichten muss.

2.2 Fehlschlüsse bei Überschreiten des Signifikanzniveaus

Problembeschreibung: Im Rahmen einer Regression werden üblicherweise die Koeffizienten der Regressoren auf Signifikanz überprüft. Bei einer linearen Regression $Y = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \ldots + \beta_n \cdot X_n$ bezeichnet der p-Wert eines β-Koeffizienten bspw. die Wahrscheinlichkeit, den durch den Koeffizienten ausgedrückten Einfluss (oder einen noch größeren) als Zufallsbefund zu finden,

Bereitgestellt von | De Gruyter / TCS
Angemeldet
Heruntergeladen am | 07.03.17 15:30

wenn er gar nicht da wäre. Die Einhaltung eines Signifikanzniveaus von 0,05 wird häufig als Kriterium für die Ablehnung der Nullhypothese akzeptiert. Wenn man dem folgt, stellt sich die Frage, wie *p*-Werte über 5% (nicht statistisch signifikante Ergebnisse) zu interpretieren sind. Auch hier kommt es gelegentlich zu Denkfehlern, bei denen sprachliche Formulierungen eine Rolle spielen. Bei einem *p*-Wert oberhalb des Signifikanzniveaus von 0,05 lautet die inhaltlich richtige und sprachlich unmissverständliche Formulierung wie folgt:

> Die Nullhypothese, dass der Regressor X_1 keinen Einfluss auf Y hat, kann nicht mit dem üblicherweise geforderten Signifikanzniveau von maximal 0,05 abgelehnt werden.

Diese Formulierung entspricht dem „Satz vom ausgeschlossenen Dritten" (*tertium non datur*), nach dem eine Aussage so formuliert sein muss, dass entweder sie selbst oder ihre Verneinung zutrifft. Die Aussage „Hans ist entweder blond oder nicht blond" ist demzufolge richtig. Die Aussage „Hans ist entweder blond oder schwarzhaarig" (oder analog: „Wenn Hans nicht blond ist, ist er schwarzhaarig") ist dagegen eine Verletzung des Satzes vom ausgeschlossenen Dritten. Es wird eine falsche Dichotomie hergestellt, die außer Acht lässt, dass es etwas Drittes geben kann, nämlich dass Hans eine andere Haarfarbe als blond oder schwarz hat. Ein analoger Trugschluss droht bei der Interpretation von *p*-Werten oberhalb des akzeptierten Signifikanzniveaus, wenn laxe, aber weit verbreitete Formulierungen wie die folgenden genutzt werden:

> Der Einfluss von X auf Y ist **n**icht **s**tatistisch **s**ignifikant. [**n**-**s**-**s**]
> Der Einfluss von X auf Y ist **s**tatistisch **n**icht **s**ignifikant. [**s**-**n**-**s**]
> Der Einfluss von X auf Y ist **n**icht **s**ignifikant. [**n**-**s**]

Von der letzten Formulierung aus, die bereits nahelegt, dass man gefunden habe, dass kein (relevanter) Effekt da ist, ist es nur ein kurzer Weg zur falschen Schlussfolgerung:

> Unsere Studie zeigt, dass ein (relevanter) Einfluss von X auf Y **n**icht vorhanden ist. [**n**]

Richtig war, dass man die Nullhypothese (kein Effekt) nicht mit dem üblicherweise geforderten Signifikanzniveau von 0,05 ablehnen konnte. Falsch ist jedoch die Schlussfolgerung, dass man die Nullhypothese bestätigt habe (Sedlmeier/Gigerenzer 1989; Wasserstein/Lazar 2016). Die Gefahr dieses Fehlschlusses entsteht, wenn sprachliche Formulierungen, was man aus *p*-Werten unter und über 0,05 ableiten kann, die falsche Dichotomie „*entweder* Ablehnung der Nullhypothese *oder* Annahme der Nullhypothese" nahelegen. Nicht statistisch signifikante Ergebnisse als Bestätigung der Nullhypothese zu interpretieren, ist ein Trugschluss, der sich auch in Formulierungen findet, bei denen er nicht auf den

Bereitgestellt von | De Gruyter / TCS
Angemeldet
Heruntergeladen am | 07.03.17 15:30

ersten Blick offensichtlich ist. So werden nicht-signifikante Ergebnisse gelegentlich dahingehend kommentiert, dass sie „im Gegensatz zu theoretischen Vorhersagen stehen", die die Existenz des Effekts nahelegen. Das ist eine unzulässige Schlussfolgerung. Man könnte das nur dann sagen, wenn man *p*-Werte oberhalb des Signifikanzniveaus als Bestätigung der Nullhypothese ansehen könnte.[8]

Lösungsansatz: Da der beschriebene Denkfehler auf der Grundlage der Logik einfach zu durchdringen ist, sollten Studierende und Nachwuchswissenschafter gezielt mit dem Satz vom ausgeschlossenen Dritten vertraut gemacht werden. Auf Journalebene sollten die Gutachter konsequent alle sprachlichen Formulierungen beanstanden, die bei $p > 0{,}05$ den Trugschluss „Bestätigung der Nullhypothese" nahelegen. Da dieser Trugschluss einfach identifiziert werden kann, ist ein strenger Maßstab im wissenschaftlichen Begutachtungsprozess leicht sicherzustellen. Außerhalb der Wissenschaft können falsche Dichotomien allerdings ein virulentes Problem sein. Wenn es um die Rezeption von Forschungsergebnissen durch die interessierte Öffentlichkeit oder um Politikberatung geht, ist den Nutzern (Fachjournalisten, Politikern) oft schwer zu vermitteln, dass „*kein signifikanter Effekt*" nicht bedeutet, dass man einen Hinweis dafür gefunden hat, dass kein Effekt (oder nur ein unbedeutender Effekt) da ist. Das besondere Problem bei der öffentlichen Rezeption von Forschungsergebnissen besteht möglicherweise darin, dass im Kampf um die öffentliche Wahrnehmung eine interessant klingende (wenngleich falsche) Meldung „*X hat keinen Einfluss auf Y!*" vielen Beteiligten lieber ist als die „langweilige" Meldung, dass man keine Aussage machen kann. Insbesondere wenn Studienergebnisse im Rahmen wichtiger öffentlicher Debatten falsch rezipiert werden, müssen sich Forscher aber immer wieder der Aufgabe stellen, vorschnelle Interpretationen richtig zu stellen.

2.3 *p-hacking*

Problembeschreibung: Es gibt bei statistischen Analysen eine Vielzahl von „Gestaltungsmöglichkeiten", die zu Verzerrungen und letztlich falschen Schlussfolgerungen führen können.[9] Werden derartige Eingriffe von Seiten der

8 Goodman (2008: 137) greift das Problem explizit auf und bezeichnet die Annahme, dass „studies with P values on opposite sides of .05 are conflicting", als ein zentrales Missverständnis bei der Interpretation des *p*-Wertes.

9 Ioannidis (2005: 0696) greift das Problem unter dem provokanten Titel „*Why Most Published Research Findings are False*" auf und warnt vor Verzerrungen: „[...], for many current scientific fields, claimed research findings may often be simply accurate measures of the prevailing bias."

Bereitgestellt von | De Gruyter / TCS
Angemeldet
Heruntergeladen am | 07.03.17 15:30

Forscher gezielt vorgenommen, um „veröffentlichungsfähige“ *p*-Werte zu erzielen, liegt *p-hacking* vor.[10] Da es sich bei der Auswahl von Daten und Analysemethoden im Forschungsprozess selten um eindeutige Entscheidungen handelt, ist *p-hacking* schwer zu identifizieren. Eine inhaltlich begründete Auswahl des informativsten Datenmaterials und der adäquatesten Analysemethode für die untersuchte Fragestellung stellt kein *p-hacking* dar. Eine nicht transparent kommunizierte selektive Auswahl und Darstellung derjenigen Analysevariante, die im Sinne der Produktion geringer *p*-Werte im Vergleich zu anderen Varianten am besten „funktioniert“, ist dagegen *p-hacking*. Für Außenstehende ist es deshalb schwierig zu unterscheiden, ob eine bestimmte Herangehensweise Verzerrungen verhindert oder geradezu hervorruft. So ist es sinnvoll, offensichtlich unsinnige Werte (z. B. ein PKW-Kraftstoffverbrauch von 95 l/100 km) zu korrigieren oder den betreffenden Datensatz mangels Zuverlässigkeit ganz aus dem Datenset zu entfernen. Wenn man dagegen gezielt die 10 % niedrigsten und/oder höchsten Beobachtungswerte aus dem Datenset herausnimmt und dann schaut, ob sich so rechnerisch ein statistisch signifikantes Ergebnis herstellen lässt, handelt es sich um *p-hacking*. Die gezielte Suche nach Auswertungsmethoden, die zu den gewünschten statistisch signifikanten Ergebnissen führen, wird von vielen Forschern gar nicht als Problem wahrgenommen, obwohl jede selektive Darstellung einer Auswertung, die „funktioniert“, der Produktion einer Verzerrung entspricht (Simmons et al. 2011). Abbildung 1 vermittelt einen Eindruck, welche „Gestaltungsmöglichkeiten“ bei statistischen Analysen im Rahmen von *p-hacking* genutzt werden können.

a) *Nicht sachlich begründete Verringerung der Stichprobengröße*: Bei der Stichprobengröße gibt es zwei Ansatzpunkte für *p-hacking*. Erstens kann man, wie bereits angesprochen, die Stichprobe um „Ausreißer“ bereinigen und herumprobieren, wie sich die *p*-Werte dadurch ändern.[11] Zweitens kann man insbesondere bei hohen *N* ausprobieren, wie sich getrennte Analysen

10 Diese Begriffsbildung wird Simmons et al. (2011: 1359) zugeschrieben: „[...] it is unacceptably easy to publish ‘statistically significant’ evidence consistent with *any* hypothesis. The culprit is a construct we refer to as *researcher degrees of freedom*. 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 %.“

11 In SPSS wird bspw. eine über die Cook-Distanz ausgeführte „Ausreißerbereinigung“ unter dem Stichwort „Automatische lineare Modellierung“ angeboten. Die Tatsache, dass derartige Vorgehensweisen in gängigen Programmroutinen standardmäßig angeboten werden, erhöht die Gefahr, dass wenig versierte Nutzer solche Tools unkritisch als richtige Standardvorgehensweise zur Verringerung von *p*-Werten wahrnehmen.

Bereitgestellt von | De Gruyter / TCS
Angemeldet
Heruntergeladen am | 07.03.17 15:30

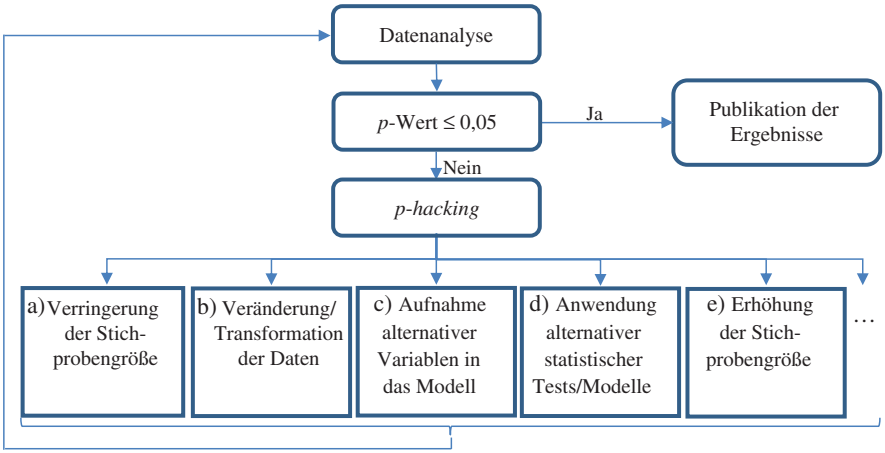

Abbildung 1: Verschiedene Möglichkeiten des *p-hacking*.
Quelle: Eigene Abbildung in Anlehnung an Motulsky (2014: Abbildung 1).

von Datenteilmengen auf die *p*-Werte auswirken. Analysiert man 20 Teilgruppen getrennt, so ist geradezu zu erwarten, dass auch ein nicht-vorhandener Effekt rein zufallsbedingt einmal als signifikant ausgewiesen wird. Wenn man genau dieses Ergebnis selektiv ausweist, hat man einen schweren Fall von *p-hacking*.[12]

b) *Nicht sachlich begründete Transformation der Daten*: Auch wenn der Stichprobenumfang feststeht, kann man *p-hacking* betreiben und ausprobieren, ob sich die *p*-Werte verringern, wenn man die Daten in irgendeiner Weise transformiert. Dazu zählen die Reduzierung des Skalenniveaus (z. B. Einkommensklassen statt Einkommen) und die Bildung neuer Variablen bspw. in Form von Relativgrößen (z. B. Gewicht geteilt durch Größe). Grundsätzlich gilt auch in diesem Zusammenhang: Jede dieser Maßnahmen kann inhaltlich begründet sein. Die gezielte Suche nach und der selektive Ausweis derjenigen Variante, die in gewünschtem Umfang signifikante Ergebnisse produziert, führt aber zu einer Überbewertung der empirischen Evidenz.

12 Kerr (1998) nennt das Problem HARKing (Hypothesizing After the Results are Known) und Motulsky (2014: 201) führt dazu aus: „This is when you analyze the data many different ways (say different subgroups), discover an intriguing relationship, and then publish the data so it appears that the hypothesis was stated before the data were collected [...]. This is a form of multiple comparisons (Berry, 2007). Kriegeskorte and colleagues (2009) call this double dipping, as you are using the same data both to generate a hypothesis and to test it.“

Bereitgestellt von | De Gruyter / TCS
Angemeldet
Heruntergeladen am | 07.03.17 15:30

c) *Nicht sachlich begründete Aufnahme/Entfernung von Variablen*: Auch bei der Auswahl der Variablen, die im Schätzmodell berücksichtigt werden, gibt es „Gestaltungsmöglichkeiten". Dies gilt zunächst für die Kontrollvariablen, deren Anzahl und Art man verändern kann, um zu sehen, mit welchem Variablenset man die gewünschten statistischen Signifikanzen am besten erzielt. Die Möglichkeit, das Variablenset zu erweitern/reduzieren oder gewisse Variablen durch andere zu substituieren, hat man aber auch bei den manifesten Variablen, mit denen im Rahmen einer hypothesengeleiteten Vorgehensweise die latenten Variablen (theoretischen Konstrukte) operationalisiert werden. Man stelle sich bspw. vor, man wolle überprüfen, ob die Einstellung gegenüber der ökologischen Landwirtschaft die Zahlungsbereitschaft für Öko-Produkte beeinflusst. Die Einstellung (=latente Variable) sei in einer Befragung durch verschiedene Fragen (Items) erhoben worden. Wenn man solange probiert, bis man ein Item zur Einstellung gefunden hat, mit der sich im Modell ein signifikantes Ergebnis ergibt, hat man eine Verzerrung.[13] Das wird aber nicht offensichtlich, wenn man aus Vermarktungsgründen nur die Auswertung publiziert, mit der ein signifikantes Ergebnis produziert werden konnte.

d) *Nicht sachlich begründete Anwendung statistischer Tests und Schätzmodelle*: Auch beim Testen von Verteilungsannahmen und der Entscheidung für ein Schätzmodell gibt es Gestaltungsspielräume, die im Rahmen des *p-hacking* missbraucht werden können. Man stelle sich vor, es sei vorab nicht eindeutig, ob man eine einfache OLS-Schätzung oder ein Paneldatenmodell nutzen soll. *p-hacking* wird daraus, wenn man beide Schätzmodelle ausprobiert und dann selektiv dasjenige darstellt, bei dem die gewünschten Signifikanzen am besten herauskommen. Die Transparenz im wissenschaftlichen Kommunikationsprozess geht verloren, wenn man die verschiedenen Modelle nicht explizit gegenübergestellt und vergleichend diskutiert. Anders gesagt: *p-hacking* bei der Auswahl von Tests und Schätzmodellen führt ebenfalls zu einer Verzerrung und einer Inflationierung der empirischen Evidenz.

e) *Nicht sachlich begründete Erhöhung der Stichprobengröße*: Ähnlich wie die Suche nach einem anderen Modell, das „funktioniert", wirkt sich die Suche nach einem größeren Datenset aus, das vielleicht „funktioniert", wenn das ursprüngliche Datenset keine „zufriedenstellenden" Signifikanzen erbracht hat. Angenommen, man habe in einem ökonomischen Experiment mit der

13 Eine *p*-Wert orientierte Auswahl von Variablen wird in vielen gängigen Programmroutinen angeboten (z. B. *Forward and backward effect selection* in SAS). Diese Funktion verringert möglicherweise den selbstkritischen Umgang mit der Frage, welche Variablen *inhaltlich* begründet in der Analyse genutzt werden sollten.

Bereitgestellt von | De Gruyter / TCS
Angemeldet
Heruntergeladen am | 07.03.17 15:30

ursprünglichen Stichprobe der Größe *N* keine statistisch signifikanten Ergebnisse gefunden. Häufig wird es in einem solchen Fall als unproblematisch wahrgenommen, die Stichprobe ad hoc zu vergrößern und dann ggf. die in der vergrößerten Stichprobe erzielten signifikanten Ergebnisse zu publizieren (Motulsky 2014). Das Problem ist, dass man auch hierdurch eine Verzerrung produziert und die empirische Evidenz überbewertet, da man nur nacherhebt, wenn man im ursprünglichen Datenset keine Signifikanzen gefunden hat.

Grundsätzlich ist bei der Bewertung des *p-hacking* Problems zu fragen, ob man es mit einer explorativen Studie zur Generierung von Hypothesen oder mit einer konfirmatorischen Studie zur Überprüfung von Hypothesen zu tun hat. Ein explorativ-entdeckendes Suchen nach korrelativen Zusammenhängen, die die Generierung von Hypothesen ermöglichen sollen, ist ein sinnvoller und unabdingbarer Schritt im Forschungsprozess. *p-hacking* kann damit als Problem verstanden werden, das sich ergibt, wenn explorative und konfirmatorische Datenanalyse nicht klar voneinander unterschieden werden. Eine Überprüfung von Hypothesen ist zwingend mit neuen Daten durchzuführen; und die Verwendung von Begriffen wie „Hypothesen*test*“ und „statistisch signifikant“ sollte bei explorativen Ansätzen, die ja nur erste Hinweise für die *Bildung* von Hypothesen liefern können, vermieden werden (Gigerenzer/Marewski 2015). Dies stände im Einklang mit Fishers Diktum, dass ein geringer *p*-Wert eigentlich nur „worthy a second look“ bedeutet (Nuzzo 2014: 151) und an*zeigen* soll, ob sich weitere Studien lohnen.

Lösungsansatz: Der Umgang mit *p-hacking* ist schwierig, da es nicht um Denkfehler geht, sondern um einen sorglosen Umgang mit guter wissenschaftlicher Praxis oder wissenschaftliches Fehlverhalten im Forschungsprozess selbst. Neben der Sensibilisierung von Studierenden und Nachwuchsforschern werden verschiedene Forderungen für die Veröffentlichungspraxis diskutiert. (i) Eine erste Forderung ist, dass bei jeder Untersuchung explizit gesagt werden sollte, ob es sich um eine explorativ-entdeckende Studie zur Identifizierung von korrelativen Zusammenhängen und zur *Generierung* von Hypothesen oder um eine Studie zur *Überprüfung* von Hypothesen handelt. Beides darf nicht vermischt werden (Marino 2014; Motulsky 2014). Eine weitergehende Forderung wäre, bei jeder Studie eine interne Replikation mit neuen Daten zu fordern. (ii) Eine zweite Forderung ist, auf Journalebene nicht nur alle Rohdaten zugänglich zu machen, sondern eine genaue und transparente Dokumentation und Veröffentlichung aller Arbeitsschritte (inkl. der Umkodierung und Transformation der Daten) zu hinterlegen (Simmons et al. 2011) oder sogar vor Durchführung einer Studie das Forschungsdesign und das gesamte Datenmaterial registrieren zu lassen. Dieser Vorschlag wirft die Frage auf, wer sich die Zeit nehmen wird oder kann, das hinterlegte Material durchzusehen

Bereitgestellt von | De Gruyter / TCS
Angemeldet
Heruntergeladen am | 07.03.17 15:30

und zu kontrollieren. (iii) Eine dritte Forderung geht dahin, von den Autoren eine explizite *no-p-hacking* Erklärung zu fordern (Simmons et al. 2012). Davon erhofft man sich eine Verstärkung des Normappells der guten wissenschaftlichen Praxis. Das Problem dabei ist, dass es zwar extreme Herangehensweisen gibt, die eindeutig *p-hacking* darstellen, es sich bei der Auswahl von Analysemethoden aber oft nicht um inhaltlich zwingende Entscheidungen handelt. Deshalb ist es schwer, einen Katalog „indexierter" Herangehensweisen zu erstellen. Angesichts des systembedingten Veröffentlichungsdrucks ist zudem fraglich, ob ein verstärkter Normappell ausreicht, um das Problem zu beheben. (iv) Eine vierte Forderung knüpft an die allgemeine Diskussion zu Verzerrungen in der wissenschaftlichen Veröffentlichungspraxis an und verlangt, dass Beiträgen mit negativen Ergebnissen und Replikationsstudien ein höherer wissenschaftlicher Stellenwert eingeräumt und eine Chance zur Publikation gegeben werden soll.[14]

2.4 Gleichsetzung der „Irrtumswahrscheinlichkeit" mit der false discovery rate

Problembeschreibung: Ein weiterer Grund für die Replikationskrise wird darin gesehen, dass der *p*-Wert *als solcher* häufig falsch interpretiert und als

14 Ausgehend von dem Problem der Verzerrung wissenschaftlicher Publikationen zugunsten positiver Ergebnisse geben Duvendack et al. (2015) einen historischen Abriss der Bemühungen um verbesserte Replikation auf der Ebene ökonomischer Journale. Sie beschreiben zudem die Replikationspolitik aller 333 Web-of-Science Ökonomiejournale und kommen zum Schluss, dass Replikation immer noch einen geringen Stellenwert hat. So publizieren nur 27 Journale regelmäßig (d. h., bei mehr als 50 % der empirischen Studien) die Daten und Auswertungscodes und nur zehn Journale schaffen Anreize für Autoren, indem sie auf ihrer Website erwähnen, dass sie Replikationsstudien publizieren. Die *Jahrbücher für Nationalökonomie und Statistik* gehören in beiden Fällen dazu. Die Häufigkeit von Replikationsstudien ist bislang aber gering und institutionalisierte Bemühungen zur Förderung von Replikationen stehen u. W. in den Wirtschaftswissenschaften noch am Anfang. Eine erwähnenswerte Initiative in Deutschland ist das Projekt „Replication in Economics" in Göttingen, das nicht nur eine Datenbasis für Replikationen bereitstellt, sondern auch die Möglichkeit bietet, die Ergebnisse als „Replication Working Papers" des Zentrums für Statistik der Universität Göttingen zu publizieren. Außerhalb der Ökonomie scheinen diese Bemühungen stärker verankert zu sein. So wurde im Bereich der medizinischen Forschung im Jahr 2013 eine globale Initiative „All Trials Registered/All Results Reported" gestartet, die sehr breite Unterstützung von Wissenschaftlern und Wissenschaftsorganisationen gefunden hat. In manchen Bereichen gibt es sogar eigenständige Journals, 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 (daher der Name) zu veröffentlichen.

Bereitgestellt von | De Gruyter / TCS
Angemeldet
Heruntergeladen am | 07.03.17 15:30

Wahrscheinlichkeit der Nullhypothese verstanden wird.[15] Anders gesagt: Neben der Verwechslung von „signifikant“ und „wichtig“ gibt es ein weiteres semantisches Missverständnis, das durch die Konvention ausgelöst wird, den *p*-Wert als „Irrtumswahrscheinlichkeit“ oder „Wahrscheinlichkeit eines Fehlers 1. Art“ zu bezeichnen. Trotz dieser Benennung ist der *p*-Wert *nicht* die hier als *false discovery rate* bezeichnete Wahrscheinlichkeit, bei Ablehnung der Nullhypothese einen Irrtum in Form eines Fehlers 1. Art zu begehen. Der *p*-Wert wird vielmehr als bedingte Wahrscheinlichkeit berechnet unter der *Annahme*, dass die Nullhypothese zutrifft. Eine Schlussfolgerung über die *Wahrscheinlichkeit* der Nullhypothese kann man aus dem *p*-Wert nicht ziehen (Kline 2013; Nuzzo 2014). Deshalb ist auch die Aussage, dass *p*-Werte genutzt werden können, um Hypothesen zu testen, nur bedingt richtig. Trotz des Begriffs „Hypothesentest“ testet man mit *p*-Werten nicht Hypothesen, sondern *p*-Werte weisen aus, wie kompatibel die Daten mit dem durch die Nullhypothese spezifizierten statistischen Modell sind (Wasserstein/Lazar 2016).

Der Sachverhalt lässt sich an einem Münzwurfbeispiel illustrieren, bei dem man vorab mit 1 %iger Wahrscheinlichkeit eine manipulierte Münze [$P(Kopf) = 0{,}75$] und mit 99 %iger Wahrscheinlichkeit eine nicht-manipulierte Münze [$P(Kopf) = 0{,}5$] zieht. Nun wirft man die Münze fünfmal und beobachtet fünfmal Kopf. Wenn die Münze ideal wäre (=kein Effekt), wäre bei vielen Wiederholungen des Experiments „fünfmaliger Münzwurf“ nur in 3,125 % ($=0{,}5^5$) der Fälle fünfmal Kopf zu erwarten. Diese bedingte Wahrscheinlichkeit, die auch als Falsch-positiv-Rate bezeichnet wird, entspricht dem *p*-Wert. Es ist aber nicht die Wahrscheinlichkeit der Nullhypothese „ideale Münze“ und damit auch nicht die Wahrscheinlichkeit, bei Verwerfung der Nullhypothese einen Fehler zu machen. Hierfür muss man zusätzlich wissen, wie hoch bei der manipulierten Münze die Wahrscheinlichkeit für fünfmal Kopf ist.

15 Schon vor über 25 Jahren berichten Sedlmeier und Gigerenzer (1989: 314) von einer Studie, die die Verbreitung dieses Missverständnisses belegt: „Oakes (1986) tested 70 academic psychologists and reported that 96 % held the erroneous opinion that the level of significance specified the probability that either H_0 or H_1 was true.“ Auch dieses Missverständnis scheint „nachhaltig“ zu sein. Cohen (1994) nutzt dafür die treffende Bezeichnung „inverse probability error“ und Nickerson (2000: 251) schreibt: „The belief that p is the probability that the null hypothesis is true is unquestionably false. However, as Berger and Sellke (1987) have pointed out ‘like it or not, people do hypothesis testing to obtain evidence as to whether or not the hypotheses are true, and it is hard to fault the vast majority of nonspecialists [...]. This is especially so since we know of no elementary textbooks that teach that $p = 0.05$ is at best very weak evidence against H_0’ (p. 114).“ Krämer (2011) weist darauf hin, dass diese Fehlinterpretation sogar explizit in weit verbreiteten Statistiklehrbüchern reproduziert wird, und auch im Jahr 2014 fasst Motulsky (2014: 204) die Lage wie folgt zusammen: „The 5 % significance threshold is often misunderstood. [...] Many scientists mistakenly believe that the chance of making a false-positive conclusion is 5 %. In fact, in many situations, the chance of making a type I false-positive conclusion is much higher than 5 %.“

Bereitgestellt von | De Gruyter / TCS
Angemeldet
Heruntergeladen am | 07.03.17 15:30

Diese auch als Richtig-positiv-Rate oder Power bezeichnete Wahrscheinlichkeit beträgt 23,73 % (=$0{,}75^5$). Man muss zudem die auch als „Prior" bezeichneten A-priori-Wahrscheinlichkeiten 1 % und 99 % berücksichtigen, dass man anfangs eine manipulierte bzw. eine ideale Münze gezogen hatte. Nach dem Satz von Bayes kommt man auf die auch als A-posteriori-Wahrscheinlichkeit bezeichnete *false discovery rate* von 92,88 % [$= 0{,}03125 \cdot 0{,}99/(0{,}03125 \cdot 0{,}99 + 0{,}237 \cdot 0{,}01)$], einen Irrtum zu begehen, wenn man die Nullhypothese „ideale Münze" verwirft. Trotz des geringen p-Werts wird man also die Nullhypothese nicht verwerfen. Der Informationsgehalt der durch das Wurfexperiment gewonnenen Daten führt lediglich dazu, dass man die A-priori-Wahrscheinlichkeit von 99 % revidiert und a posteriori (d. h. *nach* Auswertung der experimentellen Daten) nur noch mit 92,88 %iger Wahrscheinlichkeit davon ausgeht, dass man es mit einer idealen Münze zu tun hat.

Zur Vermeidung von Missverständnissen, die verständlicherweise durch die irreführende, aber geläufige Bezeichnung des p-Werts als „Irrtumswahrscheinlichkeit" oder als „Wahrscheinlichkeit eines Fehlers 1. Art" entstehen können, ist Folgendes festzuhalten:

1. Trotz ihres Namens gibt die „Irrtumswahrscheinlichkeit" *nicht* die Wahrscheinlichkeit an, einen Irrtum zu begehen, wenn man die Nullhypothese ablehnt. Anders gesagt: Der p-Wert, obwohl er auch als „Wahrscheinlichkeit eines Fehlers 1. Art" bezeichnet wird, ist *nicht* die eigentlich interessierende *false discovery rate*, d. h. die sich nach der Analyse ergebende A-posteriori-Wahrscheinlichkeit, bei Verwerfung der Null einen Fehler 1. Art zu machen.
2. Für die Bestimmung der *false discovery rate* benötigt man neben dem auch als Falsch-positiv-Rate bezeichneten p-Wert die Richtig-positiv-Rate oder Power $1-\beta$ einer konkreten Alternativhypothese H_A; und für die Alternativhypothese H_A und die Nullhypothese H_0 braucht man Wahrscheinlichkeitsinformationen von außerhalb der Stichprobe in Form der Priors $P(H_A) = 1 - P(H_0)$. Ohne Power und Prior ist die Bestimmung der *false discovery rate* grundsätzlich *nicht* möglich (Motulsky 2014).[16]

16 Wie bei der häufig mit α bezeichneten Obergrenze des p-Werts von 0,05 wird auch für die Power $1-\beta$ bei expliziten Hypothesentests ein Grenzwert vorgeschlagen. Dieser wird in der Literatur im Allgemeinen mit 0,8 angegeben (List et al. 2011; Colquhoun 2014). Das entspricht der Forderung, dass die Falsch-negativ-Rate β nicht mehr als viermal so hoch sein soll wie die Falsch-positiv-Rate α. Unterstellt man im Satz von Bayes die Werte $p = \alpha = 0{,}05$, $\beta = 0{,}2$ sowie Laplace Priors $P(H_A) = P(H_0) = 0{,}5$, so kommt man zu einer *false discovery rate* von 5,9 %. Durch die fälschliche Gleichsetzung des p-Werts mit der *false discovery rate* macht man also bei dieser Konstellation keinen großen Fehler. Liegt keinerlei Vorwissen vor (*flat prior*) fallen die beiden Wahrscheinlichkeiten zusammen. Wohl deshalb spekuliert Nickerson (2000: 289): „Perhaps it is

Bereitgestellt von | De Gruyter / TCS
Angemeldet
Heruntergeladen am | 07.03.17 15:30

Wenn man versteht, dass p-Werte lediglich ausweisen, wie kompatibel die Daten mit dem durch die Nullhypothese spezifizierten statistischen Modell sind, versteht man auch, dass die Praxis, die Einhaltung von $p \leq 0,05$ als Bedingung für die Ablehnung der Nullhypothese anzusehen, eine verbreitete Konvention darstellt, aber nicht mit der Einhaltung einer einheitlichen und akzeptablen Obergrenze für die *false discovery rate* im Einklang steht. Wie das Wurfbeispiel zeigt, können auch geringe p-Werte mit inakzeptabel hohen *false discovery rates* zusammenfallen. Obwohl p-Werte nicht die Wahrscheinlichkeit der Nullhypothese angeben, kann aber die qualitative Aussage gemacht werden, dass mit geringeren p-Werten c.p. auch die Wahrscheinlichkeit der Nullhypothese sinkt.

Die Forderung, statistische Inferenzen grundsätzlich mit dem Satz von Bayes zu überprüfen, kann kritisch gesehen werden. Zwar müssen bei einer Bayesschen Analyse die Priors explizit benannt und damit transparent gemacht werden. Dennoch eröffnet sich ein weites Feld für subjektive Bewertungen von Seiten des Forschers (Simmons et al. 2011). Es gilt ja: Je geringer man a priori (vor der Analyse) die Wahrscheinlichkeit $P(H_A)$ einschätzt, desto höher ist a posteriori (nach der Analyse) die Wahrscheinlichkeit, bei Ablehnung der Nullhypothese H_0 einen Irrtum zu begehen. Man könnte daher versucht sein, zu kritisieren, dass man mit der subjektiven Wahl des Priors die Ablehnung/Nicht-Ablehnung der Nullhypothese willkürlich steuern kann. Es ist zwar zunächst richtig, dass c.p. ein geringerer Prior $P(H_A)$ zu einer Erhöhung der *false discovery rate* führt. Eine derartige isolierte Änderung des Priors ist aber von der Sache her nicht plausibel. Der Prior muss den a priori vorhandenen wissenschaftlichen Kenntnisstand widerspiegeln. Mit einer Verringerung des Priors $P(H_A)$ muss deshalb auch die Spezifizierung einer unwahrscheinlicheren Alternativhypothese H_A einhergehen, die ihrerseits die Power $1-\beta$ erhöht. Eine erhöhte Power wiederum verringert *die false discovery rate*. Bei konsistenter Festlegung von Prior und Alternativhypothese gibt es also einen gegenläufigen Effekt.

Interessanterweise scheint es nicht zu der Gleichsetzung des p-Werts mit der *false discovery rate* zu kommen, wenn andere Forscher signifikante Ergebnisse finden, die im Widerspruch zum bisherigen Kenntnisstand stehen. In solchen Fällen wird der Satz von Bayes anscheinend qualitativ „mitgedacht“. Das bedeutet, dass man auch einer kleinen Irrtumswahrscheinlichkeit (z. B. $p = 0{,}01$) kein Vertrauen schenkt und fundiertes Vorwissen nicht aufgrund einer einzelnen Studie über den Haufen wirft, auch wenn in der betreffenden Studie die Daten und Analysemethoden einer kritischen Überprüfung standhalten. Man

the case, as Rindskopf (1997) has argued, ‘that null hypothesis tests are still used because they are testing approximately the right thing under many real circumstances, even though most researchers do not know the rationale’ (p. 321).“

Bereitgestellt von | De Gruyter / TCS
Angemeldet
Heruntergeladen am | 07.03.17 15:30

interpretiert die Irrtumswahrscheinlichkeit in diesem Fall also nicht als Wahrscheinlichkeit der Nullhypothese, sondern fragt im Sinne des Satz von Bayes lediglich danach, ob bzw. wie stark man seine fundierte A-priori-Wahrscheinlichkeit im Lichte eines einzelnen Studienergebnisses updaten soll. Formal würde das der folgenden Argumentation entsprechen: Es gibt fundierte wissenschaftliche Vorkenntnisse, die einer A-priori-Wahrscheinlichkeit für die Nullhypothese von fast 100 % entsprechen. Nach dem Satz von Bayes ergibt sich in dieser Konstellation auch bei einem hochsignifikanten Ergebnis eine A-posteriori-Wahrscheinlichkeit der Nullhypothese von annähernd 100 %. Es ist deshalb auch formal richtig, die Nullhypothese bei einem geringen p-Wert nicht gleich zu verwerfen.

Lösungsansatz: Der Gleichsetzung der „Irrtumswahrscheinlichkeit" mit der „*false discovery rate*" ist schwer zu begegnen. Einerseits verführt der Begriff „Irrtumswahrscheinlichkeit" sprachlich bedingt jede neue Generation von Forschern zu der Fehlinterpretation, dass der p-Wert die Wahrscheinlichkeit bezeichne, bei Ablehnung der Nullhypothese einen Irrtum zu begehen. Andererseits haben viele Menschen Schwierigkeiten mit dem Verständnis von (bedingten) Wahrscheinlichkeiten (Gigerenzer 2002). In der Lehre muss dem abgeholfen werden, indem durch anschauliche Beispiele das Verständnis geschärft wird, was bedingte Wahrscheinlichkeiten und damit p-Werte aussagen und was nicht. Was in der wissenschaftlichen Veröffentlichungspraxis verändert werden soll, ist dagegen weniger offensichtlich. Ein zentrales Argument dafür, alles beim Alten zu belassen, ist die bequeme Natur der konventionellen Interpretation des p-Wertes, die sowohl Forschern als auch Gutachtern eine (scheinbar) klare Leitlinie zur Beurteilung der Validität von Ergebnissen verschafft. Bei ausschließlichem Rückgriff auf den p-Wert, der ja alleine auf der Grundlage der jeweiligen Stichprobe berechnet wird, abstrahiert man allerdings von jeglichem Vorwissen. Wissenschaft und Erkenntnisfortschritt beruhen aber auf Vorarbeiten, die mit den eigenen Ergebnissen kontrastiert und zusammengeführt werden müssen – ein Sachverhalt, der prägnant durch die Metapher von den „Zwergen auf den Schultern von Riesen" ausgedrückt wird. Dementsprechend werden die meisten Forscher Ergebnisse, die nicht statistisch signifikant sind, aber bereits etablierte Erkenntnisse reproduzieren, als weitere kleine Bestätigung werten. Dagegen werden sie statistisch signifikante, aber überraschende Ergebnisse skeptisch beurteilen, auch wenn die Daten- und Modellwahl adäquat und gut begründet erscheint. Dies lässt sich als gesundes wissenschaftliches Misstrauen im Sinne einer impliziten Anwendung des Satzes von Bayes interpretieren. Bei konsequenter Formalisierung ließe sich dieses Misstrauen mit dem Satz von Bayes begründen, durch den man zu einer Einschätzung der *false discovery rate* kommen kann. Mit Blick auf

Bereitgestellt von | De Gruyter / TCS
Angemeldet
Heruntergeladen am | 07.03.17 15:30

die Akzeptanz von Veränderungsvorschlägen wird deshalb gelegentlich eine stufenweise Kombination der herkömmlichen, auf p-Werten basierenden Vorgehensweise und einer umfassenderen, auf dem Satz von Bayes beruhenden Validierung der Ergebnisse vorgeschlagen (Nickerson 2000: 290–291). Man will also das eine tun, ohne das andere zu lassen. Die Beibehaltung der gewohnten Vorgehensweise würde den Forschern entgegenkommen, die ein Unwohlsein gegenüber einer unvermeidbar subjektiven Festlegung von Priors haben. Eine zusätzliche formale Bayes-Argumentation hätte den Vorteil, dass sowohl eine zu geringe als auch eine übertriebene Skepsis gegenüber neuen Forschungsergebnissen intersubjektiv nachvollziehbar wird, da die zugrunde gelegten Priors transparent gemacht werden müssen. Man könnte auch Variantenrechnungen bzgl. der Priors durchführen und wüsste, worüber man zu streiten hat (Zyphur/Oswald 2015). Neben deutlich erhöhten methodischen Anforderungen ist mit Blick auf die Praktikabilität allerdings unklar, wie eine Darstellung beider Ansätze z. B. bei einer multiplen Regression aussehen könnte, ohne das Platzangebot von Journalen zu sprengen.

3 Rückblick und Ausblick

Als bemerkenswerter Fakt ist festzuhalten, dass auf die Fehlinterpretationen des p-Wertes sowie auswertungsbedingte Verzerrungsprobleme über die Jahrzehnte hinweg oft hingewiesen wurde, sich aber anscheinend in vielen Disziplinen wenig an den gängigen Fehlinterpretationen und der üblichen Praxis verändert hat. Dies liegt zum einen an der zum Teil disparaten, über die Disziplinen verstreuten, unübersichtlichen und oft auf Einzelaspekte abzielenden Literatur zur p-Wert-Problematik. Zum anderen erfordert die Einarbeitung in diese nicht-triviale Thematik einen hohen Zeitaufwand und man kann sich den Worten von Nickerson (2000: 290) anschließen: „The situation is not simple–it is confused and confusing–and a nonsuperficial understanding of the issues requires a considerable investment of time and effort.“

Auch wir, die Verfasser dieses Beitrags, sind in der Vergangenheit in unterschiedlichem Maße den angesprochenen Missverständnissen unterlegen, bevor wir uns vor dem Hintergrund der offensichtlichen Probleme näher mit dem Thema auseinandergesetzt haben. Inhaltlich teilen wir die Einschätzung von Nickerson (2000: 262), der sich bereits vor mehr als 15 Jahren wie folgt äußerte: „I believe that much of the confusion [...] about what p values mean derives from [...] ambiguities in casual language, some of which can be quite subtle.“ Vor diesem Hintergrund hoffen wir, dass der vorliegende Methodenkommentar Wirtschaftswissenschaftlern die verschiedenen Aspekte

Bereitgestellt von | De Gruyter / TCS
Angemeldet
Heruntergeladen am | 07.03.17 15:30

des Problems systematisch, komprimiert und zeitsparend näher bringt, eine Hilfestellung in der Lehre und Doktorandenausbildung ist und letztlich zu einer Reduzierung der *false discovery rate* beiträgt.

Literatur

APA (American Psychological Association) (2010), Publication Manual of the American Psychological Association. Washington, American Psychological Association (6. Auflage).

Armstrong, J.S. (2007), Significance Tests Harm Progress in Forecasting. International Journal of Forecasting 23 (2): 321–327.

Berger, J.O., T. Sellke (1987), Testing a Point Null Hypothesis: The Irreconcilability of *P* values and Evidence. Journal of the American Statistical Association 82 (397): 112–122.

Berry, D.A. (2007), The Difficult and Ubiquitous Problems of Multiplicities. Pharmaceutical Statistics 6 (3): 155–160.

Cohen, J. (1994), The Earth is Round ($p<0.05$). American Psychologist 49 (12): 997–1003.

Colquhoun, D. (2014), An Investigation of the False Discovery Rate and the Misinterpretation of *p*-values. Royal Society Open Science. Erhältlich unter: http://dx.doi.org/10.1098/rsos.140216.

Cumming, G. (2012), Understanding The New Statistics. Effect Sizes, Confidence Intervals, and Meta-Analysis. New York, Routledge.

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

Gigerenzer, G. (2002), Das Einmaleins der Skepsis. Über den richtigen Umgang mit Zahlen und Risiken. Berlin, Berlin Verlag.

Gigerenzer, G., J.N. Marewski (2015), Surrogate Science: The Idol of a Universal Method for Statistical Inference. Journal of Management 41 (2): 421–440.

Goodman, S. (2008), A Dirty Dozen: Twelve *p*-value Misconceptions. Seminars in Hematology 45: 135–140.

Ioannidis, J.P.A. (2005), Why Most Published Research Findings are False. PLoS Medicine 2 (8): e124: 0696–0701.

Kerr, N.L. (1998), HARKing: Hypothesizing After the Results are Known. Personality and Social Psychology Review 2 (3): 196–217.

Kirk, R.E. (1996), Practical Significance: A Concept Whose Time has Come. Educational and Psychological Measurement 56 (5): 746–759.

Kline, R.B. (2013), Beyond Significance Testing: Statistics Reform in the Behavioral Sciences. Washington, American Psychological Association (2.Auflage).

Krämer, W. (2011), The Cult of Statistical Significance – What Economists Should and Should Not Do to Make their Data Talk. Schmollers Jahrbuch 131 (3): 455–468.

Kriegeskorte, N., W.K. Simmons, P.S.F. Bellgowan, C.I., Baker (2009), Circular Analysis in Systems Neuroscience: The Dangers of Double Dipping. Nature Neuroscience 12 (5): 535–540.

List, J.A., S. Sadoff, M. Wagner (2011), So You Want to Run an Experiment, Now What? Some Simple Rules of Thumb for Optimal Experimental Design. Experimental Economics 14 (4): 439–457.

Bereitgestellt von | De Gruyter / TCS
Angemeldet
Heruntergeladen am | 07.03.17 15:30

Marino, M.J. (2014), The Use and Misuse of Statistical Methodologies in Pharmacology Research. Biochemical Pharmacology 87 (1): 78–92.
McCloskey, D.N., S.T. Ziliak (1996), The Standard Error of Regressions. Journal of Economic Literature 34 (1): 97–114.
Mittag, K.C., B. Thompson (2000), A National Survey of AERA Members' Perceptions of Statistical Significance Tests and Other Statistical Issues. Educational Researcher 29 (4): 14–20.
Motulsky, H.J. (2014), Common Misconceptions about Data Analysis and Statistics. The Journal of Pharmacology and Experimental Theurapeutics 351 (8): 200–205.
Motulsky, H. (2016), Essential Biostatistics. A Nonmathematical Approach. Oxford, Oxford University Press.
Nickerson, R.S. (2000), Null Hypothesis Significance Testing: A Review of an Old and Continuing Controversy. Psychological Methods 5 (2): 241–301.
Nuzzo, R. (2014), Statistical Errors. *P*-values, the 'Gold Standard' of Statistical Validity, Are Not as Reliable as Many Scientists Assume. Nature 506 (7487): 150–152.
Oakes, M. (1986), Statistical Inference: A Commentary for the Social and Behavioural Sciences. New York, Wiley.
Rindskopf, D.M. (1997), Testing "Small," Not Null, Hypotheses: Classical and Bayesian Approaches. S. 319–332 in: L.L. Harlow, S.A. Mulaik, J.H. Steiger (Hrsg.), What If There Were No Significance Tests? Hillsdale, Erlbaum.
Sedlmeier, P., G. Gigerenzer (1989), Do Studies of Statistical Power Have an Effect on the Power of Studies? Psychological Bulletin 105 (2): 309–316.
Sellke, T., M.J. Bayarri, J.O. Berger (2001), Calibration of *p*-Values for Testing Precise Null Hypotheses. The American Statistician 55 (1): 61–71.
Siegfried, T. (2014), To Make Science Better, Watch Out for Statistical Flaws. Science News.
Simmons, J.P., L.D. Nelson, U. Simonsohn (2011), False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. Psychological Science 22 (11): 1359–1366.
Simmons, J.P., L.D. Nelson, U. Simonsohn (2012), A 21 Word Solution. Dialogue. The Official Newsletter of the Society for Personality and Social Psychology 26 (2): 4–7.
Sterne, J., D. Smith (2001), Sifting the Evidence – What's Wrong with significance tests? British Medical Journal 322: 226–231.
Storey, J.D. (2011), False Discovery Rates. S. 504–508 in: M.Lovric (Hrsg.): International Encyclopedia of Statistical Science. Berlin und Heidelberg, Springer.
Trafimow, D., M. Marks (2015), Editorial. Basic and Applied Social Psychology 37: 1f.
Wasserstein, R.L., N.A. Lazar (2016), The ASA's Statement on *p*-values: Context, Process, and Purpose. The American Statistician 70 (2): 129–133.
Ziliak, S.T., D.N. McCloskey (2008), The Cult of Statistical Significance. How the Standard Error Costs Us Jobs, Justice, and Lives. Michigan, The University of Michigan Press.
Zyphur, M.J., F.L. Oswald (2015), Bayesian Estimation and Inference: A User's Guide. Journal of Management 41 (2): 390–420.

Bereitgestellt von | De Gruyter / TCS
Angemeldet
Heruntergeladen am | 07.03.17 15:30