P(H|D) versus $P(D|H_0)$? Wie man das Testen von Hypothesen – lieber doch nicht – einführen sollte

RAPHAEL DIEPGEN, BOCHUM

Zusammenfassung: *Der Autor kritisiert den Vorschlag von Krauss und Wassner, den weitverbreiteten Missverständnissen über den Signifikanztest durch einen Unterricht vorzubeugen, der vor allem das Bayessche Konzept P(H|D) dem (angeblich) signifikanztestspezifischen Konzept $P(D|H_0)$ gegenüberstellt.*

1 Einleitung

Die Debatte über Sinn und Unsinn des Signifikanztestes insbesondere in den empiri(sti)schen Humanwissenschaften ist seit Jahrzehnten ein publikations- und karrierefördernder Dauerbrenner. Beliebt sind in diesem Rahmen immer wieder auch Untersuchungen mit dem – überaus leicht zu erreichenden – Ziel nachzuweisen, dass nicht nur die den Signifikanztest anwendenden Wissenschaftler und ihre Studenten, sondern auch die den Signifikanztest lehrenden Dozenten die mathematische Logik dieses Testens nicht richtig verstanden haben (Haller & Krauss, 2002). Diese Debatte nährt so manchen Kritiker – und zwar, das ist das Schöne, auf Dauer: Sie bleibt nämlich für die statistische Praxis in den genannten Wissenschaften völlig folgenlos und kann daher alle paar Jahre wieder neu gestartet werden. Folgenlos bleibt diese Debatte, weil sie in naiver Weise so tut, als ergäbe sich der Sinn des Signifikanztestens aus seinem mathematischen Rationale, wie es etwa zunächst von Fisher, dann von Neyman und Pearson entwickelt wurde. Tatsächlich aber dürften der „Sinn" und die Funktion des Signifikanztestens – ziemlich unabhängig von dieser Logik – in diesen Wissenschaften ganz anders zu fassen sein, etwa: die reputationsfördernde Verkleidung als mathematisierte Naturwissenschaft, die Entlastung von Unsicherheit durch Ritualisierung, die Ersatzbeschäftigung mit Methoden statt mit relevanten Inhalten in Ermangelung derselben, die Bereitstellung objektiver Beurteilungskriterien für wissenschaftliche Qualifikationsarbeiten, die Verkleinbürgerlichung durch Senkung des ursprünglich elitär-bildungsbürgerlichen auf ein mathematisiert-technisches Sprachniveau. Und so weiter und so fort. Diesen „Sinn", diese Funktionen kann das Signifikanztesten aber vermutlich überhaupt nur erfüllen, wenn seine mathematische Logik – und damit seine überaus bescheidene Begrenzung – nicht richtig verstanden wird. Die große Beliebtheit des Signifikanztestens bei den empirischen Humanwissenschaftlern – überaus brüchige didaktische Legitimation für das Thema Hypothesentesten auf der Schule – ist zum großen Teil überhaupt nur zu verstehen unter der skeptischen Annahme: „Denn sie wissen nicht, was sie tun." Dass gerade der diesen „mächtigen" Wissenschaften hinterherlaufende „schwache" Stochastikunterricht auf der Schule in der Lage sein soll, durch Aufklärung über die mathematische Logik des Signifikanztests diesen seines tatsächlichen „Sinns" und seiner Funktionen für diese Wissenschaften zu berauben, dies allerdings erscheint ziemlich zweifelhaft.[1]

Von solchen Zweifeln unbekümmert unterbreiten nun Krauss und Wassner (2001) einen Vorschlag für die Einführung des Hypothesentestens in der Schule, der den verbreiteten Fehlinterpretationen statistischer Signifikanz – beispielsweise in der Psychologie – Einhalt gebieten soll.

2 Der Vorschlag von Krauss und Wassner (2001)

Krauss und Wassner (2001) gehen davon aus, dass die verbreiteten Missverständnisse[2] über die Bedeutung eines signifikanten Befundes im Wesentlichen auf dem Irrtum beruhen, ein Signifikanztest liefere Aussagen über die Wahrscheinlichkeit von Hypo-

[1] Hier war auch ich zugegebenermaßen in meinen von Krauss und Wassner zitierten Beiträgen aus den achtziger Jahren etwas optimistischer. Diesen Optimismus habe ich längst verloren, sowohl aufgrund meiner Erfahrungen in den empirischen Humanwissenschaften, als auch aufgrund meiner Erfahrungen mit dem Mathematikunterricht in Deutschland: Ich habe das Vermögen beider Bereiche unterschätzt, auch an den sinnlosesten Dingen über Jahrzehnte festzuhalten.

[2] Untersucht wurden von Krauss und Wassner (2001) die folgenden sechs Missinterpretationen eines auf dem 1%-Niveau signifikanten Befundes: 1) Es ist eindeutig bewiesen, dass die Nullhypothese falsch ist. 2) Die Wahrscheinlichkeit des Zutreffens der Nullhypothese ist gefunden worden. 3) Es ist eindeutig bewiesen, dass die Alternativhypothese wahr ist. 4) Man kann nun die Wahrscheinlichkeit ableiten, dass die Alternativhypothese richtig ist. 5) Entscheidet man sich nun, die Nullhypothese zu verwerfen, dann kennt man jetzt die Wahrscheinlichkeit dafür, dass diese Entscheidung falsch ist. 6) Wenn man das Experiment sehr oft wiederholen würde, würde man in 99% der Fälle ein signifikantes Ergebnis bekommen.

thesen. Da solche Wahrscheinlichkeiten von Hypothesen aber überhaupt nur im Rahmen der Bayes-Statistik behandelt werden, sei es sinnvoll, Bayes-Statistik und Signifikanztests kontrastierend einzuführen. Dabei sei freilich erstens anders als üblich die grundlegende Formel von Bayes nicht für irgendwelche Ereignisse A und B zu erarbeiten, sondern für eine Hypothese H und Daten D, also

$$P(H \mid D) = \frac{P(D \mid H) \cdot P(H)}{P(D \mid H) \cdot P(H) + P(D \mid \overline{H}) \cdot P(\overline{H})},$$

und zweitens auch bei der Einführung des Signifikanztestes anders als üblich die Notation mit bedingten Wahrscheinlichkeiten zu benutzen. „Der didaktische Kniff zur besseren Durchdringung der Signifikanztests besteht nun darin, beide Testverfahren (klassisches und Bayessches) in Abhängigkeit von Daten und Hypothesen als bedingte Wahrscheinlichkeiten auszudrücken und vergleichend gegenüberzustellen:

(a) P(H|D) liegt Bayesschen Testverfahren zugrunde

(b) $P(D|H_0)$ liegt Signifikanztests zugrunde“

(S. 32). Im Falle (b) stehe dabei D für die „vorliegenden oder noch unwahrscheinlichere Daten“ (S. 30). Der Vergleich von (a) und (b) zeige dabei dem Schüler von Anfang an, „dass es zwei grundsätzlich verschiedene Arten Beurteilender Statistik gibt: Man kann die Wahrscheinlichkeit der Hypothese bei vorgegebenen Daten berechnen oder man kann die Wahrscheinlichkeit der Daten bei gegebener Nullhypothese betrachten.“ (S. 32) Schließlich solle man im Unterricht den Signifikanztest nicht mehr in einer Entscheidung bezogen auf ein vorgewähltes Signifikanzniveau münden lassen, sondern lediglich in die Berechnung des „informativeren“ p-Wertes, also der oben unter (b) genannten Wahrscheinlichkeit $P(D|H_0)$.

3 Kritik

Weder die Idee, die Formel von Bayes nicht auf irgendwelche Ereignisse, sondern auf Hypothesen und Daten zu beziehen, noch der Vorschlag, auch bei Signifikanztests bedingte Wahrscheinlichkeiten zu notieren, ist neu; dies wird selbstverständlich bereits seit vielen Jahren in weitverbreiteten Schulbüchern so praktiziert (z.B. bei Diepgen u.a., 1993). Auch die Gegenüberstellung klassischer und bayesscher Statistik findet sich in Schulbüchern (Diepgen u.a., 1993, S. 212-216), sogar schon auf der Sekundarstufe I (Lauter u.a., 1995, S. 221). Problematisch erscheint es aber, den Kontrast zwischen diesen beiden Denkschulen in der genannten Weise auf (a) P(H|D) versus (b) $P(D|H_0)$ zuzuspitzen. Warum?

1. Zunächst etwas Grundsätzliches: Likelihoods P(D|H), also die Wahrscheinlichkeiten für die beobachteten Daten D unter der Bedingung – oder klassisch: der Annahme – der Geltung von Hypothesen H spielen unbeschadet ggf. unterschiedlicher Begrifflichkeiten und Notationen in beiden Arten von Statistik gleichermaßen eine wichtige Rolle; sie unterscheiden damit insbesondere nicht den klassischen Signifikanztest von der Bayes-Statistik. Im Fundamentallemma von Neyman und Pearson wird der klassische („optimale“) Test zweier (einfacher) Hypothesen gegeneinander durch einen kritischen Wert für das Verhältnis der Likelihoods der konkurrierenden Hypothesen charakterisiert, ebenso wie auch die Formel von Bayes auf den Likelihoods der konkurrierenden Hypothesen beruht. Der Unterschied zwischen klassischer und Bayesscher Statistik wird also gerade nicht dadurch markiert, dass man nur in der einen Statistik nach den Wahrscheinlichkeiten von Daten unter Hypothesen fragte, in der anderen aber nicht – ganz im Gegenteil. Der Unterschied wird vielmehr dadurch markiert, dass man über die gemeinsame Berücksichtigung von Likelihoods hinaus nur in der Bayesstatistik auch Hypothesen Wahrscheinlichkeiten zuordnet, dies freilich notgedrungen schon a priori, nicht erst a posteriori, während man dies in der klassischen Statistik eben nicht tut (und sich deshalb dort schwer tut, Hypothesen überhaupt als Ereignisse zu interpretieren und ihnen die Rolle von „Bedingungen“ bei bedingten Wahrscheinlichkeiten zuzubilligen). Vor diesem theoretischen Background wirkt die kontrastierende Charakterisierung von (a) Bayes durch P(H|D) versus (b) Signifikanztests durch $P(D|H_0)$ seltsam schief. Auf dieser theorieorientierten Folie müsste es allenfalls vielmehr heißen: (a) bei Bayes sowohl P(H) als auch P(H|D) (für alle konkurrierenden Hypothesen) versus (b) beim Signifikanztest weder $P(H_0)$ noch $P(H_0|D)$.

2. Irritiert dies zunächst nur den theoretisch-statistisch vorgebildeten Fachmann, so dürfte auch den Schüler verwirren, dass bei dieser suggestiven Kontrastierung derselbe Platzhalter D zwei völlig verschiedene Bedeutungen hat: Bei (a) bedeutet D das tatsächlich beobachtete Datum, bei (b) aber die Vereinigung des tatsächlich beobachteten mit allen „noch extremeren“, von dem unter der Nullhypothese zu Erwartenden „noch weiter entfernten“ möglichen Daten[3]. Auch wenn man diese formale

[3] Ich benutze hier andere Formulierungen als Krauss und Wassner (2001). Wenn diese nämlich D bei (b) für die „vorliegenden oder noch unwahrscheinlichere Daten“ setzen, bedeutet dies eine nicht ganz unproblematische

Verwirrung durch Verwendung verschiedener Buchstaben auflöst – und damit freilich dem Kontrast (a) versus (b) seine vermeintliche Prägnanz nimmt –, dürfte es den nachdenklichen Schüler ziemlich ratlos machen, dass, nachdem man bei Bayes sinnvollerweise nur die bedingten Wahrscheinlichkeiten der tatsächlich beobachteten Daten verrechnet hat, nun beim Signifikanztest auch die bedingten Wahrscheinlichkeiten von Daten eine Rolle spielen sollen, die tatsächlich überhaupt nicht beobachtet worden, sondern nur möglich sind.[4] Es ist eben nicht so, wie Krauss und Wassner suggerieren, dass man bei Bayes „die Wahrscheinlichkeit der Hypothese bei vorgegebenen Daten" berechnet und beim Signifikanztest „die Wahrscheinlichkeit der Daten bei gegebener Nullhypothese" betrachtet. Wäre dies so, erschöpfte sich der Signifikanztest lediglich in der Berechnung und Bewertung der Likelihood einer isolierten Hypothese – ohne Rücksicht darauf, welche Wahrscheinlichkeiten die beobachteten Daten unter anderen Hypothesen haben. Der Schüler könnte auf dieser Folie dann beispielsweise kaum begründen, warum man, wenn man beim Binomialtest der Nullhypothese H_0: $p=0{,}5$ über die Wahrscheinlichkeit p eines Merkmals in der Zufallsstichprobe vom Umfang $n=1000$ genau 500 Merkmalsträger gefunden hat, H_0 nicht verwerfen sollte, obwohl doch die Wahrscheinlichkeit für dieses Datum unter H_0 – also die Likelihood von H_0 – verschwindend gering ist.

3. Hier wird deutlich: Der Vorschlag von Krauss und Wassner orientiert sich leider – genau so wie die von ihnen eigentlich kritisierte inferenzstatistische Praxis, also das dem Neyman-Pearsonschen „Über-Ich" nicht genügende „Ich" des Praktikers (Gigerenzer, 1993a)[5] – nur an der rudimentären Signifikanztestkonzeption von Fisher[6] mit all ihren Begründungslücken und Rationalitätsdefiziten. Dieser Vorschlag ignoriert die präzisierende Weiterentwicklung in der mathematischen Statistik von Neyman und Pearson, durch die überhaupt erst diese Lücken und Defizite begrifflich gefasst und behoben werden können. Für diese mathematische Statistik ist ein Hypothesentest eben nicht wie bei

---

Verengung. Diese Formulierung mag vielleicht für die ursprüngliche Signifikanztestkonzeption von Fisher passend sein – sofern man bei „noch unwahrscheinlichere" die Bedingung oder Annahme „unter H_0" ergänzt –, sicherlich aber nicht für die modernere Konzeption von Neyman und Pearson: Dass man beim klassischen Neyman-Pearson-Testen (für eindimensionale Prüfgrößen) die Ablehnungsbereiche als Randbereiche konstruiert, hat seinen Grund zunächst nicht darin, dass die datenbasierten (eindimensionalen) Prüfgrößenwerte unter der Nullhypothese in streng monotoner Weise immer unwahrscheinlicher werden, je extremer sie werden, sondern darin, dass man bei einer anderen Konstruktion der Ablehnungsbereiche zu große Wahrscheinlichkeiten erhielte für eine falsche Entscheidung zugunsten der Nullhypothese für „wahre" Werte, die sehr weit von dem nullhypothetischen Wert entfernt sind. Kurzum: Nicht, dass die potentiellen Daten (unter der Nullhypothese) noch unwahrscheinlicher sind als das beobachtete Datum, ist in dieser Konzeption der eigentliche Grund dafür, ihre Wahrscheinlichkeiten zu summieren, sondern zunächst nur, dass sie – grob formuliert – „noch weiter" von der Nullhypothese „entfernt" sind als das beobachtete Datum. Es kommt hier sozusagen zunächst nur darauf an, dass die potentiellen Daten extremer, nicht aber, dass sie unwahrscheinlicher als das beobachtete Datum sind.

[4] Tatsächlich verbirgt sich hier eine für die Fishersche Konzeption gravierende Problematik. So macht der Fishersche Begriff der Überschreitungswahrscheinlichkeit („p-Wert") ja überhaupt nur bei einer eindimensionalen Prüfverteilung Sinn, also dann, wenn sich die in den Daten enthaltene „relevante" Information auf eine eindimensionale Prüfstatistik reduzieren lässt (Stichwort „Suffizienz") – was, anders als von den meisten empirischen Humanwissenschaftlern (und vermutlich auch Mathematiklehrern) geglaubt, eher die Ausnahme denn die Regel sein dürfte. Es ist also im Allgemeinen überhaupt nicht klar, welche Likelihoods sinnvollerweise zu der „Überschreitungswahrscheinlichkeit" zu summieren sind, auf der das ganze Konzept beruht.

[5] Von Gigerenzer stammt die reizvolle Interpretation verhaltenswissenschaftlicher Inferenzstatistik in „psychoanalytischen" Begriffen: Es gibt da einerseits zunächst das Bayessche Es, insbesondere also das natürliche Bedürfnis der Wissenschaftler, so etwas wie die Wahrscheinlichkeit ihrer Hypothesen zu kennen und durch Daten steigern zu können. Auf der anderen Seite steht das Neyman-Pearsonsche Über-Ich, nämlich die rational-mathematische Konzeption einer Entscheidungsregel mit kontrollierten Fehlerwahrscheinlichkeiten zweier Art. Dazwischen laviert das Fishersche Ich, das in mechanisierter Weise ex post p-Werte berechnet und mit den konventionellen Signifikanzniveaus von 1% oder 5% vergleicht, mündend in einer dichotomen Ja-Nein-Entscheidung über die Hypothesen. Das Neyman-Pearsonsche Über-Ich dominiert die formale Statistikausbildung, die moderneren Lehrbücher, vor allem die Rhetorik der methodischen Sonntagsreden, mit denen man sich nach außen präsentiert. Das Bayessche Es bestimmt die inneren Wünsche und Phantasien und führt dann beispielsweise dazu, dass man signifikante Ergebnisse – wie von Krauss und Wassner moniert – im Bayesschen Sinne als Aussagen über Hypothesenwahrscheinlichkeiten missinterpretiert. Das Fishersche Ich schließlich bestimmt die weitgehend automatisierte alltägliche statistische Routine im Wissenschaftsbetrieb, heute weitestgehend dem Computer überlassen.

[6] Fisher steht hier nur als Label für die Auffassung, es sei sinnvoll, eine skeptische Nullhypothese aufzugeben dann, wenn die Überschreitungswahrscheinlichkeit beobachteter Daten unter dieser Nullhypothese hinreichend gering sei. Ob und inwieweit diese Fassung Sir Ronald Aylmer Fisher gerecht wird, sei dahingestellt.

Krauss und Wassner eine Bewertung einer isolierten Nullhypothese nach der Datenerhebung anhand eines p-Wertes, sondern ein vor der Datenerhebung definiertes datengestütztes Entscheidungsverfahren zur Entscheidung zwischen zwei konkurrierenden Hypothesen, welches die (bedingte) Wahrscheinlichkeit einer (fälschlichen) Entscheidung für die Alternativhypothese, falls die Nullhypothese gilt, durch eine vorgewählte Schranke α limitiert (und unter dieser Nebenbedingung die Fehlerwahrscheinlichkeiten zweiter Art in einem bestimmten Sinne minimiert). Es ist schwer vorstellbar, dass man die statistische Praxis rationaler machen könnte ausgerechnet dadurch, dass man den Schülern diese – jedenfalls im Vergleich – rationalen Konzepte der mathematischen Statistik im Unterricht vorenthält und ihnen nur eine alte Konzeption des Hypothesentestens präsentiert, an deren Rationalitätslücken sich die mathematische Statistik längst abgearbeitet hat.

4. Der Schüler, für den der Signifikanztest nur eine ex post-Bewertung einer isolierten Nullhypothese anhand ihres p-Wertes im Sinne Fishers bleibt – wie von Krauss und Wassner vorgeschlagen –, dem fehlt die Begrifflichkeit, um rational mit folgenden Fragen umzugehen: Wie groß sollte die Stichprobe sein? Wie groß das Signifikanzniveau? Welche Unterschiede von der Nullhypothese werden mit welcher Wahrscheinlichkeit entdeckt? Wie sollte man vernünftigerweise verfahren, wenn man eine einzige globale Nullhypothese nur über mehrere Tests verschiedener Teilnullhypothesen prüfen kann? Wie begegnet man also dem Problem der Alpha-Fehler-Inflation? Und so weiter. Wer aber in der Schule nicht einmal die Begriffe lernt, mit denen man überhaupt nur sinnvoll über diese Fragen reden kann – und das sind die Grundbegriffe der Neyman-Pearson-Statistik wie Entscheidungsverfahren, Fehler erster und zweiter Art und ihre gegenläufigen und ggf. vom Stichprobenumfang abhängigen (bedingten) Wahrscheinlichkeiten –, der hat keine Chance, mit diesen für jede ernsthafte und nicht nur ritualisierte inferenzstatistische Praxis wichtigen Fragen umzugehen. Es steht daher zu befürchten, dass der Vorschlag von Krauss und Wassner die von ihnen beklagte Irrationalität der inferenzstatistischen Praxis eher zu befördern denn einzudämmen geeignet ist. Und es dürfte dann in Zukunft nicht schwer sein nachzuweisen, dass die nach dem Vorschlag von Krauss und Wassner ausgebildeten Schüler (und Wissenschaftler) zwar nicht mehr den von Krauss und Wassner untersuchten formalen Missverständnissen über das Hypothesentesten unterliegen, dafür aber anderen[7] – und zwar solchen, die nicht nur für die nachträgliche Bewertung statistischer Ergebnisse Bedeutung haben, sondern auch für die Planung des statistischen Vorgehens selbst.

5. Ein Letztes: Wenn man im Unterricht Bayessche Statistik dem Signifikanztest gegenüberstellen will, dann macht dies kaum Sinn, solange man sich bei der Formel von Bayes – wie Krauss und Wassner – auf den Fall zweier einfacher Hypothesen beschränkt, wie dies leider zumeist auch in der Schule geschieht. Denn dann ist Bayesstatistik überhaupt keine Alternative zum üblichen Signifikanztest, bei dem typischerweise eine einfache Nullhypothese einer *zusammengesetzten* Alternativhypothese gegenübersteht, also einer unendlich großen Menge von dicht liegenden punktförmigen Alternativen. Die Idee von Krauss und Wassner kann aber überhaupt nur greifen, wenn man Bayes-Statistik auch für diesen Fall konzipiert. Dies wäre auf der Schule formaler Schwierigkeiten wegen kaum machbar – man müsste mit stetigen Wahrscheinlichkeitsverteilungen, also Dichten arbeiten –, wenn man diese dichte Unendlichkeit „exakt" in infinitesimaler Begrifflichkeit repräsentieren wollte. Tatsächlich aber kann man sich dieser Problematik entledigen, wenn man den Hypothesenraum der unendlich vielen dicht liegenden Alternativen auf endlich viele vergröbert (wie es dann ja auch letzten Endes die Bayes-Statistik in der Praxis macht, wenn es um numerische Berechnungen im Computer geht.) Die Erweiterung der Formel von Bayes von zwei auf endliche viele Hypothesen ist auch auf der Schule einfach. Dann etwa könnte man im Unterricht vergleichen (a) die Berechnung der a posteriori-Wahrscheinlichkeiten der vergröbert elf Hypothesen H_{00}: $p=0,0$, H_{01}: $p=0,1$, H_{02}: $p=0,2$, ..., H_{10}: $p=1,0$ über eine Merkmalswahrscheinlichkeit p nach Beobachtung eines Datums D, beispielsweise „65 Merkmalsträger in der Zufallsstichprobe von n=100", auf der Basis von a priori-Wahrscheinlichkeiten mit (b) der Durchführung eines einseitigen Signifikanztest der Hypothese H_0: $p=0,5$.

[7] Etwa: 1) Ein signifikanter Befund weist auf relevante, also hinreichend große Abweichungen der Realität von der Nullhypothese hin. 2) Es kommt nur auf die Signifikanz eines Befundes an, nicht aber auf die zugrundeliegende Stichprobengröße. 3) Behauptet eine Nullhypothese H_0: a=b=c die Gleichheit von drei Populationsparametern a, b und c, dann ist diese Nullhypothese auf dem Niveau α zu verwerfen, sobald sich beim Testen der beiden Einzelhypothese H_{01}: a=b und H_{02}: b=c mindestens ein p-Wert kleiner α ergibt.

4 Vorschlag

Wenn denn der schulische Stochastikunterricht auch den von Krauss und Wassner kritisierten Missinterpretationen durch Aufklärung vorbeugen will, dann sollte er zunächst vor allem anderen den prinzipiell probabilistischen Charakter des Testens betonen. Denn gerade über diesen prinzipiell unsicheren Charakter täuscht die „Mechanisierung" statistischer Methoden (Gigerenzer, 1993b) hinweg. Dagegen hilft aber sicherlich weniger eine formal-abstrakte Gegenüberstellung von P(H|D) versus $P(D|H_0)$; das Problem dürfte nämlich weniger sein, dass in den Humanwissenschaften diese beiden Wahrscheinlichkeiten verwechselt werden (und dass dies so furchtbar schlimm wäre), sondern viel mehr, dass man dort vergessen oder verdrängt hat, dass es sich überhaupt in beiden Fällen um Wahrscheinlichkeiten handelt. Nun weiß jeder Lehrer – auch ohne die angekündigten Forschungsergebnisse von Krauss und Wassner aus einem einschlägigen DFG-Schwerpunktprogramme abzuwarten –, dass sich (bedingte) Wahrscheinlichkeiten „verständnisfördernd" durch Häufigkeiten repräsentieren oder „erleben" lassen. Für die Fehlerwahrscheinlichkeiten beim Signifikanztest – notfalls auch für die von Krauss und Wassner favorisierten p-Werte – könnte man dies durch Simulation vieler Tests bei geltender Nullhypothese bewerkstelligen: Die Schüler erleben dann, dass im Falle der Geltung der Nullhypothese auf lange Sicht die relative Häufigkeit signifikanter Befunde gleich dem Signifikanzniveau α ist, insbesondere also, dass gelegentlich Fehlentscheidungen fallen. Obwohl Krauss und Wassner (2001, S. 33) seltsamerweise[8] den didaktischen Wert von Häufigkeitsinterpretationen gerade für den Unterricht über die Formel von Bayes preisen, gelingt diese Interpretation für die Bayes-Statistik nicht: Hier müsste man schließlich mit der a priori-Verteilung entsprechenden relativen Häufigkeiten verschiedene hypothetische Zustände der Welt simulieren – eine auch für den Schüler wohl wenig überzeugende Simulation, insofern dieser wohl nicht zu Unrecht genau einen – leider unbekannten – „wahren" Zustand der Welt voraussetzen dürfte.

Besser aber noch als die demonstrierende fiktive Simulation mehrerer Tests, von denen man in der Forschungsrealität später dann doch immer nur einen durchführt, wäre es, wenn der Schüler schon bei der Durchführung eines einzelnen Tests dessen probabilistischen Charakter erlebte. Dies wäre aber durchaus möglich, wenn man auf der Schule nicht das klassische Hypothesentesten für feste Stichprobengrößen einführte, sondern stattdessen das sequentielle Hypothesentesten, organisiert in Form eines graphischen „random walk", bei dessen Durchlaufen jeder Schüler unmittelbar erlebt, dass es sich beim Hypothesentesten offensichtlich prinzipiell um einen zufallsabhängigen, also – wie auch das Publizieren stochastikdidaktischer Artikel – fehlerrisikobehafteten Vorgang handelt (Diepgen, 1987).

[8] Seltsamerweise, weil sich doch gerade die Bayesianer ausdrücklich auf einen nichtfrequentistischen Wahrscheinlichkeitsbegriff berufen.

Literatur

Diepgen, R. (1987): Sequentielles Testen – auch didaktisch vielleicht eine gute Alternative. Stochastik in der Schule 7 (2), 9-25

Diepgen, R., Kuypers, W. & Rüdiger, K. (1993): Mathematik. Sekundarstufe II. Stochastik. Berlin: Cornelsen

Gigerenzer, G. (1993a): The Superego, the ego, and the id in statistical reasoning. In: Keren, G. & Lewis, C. (Hrsg.): A handbook for data analysis in the behavioral sciences: Methodological issues. Hillsdale, NJ: Erlbaum

Gigerenzer, G. (1993b): Über den mechanischen Umgang mit statistischen Methoden. In: Roth, E. (Hrsg.): Sozialwissenschaftliche Methoden. München: Oldenbourg

Haller, H. & Krauss, S. (2002): Misinterpretations of Significance: A Problem Students Share with their Teachers? Methods of Psychological Research Online 7 (1) (http://www.mpr-online.de)

Krauss, S. & Wassner, C. (2001): Wie man das Testen von Hypothesen einführen sollte. Stochastik in der Schule 21 (1), 29-34

Lauter, J., Bielig-Schulz, G., Diepgen, R., Jahnke, T., Kuypers, W. & Wuttke, H. (1995): Mathematik. 10. Schuljahr. Berlin: Cornelsen

Autor

Dr. Raphael Diepgen
Ruhr-Universität Bochum
Fakultät für Psychologie
44780 Bochum
E-Mail:
raphael.diepgen@ruhr-uni-bochum.de