Unklare Begriffe und Wunschdenken bei Signifikanztests

MATTHIAS MOSSBURGER

Zusammenfassung: *Einige der bisher üblichen Erklärungen von Grundbegriffen sind unzureichend oder sogar irreführend. Dadurch kann sich entgegen der Logik leicht der Wunsch durchsetzen, dass Signifikanztests fundierte Aussagen über Hypothesen liefern. Ich schlage eine Sprache vor, in der man meines Erachtens gut erkennen kann, was Signifikanztests leisten, und was nicht.* [1]

1 Empirische Ergebnisse

Die Verbreitung von Fehlvorstellungen bei Signifikanztests wurde bereits 1986 von Oakes, und 2001 von Haller, Krauss und Wassner anhand von Fragebögen untersucht, die man Psychologiestudenten und Dozenten vorlegte. Diese Fragebögen habe ich an die Schule angepasst und Schülern der 12. Jahrgangsstufe vorgelegt:

Stellen Sie sich vor, Sie testen eine Nullhypothese H_0 anhand einer Stichprobe. Es stellt sich heraus, dass das Ergebnis der Stichprobe im Ablehnungsbereich von H_0 liegt, und zwar auf einem Signifikanzniveau von 5% (Fehler 1. Art). Welche der folgenden Aussagen lassen sich aus diesem Ergebnis folgern? Es können dabei alle, oder nur 5, …, oder vielleicht auch keine der Aussagen zutreffen.

1. *Es ist bewiesen, dass H_0 falsch ist.*
2. *Die Wahrscheinlichkeit, dass H_0 gilt, beträgt höchstens 5%.*
3. *Es ist bewiesen, dass die Gegenhypothese zu H_0 gilt.*
4. *Man kann die Wahrscheinlichkeit berechnen, dass die Gegenhypothese zu H_0 gilt.*
5. *Entscheidet man sich nun, H_0 abzulehnen, dann beträgt die Wahrscheinlichkeit höchstens 5%, dass man sich dabei irrt.*
6. *Wenn man H_0 sehr oft testet, dann erhält man in etwa 5% der Fälle ein signifikantes Ergebnis.*

Untersuchungen von Haller, Krauss und Wassner ergaben: 86% der befragten Psychologiedozenten, die Statistik unterrichten, haben falsche Vorstellungen darüber, was ein signifikantes Ergebnis bedeutet. Bei Psychologiestudenten, die sich in ihrem Studium ausführlich mit Signifikanztests beschäftigen und die bereits eine Statistik-Vorlesung erfolgreich absolviert hatten, waren es 100% (siehe Krauss und Wassner, 2001).

Bei meiner kleinen Untersuchung erkannte nur einer von 45 Schülern, dass keine der 6 Aussagen folgt. Dabei unterrichteten zwei Lehrer eines Gymnasiums ihre 12. Klassen zwei Wochen zum Thema „Signifikanztest“; diese Klassen und ihre Lehrer nennen wir im Folgenden a und b. Der Unterricht in Klasse a orientierte sich am Buch von Distel und Feuerlein (2010), in Klasse b an Jahnke und Scholz (Hrsg., 2010). Lehrer und Schüler wussten, dass am Ende der zwei Wochen ein Test stattfindet. Lehrer a kannte den Inhalt des Tests bereits am Anfang der beiden Wochen, Lehrer b erst am Tag vor seiner Durchführung. Beide Lehrer sollten ihren „gewohnten“ Unterricht halten. Der Test enthielt nicht nur den oben angegebenen Fragebogen, sondern auch zwei Aufgaben zum Signifikanztest im Stil des bayerischen Abiturs. Bei den „Abitur“-Aufgaben erreichten die Schüler im Durchschnitt 81% aller Punkte: Sie können zwar Abituraufgaben lösen, aber Signifikanztests nicht richtig interpretieren.

Das verständnislose und ständig wiederholte Durchführen von Nullhypothesentests nennen Gigerenzer, Krauss und Vitouch „the null ritual“ (2004). In ihrem Artikel werden einige Ursachen für falsche Vorstellungen genannt.

2 Ein Unterrichtsgespräch

Meines Erachtens gibt es noch weitere Ursachen. Darauf weist ein Unterrichtsgespräch hin, das zwei Tage nach Durchführung des Tests in Klasse a stattfand. Ziel dieses Gesprächs war, die falschen Vorstellungen herauszufinden, die hinter den falschen Antworten der Schüler stecken. Der Lehrer hielt sich weitgehend zurück, bewertete nichts, und stellte im Wesentlichen nur ein paar Fragen. Am Anfang der Unterrichtsstunde wurde mitgeteilt:

1. Der Fragebogen wurde von der ganzen Klasse falsch beantwortet.
2. Im Buch der anderen Klasse (Jahnke und

[1] Ich bedanke mich bei Herrn Prof. Dr. Krauss für eine Reihe anregender Gespräche.

Scholz, Hrsg., 2010) stehen auf den Seiten 140 und 142 falsche Interpretationen von Signifikanztests. [2]

3. „Die Wahrscheinlichkeit für den Fehler 1. Art ist höchstens 5%“ bedeutet: „Wenn H_0 gilt, dann ist die Wahrscheinlichkeit, dass man H_0 ablehnt, höchstens 5%“; kurz: $P_{H_0}(A) \leq 5\%$.

Die Formulierung in 3. tauchte zwar auch in manchen früheren Gesprächen mit der Klasse auf, aber meistens herrschte die Redewendung des Buches von Distel und Feuerlein (2010) vor: „Die Wahrscheinlichkeit dafür, dass H_0 irrtümlich abgelehnt wird.“

Die Klasse wurde nun aufgefordert, den Begriff „Fehler 1. Art“ möglichst kurz und präzise mittels H_0 und A zu erklären. Die Schüler nannten drei Möglichkeiten: „H_0 wird irrtümlich abgelehnt“, „A, obwohl H_0“ und „A, aber H_0“. Der Lehrer forderte eine stärkere Präzisierung: Was genau bedeutet „irrtümlich“, „obwohl“, „aber“? Es sollten nur noch die in der Logik üblichen Ausdrücke „oder“, „und“, „nicht“, „wenn ..., dann ...“ und so weiter verwendet werden. Die Schüler nannten wieder drei Möglichkeiten:

1. H_0 und A.
2. Wenn H_0, dann A.
3. Wenn A, dann H_0.

In der anschließenden Diskussion, wer denn nun recht hat, wurden auch Schulbücher benutzt. Dort findet man Tabellen der folgenden Art:

	A	$\overline{A}$
H_0	Fehler 1. Art	...
$\overline{H_0}$	...	Fehler 2. Art

Am Ende der Diskussion stimmte die Klasse über zwei Fragen ab: Was verstehen unsere Bücher unter einem Fehler 1. Art? 23 von 26 Schülern waren für „H_0 und A“. Was versteht ihr selbst unter einem Fehler 1. Art? 24 Schüler waren für „Wenn H_0, dann A“; als Grund für diese Entscheidung wurde angegeben, dass H_0 bei einem Fehler 1. Art vorausgesetzt wird (vergleiche 3. Mitteilung am Anfang der Stunde).

Anschließend stellte der Lehrer folgende Fragen: Wenn man „Fehler 1. Art“ auf diese Weisen festlegt, was bedeutet dann „Wahrscheinlichkeit eines Fehlers 1. Art“? Etwa $P(H_0 \cap A)$, oder $P(H_0 \Rightarrow A)$? Welchen Sinn hat die Wahrscheinlichkeit einer logischen Folgerung? Leider blieb keine Zeit, diese Fragen zu klären. Die Klasse schien nun ziemlich ratlos zu sein.

3 Ursachen

Die Literatur beschäftigt sich seit mindestens 30 Jahren immer wieder mit Problemen bei Signifikanztests; siehe zum Beispiel Birnbaum (1982), Diepgen (1985), Stein (1993), Waldmüller (1998), Buth (2003), Gigerenzer, Krauss, Vitouch (2004), Diepgen (2012). Ich teile die Auffassung, dass ein wesentlicher Grund für falsche Interpretationen darin liegt, dass sich entgegen der Logik immer wieder ein Wunschdenken durchsetzt (siehe Gigerenzer, Krauss, Vitouch (2004), Seite 400): Eigentlich möchte man mit Sicherheit wissen, ob eine Hypothese H wahr ist. Das geht leider nicht, also möchte man wenigstens $P(H)$ wissen. Das geht aber auch nicht, zumindest nicht im Rahmen eines Signifikanztests: Er redet zwar von Wahrscheinlichkeiten und vom Test einer Hypothese, aber er liefert nicht das gewünschte $P(H)$; siehe auch Buth (2003).

Das oben beschriebene Unterrichtsgespräch enthält Hinweise für eine weitere Ursache: Einige der bisher üblichen Erklärungen der Grundbegriffe sind unzureichend. Wenn Grundbegriffe unklar bleiben, dann hat es die Logik natürlich schwer, sich gegen ein Wunschdenken durchzusetzen.

Die bisher üblichen Erklärungen des Begriffes „Fehler 1. Art“ werden von vielen Schülern als „H und A“ aufgefasst, wobei oft nicht genau zwischen einer Aussage A und einem Ereignis A unterschieden wird. Sie kennen außerdem nur *eine* Möglichkeit, Und-Verknüpfungen von Aussagen in entsprechende Verknüpfungen von Ereignissen zu übersetzen, nämlich mit Hilfe von Schnittmengen. Damit landet man schnell bei dem Missverständnis, den „Fehler 1. Art“ und die Aussage „H und A“ als Ereignis $H \cap A$ zu interpretieren. Diese Interpretation wird zusätzlich durch Tabellen wie im Unterrichtsgespräch nahegelegt, die man leicht mit sogenannten 4-Felder-Tafeln verwechseln kann:

	A	$\overline{A}$
B	$P(B \cap A)$	...
$\overline{B}$	...	...

Wenn man Andeutungen wie „A, obwohl H“ (bewusst oder unbewusst) als $H \cap A$ auffasst, dann sind

[2] Auf Seite 140 wird $P_{0,03}^{100}(X > 3) \approx 35\%$ so interpretiert: „Die Wahrscheinlichkeit dafür, dass die Maschine trotzdem ordnungsgemäß arbeitet, beträgt höchstens 35%.“ Dabei ist „Maschine arbeitet ordnungsgemäß“ die Nullhypothese, siehe Zeile 12 von unten.

„A, obwohl H“ und „H, obwohl A“ gleichbedeutend, da $H \cap A = A \cap H$. Damit ist auch nicht mehr klar, ob H, oder A, oder ob überhaupt etwas vorausgesetzt wird (siehe Vorschläge der Schüler). Verwechslungen zwischen $P_H(A)$, $P_A(H)$ und $P(H \cap A)$ sind dann kaum mehr zu vermeiden. Daraus ergibt sich ein weit verbreiteter Fehler bei der Interpretation von Signifikanztests: Man traut ihnen Aussagen über Hypothesen zu, wie etwa $P_A(H)$. Dieser Fehler steckt hinter den ersten fünf Aussagen des oben angegebenen Fragebogens. Bei der sechsten Aussage fehlt die Voraussetzung, dass H_0 gilt.

Ich sehe also zwei Hauptprobleme:

1. Ein Konflikt zwischen Wunsch und Logik
2. Unklare Begriffe.

4 Vorschläge

Meines Erachtens sollte man den Konflikt zwischen Wunsch und Logik mit Hilfe geeigneter Begriffe aufheben. Mit geeigneten Begriffen gewinnt die Logik exakte Aussagen und Folgerungen. Mit geeigneten Begriffen kann man aber auch Wünsche klar zur Sprache bringen und Voraussetzungen genau benennen, unter denen sie erfüllbar sind.

Wer zum Beispiel nur erklärt, dass $P(H)$ ein sinnloser Ausdruck ist, da H kein Ereignis ist, der spricht nur davon, wie man sein Ziel nicht erreicht. Das ist nicht nur unbefriedigend, sondern auch nur die halbe Wahrheit: Die Bayes-Statistik verwendet $P(H)$, und liefert das gewünschte $P_A(H)$. Ich teile die Ansicht von Krauss und Wassner (2001), Signifikanztests und Bayes-Statistik im Unterricht miteinander zu vergleichen. Selbstverständlich sollte man darauf achten, dass ein solcher Vergleich nicht zu Verwirrungen führt, siehe Diepgen (2012). Auch dazu braucht man klare und geeignete Begriffe.

Bei der Wahl der Grundbegriffe orientiere ich mich an Krengel (2003), §6.2. Allerdings beschränke ich mich auf den Spezialfall, dass nur Binomialverteilungen betrachtet werden.

Definition 1. Ein *Signifikanztest* (für Binomialverteilungen) wird festgelegt durch (bzw. ist ein Quadrupel bestehend aus)

1. Eine Zahl $n \in \mathbb{N}$
2. Eine Teilmenge $M \subseteq [0;1]$
3. Eine Teilmenge $H \subseteq M$
4. Eine Teilmenge $A \subseteq \{0;1;\ldots;n\}$.

Die Zahl n heißt *Stichprobenlänge*, M heißt *Menge der möglichen Trefferwahrscheinlichkeiten*, H heißt *Hypothese*, und A heißt *Ablehnungsbereich* . □

Eine solche Definition dient der Klärung und Zusammenfassung von Begriffen, über die man bereits ausführlich gesprochen hat. Sie hält zum Beispiel eine Entscheidung über die Frage fest, ob Hypothesen im mathematischen Modell durch Aussagen oder durch Mengen dargestellt werden sollen. Außerdem kann man Definition 1 als Gliederung und Anleitung zum Lösen von Aufgaben auffassen:

Beispiel 1. Gegeben sei eine Kiste, die entweder 3% oder 10% defekte Schrauben enthält. Durch eine Stichprobe der Länge 100 (mit Zurücklegen) soll die Vermutung getestet werden, dass die Kiste 3% defekte Schrauben enthält.
Einem solchen Text muss man zunächst die wesentlichen Informationen entnehmen. In Definition 1 steht, was wesentlich ist:

1. $n = 100$
2. $M = \{0{,}03\,;\,0{,}1\}$
3. $H = \{0{,}03\}$.

Danach kommt die Überlegung, bei welchen Trefferzahlen die Hypothese „nicht plausibel“ erscheint (4. Punkt in Definition 1). Im vorliegenden Beispiel hat A die Form $\{k;\ldots;100\}$. Anschließend kann man, je nach Aufgabe und gegebenen Daten, k bestimmen, oder $P_{0{,}03}(A)$ berechnen, und so weiter. □

Die Verwendung von $H \subseteq M$ und $A \subseteq \{0;1;\ldots;n\}$ macht deutlich, was man bei einem Signifikanztest berechnen kann, und was nicht: Auf M steht kein Maß zur Verfügung, also liefert ein Test keine Wahrscheinlichkeiten für Hypothesen: Weder $P(H)$ noch $P_A(H)$ sind definiert. Im Gegensatz zu M gibt es auf $\{0;1;\ldots;n\}$ die Binomialverteilungen P_m mit $m \in M$, also macht ein Test nur Wahrscheinlichkeitsaussagen über Teilmengen $A \subseteq \{0;1;\ldots;n\}$.

Außerdem kann man jetzt genauer erklären, warum Signifikanztests (nach R. A. Fisher) nur dann eingesetzt werden, wenn man fast nichts über eine Situation weiß. Was bedeutet „fast nichts wissen“? Was müsste man denn wissen? Mit Definition 1 kann man genau auf die Stelle zeigen, an der etwas fehlt: Auf M fehlt ein Maß. Wenn man ein Maß auf M hätte, dann könnte man mit Bayes das gewünschte $P_A(H)$ berechnen, siehe Beispiel 3. Damit kann auch eine deutliche Grenze zwischen Signifikanztests und Bayes-Statistik gezogen werden, die vor

Verwechslungen schützt: Wenn ein Maß auf M zur Verfügung steht, dann verwendet man Bayes; ansonsten ist man leider auf Signifikanztests angewiesen und erhält nicht das gewünschte $P_A(H)$, sondern höchstens den vagen Eindruck, dass H „nicht plausibel erscheint“.

Was ist ein Fehler 1. Art? Es liegt nahe, dass eine Antwort die folgende Form haben sollte: „Ein Fehler 1. Art ist …“, oder auch „… heißt Fehler 1. Art“, wobei in „…“ ein mathematisches Objekt angegeben wird. Definitionen in dieser Form sind in der Mathematik üblich: „Ein Viereck mit 4 gleich langen Seiten heißt Raute.“ Erstaunlicherweise findet man in der Literatur für den Begriff „Fehler 1. Art“ keine Definition in dieser Form. Zum Beispiel steht in Krengel (2003), §6.2 lediglich „Ist $\vartheta \in H$ und wird die Hypothese verworfen, so spricht man von einem Fehler erster Art.“ Was aber *ist* ein „Fehler 1. Art“?

Die Beschreibung in Krengel (2003) könnte man auch so formulieren: Ein Fehler 1. Art tritt genau dann ein, wenn $p \in H$ und $x \in A$; dabei bezeichnet p die tatsächliche Trefferwahrscheinlichkeit und x die Trefferzahl in der Stichprobe. Die Aussage „$p \in H$ und $x \in A$“ ist äquivalent zu „$(p,x) \in H \times A$“, also tritt ein Fehler 1. Art genau dann ein, wenn $(p,x) \in H \times A$. Das erinnert an die Sprechweise „Ein Ereignis B tritt genau dann ein, wenn $\omega \in B$“. Es liegt also nahe, $H \times A$ als „Fehler 1. Art“ zu bezeichnen.

Hier wird deutlich, dass im Unterricht oft nur *ein* Typ von Und-Aussagen genauer behandelt wird: „$x \in A$ und $x \in B$“ ist äquivalent zu „$x \in A \cap B$“. Die Und-Aussage beim Fehler 1. Art ist jedoch von einem anderen Typ: „$p \in H$ und $x \in A$“ kann nicht mit $H \cap A$ angegeben werden, sondern mit $H \times A$. Eine genaue Erklärung des Unterschieds zwischen diesen beiden Typen von Und-Aussagen könnte helfen, falsche Interpretationen zu vermeiden (siehe Abschnitt 3).

An dieser Stelle sollte man im Unterricht vor allem zwei Dinge klären: Die Frage, inwiefern man einen Fehler 1. Art als Ereignis deuten kann, und den Umgang mit einer Menge von Paaren. Die Schüler sollten bereits damit vertraut sein, dass Ergebnisse von zweistufigen Experimenten mit Hilfe von Paaren beschrieben werden, etwa beim zweimaligen Werfen eines Würfels: Die Aussage „Beim 1. Wurf fällt eine 5 und beim 2. Wurf fällt eine 3“ wird mit dem Paar $(5;3)$ beschrieben. An diese Schreibweise kann man im folgenden Beispiel anknüpfen.

Beispiel 2. Die Kiste in Beispiel 1 stammt aus einer Lagerhalle, in der nur Kisten mit 3% oder 10% defekten Schrauben liegen. In einem ersten Schritt wird eine Kiste zufällig ausgewählt. (Dummerweise fehlen einmal wieder sämtliche Etiketten.) Ergebnisraum des ersten Schritts: $M = \{0{,}03\,;\,0{,}1\}$. Im zweiten Schritt wird eine Stichprobe wie in Beispiel 1 entnommen. Die Aussage „In der Kiste befinden sich 3% defekte Schrauben und in der Stichprobe 15 defekte Schrauben“ wird mit dem Paar $(0{,}03\,;15)$ beschrieben. Dieses Paar beschreibt außerdem (sehr kompakt!) eine Möglichkeit für das Eintreten eines Fehlers 1. Art, wenn eine Entscheidungsregel mit Ablehnungsbereich $A := \{7;\ldots;100\}$ befolgt wird.

Der Ergebnisraum dieses zweistufigen Experiments besteht also aus allen Paaren $(p;x)$ mit $p \in M$ und $x \in \{0;1;\ldots;100\}$, das heißt $\Omega := M \times \{0;1;\ldots;100\}$. Ein Fehler 1. Art tritt genau dann ein, wenn $(p;x) \in \{(0{,}03\,;\,7),\ldots,(0{,}03\,;\,100)\} = H \times A$. Insbesondere ist $H \times A \subseteq \Omega$, also kann der Fehler 1. Art als ein Ereignis eines Experiments aufgefasst werden. □

Definition 2. Es seien H und A wie in Definition 1. Die Menge $H \times A$ heißt *Fehler 1. Art*. Sprechweise: Ein Fehler 1. Art tritt genau dann ein, wenn $(p,x) \in H \times A$; dabei ist p die tatsächliche Trefferwahrscheinlichkeit und x die Anzahl der Treffer in der Stichprobe. □

Definition 2 habe ich bisher noch nicht in der Literatur gefunden, und es liegen wohl auch noch keine Erfahrungen über ihren Einsatz im Unterricht vor. Ob man sie im Unterricht verwendet, hängt unter anderem davon ab, ob man von der Aussage „$p \in H$ und $x \in A$“ zur Aussage „$(p,x) \in H \times A$“ übergehen soll. Ich lade zu einer Diskussion über Definition 2 ein, und fasse meine Argumente zusammen:

1. In Definition 2 wird der Begriff „Fehler 1. Art“ als Name eines Objekts festgelegt. Damit wird die Frage „Was ist ein Fehler 1. Art?“ genau beantwortet.

2. Die bisher üblichen Erklärungen des Begriffes „Fehler 1. Art“ fassen viele Schüler mehr oder weniger vage als $H \cap A$ auf, da sie nur *eine* Möglichkeit kennen, Und-Aussagen mit Hilfe von Mengen darzustellen, nämlich mit Schnittmengen (siehe Abschnitt 3). Mit $H \times A$ kann deutlich hervorgehoben werden, dass der Fehler 1. Art zu einem anderen Typ von Und-Aussagen gehört.

3. Wer nur erklärt, warum man einen Fehler 1. Art nicht mit $H \cap A$ darstellen kann, der spricht nur davon, wie etwas nicht geht. Befriedigender und auf Dauer auch einprägsamer wäre es zu sagen, wie es geht: Mit $H \times A$ statt $H \cap A$.

4. Die Schreibweise $H \times A$ sollte keine Schwierigkeiten bereiten, wenn sie hauptsächlich als Abkürzung für etwas bereits Bekanntes verwendet wird: Für eine Menge von Paaren.

In der Literatur werden „Wahrscheinlichkeit eines Fehlers 1. Art“ und „Risiko 1. Art“ als gleichwertig angesehen. Die erste Formulierung halte ich jedoch für irreführend: Sie legt nahe, dass es sich dabei um die Wahrscheinlichkeit eines Ereignisses namens „Fehler 1. Art“ handelt, also um so etwas wie $P(\text{„Fehler 1. Art“})$ (siehe Ende Abschnitt 2).

Definition 3. Es seien H und A wie in Definition 1. Wenn $h \in H$, dann heißt $P_h(A)$ ein *Risiko 1. Art*. Ein *Signifikanzniveau* für einen Test ist eine obere Schranke für die Menge $\{P_h(A) \mid h \in H\}$. □

Ein Signifikanzniveau ist also eine reelle Zahl α mit folgender Eigenschaft: $h \in H \implies P_h(A) \leq \alpha$.

Die Tatsache, dass Signifikanztests (nach R. A. Fisher) höchstens ein vages „nicht plausibel“ liefern, wird auf viele Schüler noch keinen besonderen Eindruck machen, wenn sie es gewohnt sind, sich auf „plausible Argumente“ zu verlassen (die Beschränkung auf „Plausibilität“ ist zur Zeit sehr in Mode): „H ist nicht plausibel“ kann so verstanden werden, dass es „plausibel“ sei, H abzulehnen; und wenn „plausible Argumente“ auch im übrigen Mathematik-Unterricht berechtigt sind, dann könnte man H mit voller Berechtigung ablehnen. Ich fürchte, der Unterschied zwischen einer „plausiblen Folgerung“ und einer logischen Folgerung wird im Unterricht nicht klar genug herausgestellt.

Daher sollte man anhand geeigneter Beispiele zeigen, dass das vage „nicht plausibel“ eines Signifikanztests einen völlig falschen Eindruck von den tatsächlichen Verhältnissen erzeugen kann: Bei einem Signifikanzniveau von 5% ist es ohne weiteres möglich, dass $P(H)$ und $P_A(H)$ deutlich über 50% liegen. Wer sich nur auf Signifikanztests beschränkt, muss auf solche Beispiele verzichten, da dann $P(H)$ und $P_A(H)$ nicht definiert sind. Wer zum Vergleich auch Bayes-Statistik behandelt, sollte auf eine genaue Unterscheidung dieser beiden Verfahren achten. Definition 1 unterstützt eine solche Unterscheidung:

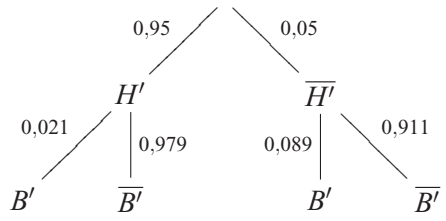
Bei einem Signifikanztest liegen Hypothese und Stichprobenergebnis nicht in einem gemeinsamen Ergebnisraum, sondern in jeweils getrennten Mengen M und $\{0;\ldots;n\}$; auf M gibt es kein Maß; damit sind Ausdrücke wie $H \cap A$ und $P_A(H)$ sinnlos. In der Bayes-Statistik verwendet man dagegen ein anderes Modell: Hypothese und Stichprobe werden als Ereignisse aufgefasst, die in einem gemeinsamen Ergebnisraum $\Omega := M \times \{0;\ldots;n\}$ liegen; auf M gibt es ein Maß, also auch auf Ω; damit ist $P_{A'}(H')$ für Teilmengen A' und H' in Ω definiert. Bei einem Signifikanztest würde es nichts nützen, Stichproben-Ereignisse als Teilmengen $A' \subseteq \Omega$ aufzufassen, da man dann keine Wahrscheinlichkeit von A' berechnen könnte: Auf Ω fehlt ein Maß.

Beispiel 3. Wenn sich in der Stichprobe des zweiten Beispiels 7 defekte Schrauben befinden, dann könnten wir die Hypothese H ablehnen, und zwar auf einem Signifikanzniveau von $P_{0,03}(A) \approx 0{,}031$. Doch mit welcher Berechtigung lehnen wir H ab? Mit welcher Sicherheit können wir sagen, dass H nicht gilt? Dazu bräuchten wir ein Maß auf M.

Es sei nun zusätzlich bekannt, dass in der Lagerhalle 95 Kisten mit 3% defekten Schrauben, und 5 Kisten mit 10% defekten Schrauben liegen. Damit hat man ein Maß auf M und auf $\Omega := M \times \{0;\ldots;100\}$.

In der Bayes-Statistik wird die Hypothese „Beim 1. Schritt erhält man eine Kiste mit 3% defekten Schrauben“ mit Elementen aus Ω dargestellt, das heißt als $H' := \{(0{,}03\,;0),\ldots,(0{,}03\,;100)\}$. Zum Verständnis hilft wieder ein Vergleich mit dem zweimaligen Würfeln: „Beim 1. Wurf fällt eine 5“ wird mit $\{(5\,;1),\ldots,(5\,;6)\}$ dargestellt.

Wenn die Trefferzahl $x = 7$ bekannt ist, dann gibt es im Rahmen der Bayes-Statistik keinen Grund, $x = 7$ durch die schwächere Aussage $x \in \{7;\ldots;100\}$ zu ersetzen. Das Ereignis „Trefferzahl 7“ wird ebenfalls als Teilmenge von Ω dargestellt: $B' := \{(0{,}03\,;7)\,,(0{,}1\,;7)\}$. Zugehöriges Baumdiagramm:



Mit Bayes folgt

$$P_{B'}(H') \approx \frac{0{,}95 \cdot 0{,}021}{0{,}95 \cdot 0{,}021 + 0{,}05 \cdot 0{,}089} \approx 0{,}82\,.$$

Wenn man also zusätzliche Informationen hat, dann könnte sich herausstellen, dass H' eine Wahrscheinlichkeit von 82% besitzt. Wenn man nur die Informationen eines Signifikanztests besitzt, dann kennt man nur $P_{0,03}(A) \approx 0{,}031$ und würde H vielleicht ablehnen, weil H „nicht plausibel erscheint“. □

Signifikanztests liefern also nur schwache Indizien für die Frage, ob Hypothesen wahr sind. R. A. Fisher, der „Vater“ der Signifikanztests, schreibt über sein eigenes Verfahren: „On the whole the ideas (a) […] and (b) that the purpose of the test is to discriminate or ‘decide’ between two or more hypotheses, have greatly obscured their understanding, when taken not as contingent possibilities but as elements essential to their logic.“ (Fisher 1959, S. 42, f.)

Zum Verständnis der Bayes-Statistik sollte man natürlich auch ihren Zweck erläutern: Es geht um eine Neubewertung von Hypothesen im Lichte neuer Daten: Vor der Stichprobe in Beispiel 3 kennt man $P(H') = 0{,}95$, nach der Stichprobe $P_{B'}(H') \approx 0{,}82$.

Die Überlegungen in Beispiel 3 kann man auch auf einfache HIV-Tests übertragen (vergleiche Gigerenzer, Krauss, Vitouch 2004, S. 396):

Beispiel 4. Eine Person wird nur 1 mal auf HIV getestet. Die Wahrscheinlichkeit, dass bei einer tatsächlich nicht infizierten Person ein HIV-Test positiv ausfällt, sei $0{,}0001$, bei einer tatsächlich infizierten Person $0{,}9999$. Zugehöriger Signifikanztest:

1. $n = 1$
2. $M = \{0{,}0001\,;\,0{,}9999\}$
3. $H = \{0{,}0001\}$ (Person nicht infiziert)
4. $A = \{1\}$ (HIV-Test positiv).

Wenn ein HIV-Test positiv ausfällt, dann könnte man H ablehnen, und zwar auf einem Signifikanzniveau von $0{,}0001$. Daraus kann man jedoch nicht schließen, dass die betreffende Person „höchstwahrscheinlich“ infiziert ist: Wenn zusätzlich $P(H') = 0{,}9999$ bekannt ist, dann folgt (mit einer Rechnung wie in Beispiel 3) $P_{A'}(H') = 0{,}5$. □

Die hier vorgeschlagene Sprache ist auch auf Situationen anwendbar in denen es zweifelhaft erscheint, was ein Wahrscheinlichkeitsmaß auf M bedeuten soll:

Beispiel 5. Eine Partei hat die Vermutung, dass ihr momentaner Stimmenanteil in der Bevölkerung höchstens 5% beträgt. Es werden 1000 Personen befragt. Signifikanztest:

1. $n = 1000$
2. $M = \{0; 0{,}01; \ldots; 1\}$ (tatsächlicher momentaner Stimmenanteil in der Bevölkerung, gerundet auf ganze Prozent)
3. $H = \{\,0\,;\,0{,}01\,;\,0{,}02\,;\,0{,}03\,;\,0{,}04\,;\,0{,}05\,\}$
4. $A = \{\,70\,;\,71\,;\,\ldots\,;\,1000\,\}$.

Man kann sich trefflich darüber streiten, was ein Wahrscheinlichkeitsmaß auf M bedeuten soll, und wie ein zugehöriges „Zufalls-Experiment“ aussehen könnte. Dem mathematischen Modell sind solche Spekulationen egal: Auf M fehlt ein Maß, wenn nur das Ergebnis der Umfrage zur Verfügung steht.

Wenn sich in der Umfrage 70 Personen für die Partei aussprechen, dann könnte man H auf einem Niveau von $P_{0{,}05}(A) \approx 0{,}003$ ablehnen. Die Partei sollte daraus jedoch nicht den Schluss ziehen, dass sie bei der nächsten Wahl „höchstwahrscheinlich“ über die 5%-Hürde kommt. Das mögen uns vielleicht zahlreiche Wahlprognosen und unser Bauchgefühl nahelegen. Man sollte „höchstwahrscheinlich über 5%“ aber nicht mit einer mathematischen Aussage in Rahmen eines Signifikanztests verwechseln. □

5 Wozu Signifikanztests?

Was weiß man über die Gültigkeit einer Hypothese, wenn eine Stichprobe im Ablehungsbereich liegt? Nichts. Was folgt aus $P_h(A) \leq 5\%$? Nichts. Wozu soll man dann $P_h(A)$ berechnen? Um eine Meinung über eine Hypothese zu begründen, die mathematisch aber nicht begründet werden kann? Sind Signifikanztests in der Schule überhaupt sinnvoll? Ich fürchte, Aussagen wie „Ich lehne H auf einem Signifikanzniveau von 5% ab“ werden in unserer Gesellschaft immer wieder dazu benutzt, mathematisch fundierte Erkenntnisse vorzugaukeln. Solange Zahlen für pseudo-mathematische Argumente missbraucht werden, solange sollte der Mathematik-Unterricht darüber aufklären, wie wenig zum Beispiel ein Signifikanzniveau aussagt.

Literatur

Birnbaum, I. (1982): Die Interpretation statistischer Signifikanz (übersetzt von Günter Fillbrun). Stochastik in der Schule, Heft 2.

Buth, M. (2003): Anmerkungen zum Testen von Hypothesen. Stochastik in der Schule, Heft 1.

Diepgen, R. (1985): Was Schüler zum Hypothesentesten wissen sollten. Stochastik in der Schule, Heft 1.

Diepgen, R. (2012): Leserbrief zum Beitrag „Hypothesentests und bedingte Wahrscheinlichkeit“ von Renate Motzer (2010). Stochastik in der Schule, Heft 2.

Distel, B. und Feuerlein, R. (2010): Mathematik 12, G8 Bayern. bsv

Fisher, R. A. (1959): Statistical methods and scientific inference. Edinburgh, UK: Oliver & Boyd.

Gigerenzer, G., Krauss, S., Vitouch, O. (2004): The null ritual. In D. Kaplan (Ed.), The Sage handbook of quantitative methodology for the social sciences (pp. 391-408). Thousand Oaks, CA: Sage.

Jahnke, T. und Scholz, D. (Hrsg., 2010): Fokus Mathematik 12, Gymnasium Bayern. Cornelsen.

Krauss, S. und Wassner, C. (2001): Wie man das Testen von Hypothesen einführen sollte. Stochastik in der Schule, Heft 1.

Krengel, U. (2003): Einführung in die Wahrscheinlichkeitstheorie und Statistik. Vieweg.

Oakes, M. (1986): Statistical inference: A commentary for the social and behavioral sciences. Chichester: Wiley.

Stein, G. (1993): Schwierigkeiten mit der Nullhypothese. Stochastik in der Schule, Heft 1.

Waldmüller, B. (1998): Was sagen signifikante Ergebnisse? - Zu einem Beispiel aus der Zeitung. Stochastik in der Schule, Heft 3.

Anschrift des Verfassers
Matthias Moßburger
Hans-Leinberger-Gymnasium
Jürgen-Schumann-Straße 20
84034 Landshut
mossburger.hlg@gmx.de