Arbeitsbericht des AK Stochastik 1999/ 2000

Manfred Borovcnik, Joachim Engel, Dieter Wickmann (Hrsg.)

Anregungen zum Stochastikunterricht:

- Die NCTM - Standards 2000
- Klassische und Bayessche Sichtweise im Vergleich

Statistisches Denken oder statistische Rituale: Was sollte man unterrichten?

Gerd Gigerenzer und Stefan Krauss

Bericht von zwei Arbeitskreistagungen des
Arbeitskreises „Stochastik in der Schule" in der
Gesellschaft für Didaktik der Mathematik e.V.
vom 29./ 30. Oktober 1999 und
10.-12. November 2000 in Berlin

©2001 by Verlag Franzbecker, Hildesheim, Berlin

Statistisches Denken oder statistische Rituale: Was sollte man unterrichten?

Gerd Gigerenzer und Stefan Krauss, Max-Planck-Institut für Bildungsforschung

Der vorliegende Beitrag beschäftigt sich mit einem wunden Punkt des derzeitigen Statistikunterrichts (1), mit dessen Folgen für die Wissenschaft (2) und mit Vorschlägen zur Verbesserung der Situation (3).

(1) Beim derzeitigen Statistikunterricht werden weder die umstrittene Stellung derSignifikanztests in der Wissenschaft noch die Aussagekraft eines signifikanten Testergebnisses klar genug dargestellt. Dies gilt für den schulischen Stochastikunterricht wie auch für den Statistikunterricht an der Universität.

(2) Dieses Versäumnis führt zu schwerwiegenden Folgen in den Sozialwissenschaften, wo die Anwendung von Signifikanztests oft an zwanghaftes Händewaschen erinnert. Wir werden diesen Punkt vor allem am Beispiel der Psychologie demonstrieren.

(3) Das Problem kann behoben werden, wenn wir beginnen, „statistisches Denken" statt „statistische Rituale" zu unterrichten. Wir werden erläutern, was die Umsetzung dieses Konzepts für den konkreten Fall des Schulunterrichts im Hypothesentesten bedeutet.

1 Signifikanz – Alles klar?

Statistik ist das meist unterrichtete Fach an deutschen Universitäten. Schätzungsweise jeder dritte Student muss im Verlaufe seines Studiums einen Kurs in Statistik belegen[1]. In den Sozialwissenschaften ist es heutzutage in der Forschung üblich, Hypothesen (soziologische, psychologische, pädagogische u.s.w.) über eine Population mit Hilfe von Signifikanztests anhand einer Stichprobe zu überprüfen. In der Psychologie ziehen derzeit nahezu alle experimentellen Arbeiten Signifikanztests zur wissenschaftlichen Untermauerung heran. Leider erinnert die Anwendung der Signifikanztests dabei meist an ein Ritual, an den gutgemeinten Versuch, die Zehn Gebote der empirischen Sozialwissenschaften (z.B. "Du sollst kein nichtsignifikantes Ergebnis publizieren", siehe Gigerenzer, 1993a) einzuhalten. Die Fragen, die uns in diesem Artikel beschäftigen, sind: Wie kam es zu der Vormachtstellung der Signifikanztests in den Sozialwissenschaften? Was sind die Schwächen des derzeitigen Statistikunterrichts an der Universität? Und schließlich: Was kann man daraus für den Stochastikunterricht in der Schule lernen?

Wir möchten Sie bitten, sich erst einmal etwas Zeit zu nehmen und über die folgende Frage nachzudenken:

"Was bedeutet die Aussage, ein statistischer Kennwert sei auf dem 5%-Niveau signifikant?"

Stellt man diese Frage Psychologiestudenten in der Vordiplomsprüfung, geben diese – nach einem intensiven Training in Statistik (gewöhnlich über zwei Semester) und wohlvorbereitet auf das Vordiplom – darauf falsche Antworten wie die folgenden[2]:

- Es bedeutet, dass er um 5% über der Zufalls-Prozentualität liegt.
- Das bedeutet, die Wahrscheinlichkeit, dass das Ergebnis ein Irrtum ist, beträgt 5%.
- Auf dem 5%-Niveau können nur 5% überzufällig auftreten. Der Rest bewegt sich im Bereich des Zufalls.
- Die Richtigkeit des Ergebnisses ist mit 95% Wahrscheinlichkeit bewiesen.
- Die Aussage ist für 95% der Fälle zutreffend.
- 5% Signifikanz bedeutet: mit 95% Wahrscheinlichkeit wird sich bei Versuchswiederholung das

[1] An den Universitäten der größten deutschen Städte ergibt sich folgendes Bild: Statistik wird unterrichtet an 7 von 11 Fachbereichen der HU Berlin, an 9 von 12 Fachbereichen der FU Berlin, an 7 von 15 Fachbereichen der TU Berlin, an 8 von 19 Fachbereichen der LMU München, an 5 von 7 Fachbereichen der Uni Köln und an 9 von 19 Fachbereichen der Uni Hamburg. Aus diesen Zahlen schätzen wir, dass in ca. 50% aller Fachbereiche der deutschen Universitäten Statistik unterrichtet wird (Quelle: Vorlesungsverzeichnisse der aufgeführten Universitäten von 1999-2001).

[2] Die folgenden Antworten sind dem Psychologievordiplom 1981 an der LMU München entnommen.

gleiche Versuchsergebnis einstellen.
- Es handelt sich annähernd um eine Normalverteilung. Es wurde eine Anzahl von weniger als 30 Versuchspersonen untersucht. Der α-Fehler beträgt .05. 95% der Aussage läßt sich aus dem Treatment erklären. H_0 ist bestätigt.
- Ein statistischer Kennwert ist mit 95% Wahrscheinlichkeit für die gesamte Population signifikant. Nur 5% der Population lassen sich nicht durch diesen Kennwert erfassen.

Es sei angemerkt, dass die meisten dieser Studenten die Prozedur der *formalen Durchführung* eines Signifikanztests fehlerfrei beherrschen.

Sind Psychologiestudenten nun die einzigen, die Schwierigkeiten mit dem Verständnis des statistischen Konzepts der Signifikanz haben? Dass auch bei wissenschaftlich arbeitenden Psychologen Konfusion bezüglich der Bedeutung eines signifikanten Testergebnisses herrscht, zeigt eine Untersuchung mit Statistikdozenten an Berliner Universitäten. Wir stellten 44 Studenten und 69 Dozenten[3] der Psychologie folgende Frage:

Stellen Sie sich vor, Sie prüfen mit Hilfe eines t-Tests[4] statistisch, ob der zwischen zwei Stichproben gefundene Unterschied signifikant ist. Es stellt sich heraus, dass der Unterschied auf dem 1%-Niveau signifikant ist. Welche der folgenden Aussagen lassen sich nun aus dieser Tatsache folgern? (Es können dabei mehrere oder auch keine Aussage zutreffen.)

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.*

Alle diese Aussagen sind falsch[5] (für eine genauere Beschreibung dieser Studie siehe Krauss & Wassner, 2001). Dennoch glaubte *jeder* der befragten Studenten und immerhin noch 86% (!) der Dozenten[6], dass eine oder mehrere dieser Aussagen richtig seien.

Die Studenten scheinen also nicht selbstverschuldet in ihre Verwirrung über die Signifikanz geraten zu sein. Doch auch mit den Dozenten ist das letzte Glied in der Schuldkette noch nicht gefunden. Im Folgenden stellen wir unseren Hauptverdächtigen für die Verschleierung der Bedeutung eines signifikanten Testergebnisses vor. Wie jede gerichtliche Schuldklärung eine genaue Analyse der Vergangenheit erfordert, so wollen wir nun die Geschichte der Signifikanztests in der Psychologie genauer durchleuchten.

2 Die Entstehung der Signifikanzkonfusion

Vor 1930 waren Nullhypothesentests in Psychologie und Pädagogik praktisch unbekannt. Bedeutende psychologische Theorien, wie die von Piaget, Köhler, Pawlow, Skinner oder Freud, entstanden ganz ohne den Einsatz von Signifikanztests (Gigerenzer, 1993a). Vor 1940 gab es insgesamt nur 17 Artikel, die Nullhypothesentests verwendeten. 1955 jedoch verwendeten bereits 80%

[3] Es muss erwähnt werden, dass sich in der Stichprobe der Dozenten auch Statistiktutoren befanden. Diese sind selbst noch Studenten (aus dem Hauptstudium) und haben die Aufgabe, Studenten aus dem Grundstudium in die Statistik einzuführen.

[4] Der t-Test zur Überprüfung von Mittelwertsunterschieden ist der von Psychologiestudenten am häufigsten benutzte Test. Die hier gestellte Frage geht auf Oakes (1986) zurück.

[5] Warum sind diese Aussagen alle falsch? Der Begriff „Signifikanz" impliziert immer „eine Welt", in der die Nullhypothese (H_0) gültig ist. Man betrachtet die *Wahrscheinlichkeit der Daten* unter der Annahme, die Nullhypothese sei richtig. Diese Wahrscheinlichkeit, p(Daten | H_0), wird auch „P-Wert" genannt. Ein Signifikanztest erlaubt keine Aussagen über die Wahrscheinlichkeit von Hypothesen, wie in Aussagen 2 und 4 oder gar Beweise von Hypothesen, wie in Aussagen 1 und 3. Aussage 5 ist ebenfalls eine Aussage über die Wahrscheinlichkeit einer Hypothese (denn wenn die Entscheidung, die Nullhypothese zu verwerfen, falsch ist, muss die Nullhypothese richtig sein). Aussagen über die Wahrscheinlichkeiten von Hypothesen (wie Aussagen 1 bis 5) sind nur in der Bayesstatistik möglich (siehe dazu Krauss & Wassner, 2001). Aussage 6 schließlich ist eine Aussage über die Wahrscheinlichkeit von signifikanten Daten statt über p(Daten | H_0).

[6] Von den von uns befragten Professoren erkannten ca. 50% alle Aussagen als falsch.

aller Artikel in führenden psychologischen Fachzeitschriften Signifikanztests. Was war in den Jahren von 1940 bis 1955 passiert?

Die Antwort darauf ist bekannt: Sir Ronald Fisher hatte 1925 und 1935 zwei Bücher geschrieben, in denen er das Prinzip des Nullhypothesentestens erläuterte. Sein erstes Buch von 1925 hatte den Titel „Statistical Methods for Research Workers". Der Titel seines zweiten, 1935 erschienenen Buches lautete „The Design of Experiments". Während sich das erste Buch hauptsächlich an Agrarwissenschaftler wandte und für den Großteil der Psychologen technisch zu anspruchsvoll war, war das zweite Buch den Bedürfnissen der Psychologie angepasst. Nach einer anfänglichen Zeit des Zögerns wurde das Konzept des Nullhypothesentestens innerhalb kurzer Zeit zur Institution in der amerikanischen Psychologie, vor allem, nachdem 1942 eine Adaption von Fishers Ideen speziell für Psychologen und Pädagogen erschien, nämlich Guilfords Lehrbuch „Fundamental Statistics in Psychology and Education" (1942). Diese damals revolutionäre Hinwendung zum Signifikanztest vollzog sich nach dem zweiten Weltkrieg auch in Deutschland. Leider wurden Signifikanztests in der Psychologie bald zu einem Ritual, das jedem Experiment übergestülpt werden musste, wenn man eine Veröffentlichung seiner Ergebnisse anstrebte. Das signifikante Testergebnis avancierte gar zum Kriterium für gute Wissenschaft. A. W. Melton, der Herausgeber des "Journal of Experimental Psychology", einer der führenden Zeitschriften auf dem Gebiet der experimentellen Psychologie, äußerte 1962 starke Bedenken, Ergebnisse zu veröffentlichen, die nur auf dem 5%-Niveau signifikant waren. Wirklich einen Platz in seinem Journal verdiente nur, wer Ergebnisse vorweisen konnte, die auf dem 1%-Niveau signifikant waren. Das Traurige an dieser Geschichte ist, dass aus Meltons Begründung deutlich wird, dass auch er die Bedeutung eines signifikanten Testergebnisses nicht verstanden hat. Melton (1962, S.553) schreibt z.B.: „The level of significance measures the confidence that the results of the experiment would be repeatable under the conditions described". Diese Täuschung, die mit Aussage 6 unseres Fragebogens verwandt ist, wird „replication fallacy" genannt. Eine ausführliche Diskussion dieser Täuschung findet sich in Falk und Greenbaum (1995).

Wie kann es sein, dass Studenten, Dozenten und sogar Herausgeber führender Zeitschriften nicht wissen, was ein signifikantes Ergebnis überhaupt bedeutet? Wir denken, dass ein inkonsistenter – oder gar ein unehrlicher – Umgang mit den Signifikanztests in den Lehrbüchern daran schuld ist. Dort wird meist so getan, als ob das Testen auf Signifikanz die *einzige* Methode des statistischen Schlussfolgerns sei. Signifikanztests werden in Lehrbüchern wie geometrische Lehrsätze dargestellt. Mit *Unehrlichkeit* meinen wir dabei: Der Satz von Pythagoras ist in der Tat konsistent. Er ist unwidersprochen und man kann mit ihm zahlreiche lückenlose Beweise führen. Die Interpretation dieser Beweise bereitet keine grundsätzlichen Schwierigkeiten und es gibt keine vergleichbare Alternative zu diesem Satz in der Geometrie. Signifikanztests dagegen sind nicht unwidersprochen und sie sind alles andere als konkurrenzlos. Trotzdem wird versucht, diese Probleme in Lehrbüchern zu verschweigen (oder zumindest zu glätten) und die Signifikanztests im Unterricht unter dem Motto „teaching mathematics as usual" vorzustellen. Dass Schüler und Studenten so den Eindruck gewinnen, mit Signifikanztests könne man – wie sonst in der Mathematik – etwas *beweisen* (siehe Aussagen 1 und 3 unseres Fragebogens), kann ihnen nicht angelastet werden. Folgenden „Mythen" über Signifikanztests müsste unserer Meinung nach in einem ehrlichen Unterricht widersprochen werden:

Mythos 1: Das "Signifikanzniveau" ist ein einheitlich definiertes Konzept

In unserer eingangs gestellten Frage interessierten wir uns für die Bedeutung eines auf dem 5%-Niveau signifikanten Kennwertes. Was würde Sir Ronald Fisher auf diese Frage antworten? Interessanterweise lieferte auch Fisher in seinen Büchern kein konsistentes Bild zu dieser Frage: Zuerst schlug er vor, das Signifikanzniveau *vor* der Durchführung des Tests im Sinne einer Konvention festzulegen (z.B. 5%) (The Design of Experiments, 1935). Später schlug er vor, man solle besser *nach* der Durchführung des Tests das Signifikanzniveau der Daten selbst, also P-Werte, angeben (Statistical Methods and Scientific Inference, 1956). Jerzy Neyman und Egon Pearson, die Ende der 20er Jahre damit begannen, Fishers Ideen weiterzuentwickeln, waren mit beiden Interpretationen

von Fisher sowie generell mit der Prozedur des Nullhypothesentestens unzufrieden. Sie schlugen vor, eine Alternativhypothese zu *konkretisieren*. In der Tat war Neyman der Meinung, Fishers Nullhypothesentests seien im mathematischen Sinne „schlechter als nutzlos" (nämlich immer dann, wenn $1-\beta < \alpha$ ist, wobei α die Wahrscheinlichkeit des Fehlers erster Art und β die Wahrscheinlichkeit des Fehlers zweiter Art ist). Fisher, bei dem es aufgrund seines Ansatzes gar kein β gab, bezeichnete seinerseits Neymans Argumentation als „kindisch" und sagte, dass solche Argumente nur von Mathematikern stammen könnten, die keinen Kontakt zu den Naturwissenschaften hätten (Gigerenzer, 1993a, S. 318). Ein weiterer Streitpunkt der beiden Parteien war, dass Fisher glaubte, mit einem signifikanten Ergebnis konkrete Erkenntnisse über das durchgeführte Einzelexperiment gewinnen zu können, während Neyman und Pearson ein signifikantes Ergebnis *behavioristisch* interpretierten: Nach ihrer Überzeugung lege ein signifikantes Ergebnis lediglich nahe, sich so zu *verhalten*, als ob die H_0 falsch sei. Das heißt, Neyman und Pearson entwickelten eine dritte Interpretation des Signifikanzniveaus: Man solle α und β *vor* der Durchführung des Tests festlegen, aber nicht gemäß einer allgemeingültigen Konvention, sondern vielmehr nach einer genauen Kosten-Nutzen-Analyse eines möglichen Fehlers erster sowie zweiter Art. Die folgende Tabelle verdeutlicht die unterschiedlichen Auffassungen des Begriffs „Signifikanzniveau":

Diese völlig voneinander abweichenden Interpretationen über die Logik eines Signifikanztests werden in Lehrbüchern meist zugunsten einer "Rezeptbuchstatistik" ausgebügelt. Ein Kochrezept zum Testen von Hypothesen sollte schließlich nicht über Gebühr zum Nachdenken animieren. Das ist letztlich auch erreicht worden: Schüler und Studenten können zwar das Ritual eines Signifikanztests durchführen, verstehen aber das Ergebnis meist nicht. Besondere Verständnisschwierigkeiten dürften Studenten haben, wenn sie nach Lehrbüchern wie "Introduction to statistics for psychology and education" von J.C. Nunally (1975) Signifikanztests lernen. In diesem Buch gibt der Autor innerhalb von 3 Seiten (S.194-196) die folgenden 8 Bedeutungen eines signifikanten Testergebnisses an, die alle falsch sind:

- the improbability of observed results being due to error
- the probability that an observed difference is real
- If the probability is low, the null hypothesis is improbable
- the statistical confidence ... with odds of 95 out of 100 that the observed difference will hold up in investigations
- the degree to which experimental results are taken "seriously"
- the danger of accepting a statistical result as real when it is actually due only to error
- the degree of faith that can be placed in the reality of the finding
- the investigator can have 95 percent confidence that the sample mean actually differs from the population mean

„Früher" Fisher (1935)	„Später" Fisher (1956)	Neyman und Pearson
Lege Signifikanzniveau *vor* der Durchführung des Tests im Sinne einer Konvention (z.B. 5%) fest. Das Signifikanzniveau ist somit eine genuine Eigenschaft des *Tests*.	Berechne *nach* der Durchführung des Tests das exakte Signifikanzniveau aus den Daten (P-Werte). Das Signifikanzniveau ist somit eine Eigenschaft der *Daten* und es wird keine willkürlich festgelegte Konvention mehr benötigt.	Lege α und β *vor* der Durchführung des Tests fest. α und β sind die relativen Häufigkeiten eines Fehlers erster bzw. zweiter Art und sind somit Eigenschaften des Tests. Man benötigt aber zu ihrer Festlegung keine Konvention, sondern eine Kosten-Nutzen-Abschätzung der möglichen Fehlentscheidungen.

Tabelle 1
Interpretationen von "Signifikanzniveau"

Am Ende versichert der Autor seinen Lesern: "All of these are different ways to say the same thing". Ist das Buch von Nunally auch ein Extremfall, nach Durchsicht aktueller Lehrbücher bleibt festzustellen: Weder Fisher noch Neyman und Pearson wären mit der inkonsistenten Vermischung ihrer Ideen in den meisten dieser Lehrbücher einverstanden gewesen.

Mythos 2: Signifikanztests sind wissenschaftlich unantastbar

Der Streit um die Signifikanztests, der bereits von Fisher auf der einen Seite und Neyman und Pearson auf der anderen Seite bis hin zu beleidigenden Äußerungen geführt wurde, war nur der Beginn der wissenschaftlichen Auseinandersetzung um diese Methode. Falls Lehrbücher diesen Streit überhaupt erwähnen, erwecken im weiteren Verlauf oft den Eindruck, als ob dieser Streit durch die spätere Forschung geklärt und beigelegt werden konnte. Die derzeit in der Psychologie geführte sogenannte „Signifikanztestdebatte" (einen aktuellen Überblick gibt Nickerson, 2000) zeigt, dass das Gegenteil der Fall ist: Es wird sowohl das mathematische Prinzip der Signifikanztests, ihr Stellenwert in der Wissenschaft als auch ihr gedankenloser Gebrauch kritisiert. Wir wollen im Folgenden aus der Fülle der Literatur zu dieser Debatte einige Artikel aufführen, deren Titel oftmals schon die Verzweiflung erkennen lassen, mit der die Autoren das zwanghafte und oftmals sinnlose Signifikanztesten anprangern:

Carver, R.P. (1978): The case against statistical significance testing.
Loftus, G.R. (1991): On the tyranny of hypothesis testing in the social sciences.
Gigerenzer, G. (1993b): Über den mechanischen Umgang mit statistischen Methoden.
Cohen, J. (1994): The earth is round ($p < .05$).
Loftus, G.R. (1994): Why psychology will never be a real science until we change the way we analyze data.
Dar, R., Serlin, D., & Omer, H. (1994): Misuse of statistical tests in three decades of psychotherapy research.
Falk, R., & Greenbaum, W. (1995): Significance tests die hard.

Wir glauben, dass Studenten, die einen oder mehrere dieser Arbeiten gelesen haben, einen viel sichereren Umgang mit Signifikanztests entwikkeln und eine richtige Vorstellung von der Bedeutung eines signifikanten Testergebnisses erhalten. Eine umfassende Einführung in die historische Entwicklung das Hypothesentestens einschließlich der Kontroversen gibt „Das Reich des Zufalls" (Gigerenzer et. al., 1999). Wir glauben, dass ein Blick auf die widersprüchlichen Ansichten über Signifikanztests Schüler und Studenten nicht etwa verunsichern würde (was gelegentlich vermutet wird), sondern dass Statistik dadurch ein interessantes und spannendes Gesicht bekäme. Festgestellt werden kann: Das derzeitige Schweigen der Lehrbücher zu den strittigen Punkten der Signifikanztests scheint der Verständlichkeit dieser Verfahren nicht förderlich gewesen zu sein.

Mythos 3: Signifikanztests sind die einzige inferenzstatistische Methode

Abgesehen von der Berechnung von Konfidenzintervallen und Effektgrößen (siehe hierzu z.B. Sedlmeier, 1996) ist vor allem die Bayesstatistik ein weiterer inferenzstatistischer Zugang neben dem Testen auf Signifikanz. Die Bayesstatistik ermöglicht es, Aussagen über die Wahrscheinlichkeiten von Hypothesen zu machen, was mit Signifikanztests nicht möglich ist. Gerade das ist aber die häufigste Fehlinterpretation eines signifikanten Testergebnisses: nämlich die Meinung, es wurde die Wahrscheinlichkeit einer Hypothese ermittelt. Dieses „Bayesianische Wunschdenken" ist in Lehrbüchern und in den Köpfen von Studenten und Forschern weit verbreitet. Seltsamerweise ist der warnende Hinweis: „Mit einem Signifikanztest kann man *nichts* über die Wahrscheinlichkeit einer Hypothese herausfinden" aber kaum in Lehrbüchern zu finden. Zur Klärung dieses Punktes könnte man ganz entscheidend beitragen, wenn man die Bayesstatistik und deren Grundprinzip im Unterricht zumindest erwähnen würde. Mit folgender, einfacher Gegenüberstellung würde beispielsweise schnell klar werden, welcher inferenzstatistische Zugang was leisten kann:

(a) $P(D \mid H_0)$ wird bei Nullhypothesentests berechnet (Fisher)
(b) $P(D \mid H_0)$ und $P(D \mid H_1)$ werden bei Neyman und Pearson verglichen

(c) $P(H|D)$ wird bei Bayesschen Testverfahren, und nur dort, ermittelt

Als einziges erlaubt es also die Bayesstatistik (c), Wahrscheinlichkeiten von Hypothesen (H) zu bestimmen, während man mit Nullhypothesentests (Fisher) die Wahrscheinlichkeit seiner Daten (D) unter Zugrundelegung der H_0 betrachtet[7] (a). Neyman und Pearson (b) fordern neben (a) die Betrachtung einer weiteren Wahrscheinlichkeit, nämlich die der vorliegenden Daten unter einer konkreten Alternativhypothese H_1. Wir denken, dass die bloße Erwähnung aller drei Paradigmen und deren Gegenüberstellung Missinterpretationen vermeiden helfen könnte – egal welches dieser Paradigmen letztlich nun unterrichtet werden soll (für einen konkreten Implementationsvorschlag eines Vergleichs von (a) und (c) im Schulunterricht siehe Krauss & Wassner, 2001).

Die Empfehlungen des US-Amerikani- schen National Council of Teachers of Mathematics (die NCTM-Standards) machen deutlich, dass deren Verfasser es auf keinen Fall gutheißen würden, dass, wie es heute oft der Fall ist, die Alternativen zum Signifikanztesten in der Schule totgeschwiegen werden: Sie fordern, dass Schüler in Mathematik grundsätzlich in der Lage sein sollten „to select and use various types of reasoning and proof as appropriate" und „to apply a wide variety of strategies". Lehrkräfte sollten bemüht sein, „various approaches to solve problems" zu diskutieren (NCTM, 2000). Engel (2000) schreibt dazu: „Es ist dabei nützlich, dass Schüler an Problemen arbeiten, die auf vielfältige Weise gelöst werden können. Dadurch wird ein Vergleich von Methoden möglich, ebenso wie das Herstellen von Querbezügen, Diskussionen über Denkprozesse und Strategien und eine Reflexion über vorgeschlagene Vorgehensweisen." Wir finden, dass die beurteilende Statistik genau solche Probleme liefern kann, und wir sollten Schülern mehr als nur *ein* statistisches Werkzeug vorstellen, mit dem man diese Probleme bearbeiten kann. Erst wenn Schülern Einblick in die ganze Werkzeugkiste gewährt wird, können sie die Bedeutung und den Nutzen der einzelnen Werkzeuge richtig verstehen.

Obwohl alle drei Mythen nicht immer *explizit* unterrichtet werden, scheinen sie sich doch in den Köpfen von Schülern und Studenten festgesetzt zu haben. Der heute übliche Statistikunterricht an Schule und Universität scheint also alle drei Mythen *implizit* zu unterstützen. Bis jetzt haben wir als Folgen dieser Mythen lediglich die Verwirrung bei der Interpretation eines signifikanten Testergebnisses genannt. Es lassen sich aber noch weitere Folgeschäden, die vor allem in der Wissenschaft zu beobachten sind, finden.

"Signifikanz" als mechanischer Beweis

Dieses Problem hängt eng mit den erwähnten Fehlinterpretationen zusammen. Wer nicht genau weiß, was ein signifikantes Testergebnis bedeutet, schreibt einem solchen Ergebnis eher eine zu große Bedeutung zu. Als Folge davon werden beispielsweise Medikamente eingeführt, die in Wahrheit gar nicht wirken. Die beiden Mediziner Beck-Bornholdt und Dubben (1998) haben sich nach jahrelanger Analyse medizinischer Fachzeitschriften ein Bild zu diesem Thema gemacht. Unter der Überschrift „Die unerträgliche Leichtigkeit der Signifikanz" schreiben sie: „In diesem Kapitel werden Sie erfahren, wie man zu belanglosen 'statistisch signifikanten' Zufallsergebnissen kommt, wie eine Unzahl von Zufallstreffern die Wissenschaft bis zum Absaufen verwässert und was man dagegen unternehmen kann."

Es gibt natürlich auch den umgekehrten Effekt: Ein erfolgversprechendes neues Medikament wird als *definitiv* unwirksam eingestuft, weil ein Experiment den P-Wert 0,06 liefert. Die American Association of Psychology (APA) schreibt beispielsweise in ihrem Publication Manual 1974: „Caution: Do not infer trends from data that fail by a small margin to meet the usual levels of significance. Such results are best interpreted as caused by chance and are best reported as such". (S. 19). Diese bürokratische Einstellung, sowie auch Meltons Meinung über publikationswerte Manuskripte (siehe oben), werden heute allerdings eher belächelt. Rosnow and Rosenthal (1989) schreiben zu diesem Thema: „Surely, God loves the .06 nearly as much as the .05".

Wir denken, man sollte Schülern und Studenten die Willkürlichkeit eines konventionellen Signifikanzniveaus (1% bzw. 5%) klar machen. Tatsäch-

[7] Genau steht D dabei für die vorliegenden oder noch unwahrscheinlichere Daten. Mit der Schreibweise als bedingte Wahrscheinlichkeit – $P(D|H_0)$ – legen wir uns auf Punkthypothesen fest (mehr hierzu siehe Krauss & Wassner, 2001).

lich scheint Fisher sich nur deshalb auf diese Werte als konventionelles Signifikanzniveau festgelegt zu haben, weil er von seinem Konkurrenten Karl Pearson (dem Vater von Egon Pearson) keine Erlaubnis bekommen hatte, dessen Tabellen, die auch andere Werte beinhalteten, nachzudrucken (Gigerenzer, 1998). Anhand von Tabelle 1 wird weiterhin klar, dass diese konventionelle Willkür nur bei Fishers früherer Interpretation des Signifikanzniveaus überhaupt eine Rolle spielt.

Ritualisierte Anwendung

Studenten sagen, sie müßten das Ritual des Signifikanztests verwenden, weil ihre Diplomarbeit sonst nicht akzeptiert würde. Doktoranden sagen, sie müßten dies tun, da ihre Dissertation sonst nicht angenommen würde. Professoren sagen, sie hätten keine Wahl, da die Herausgeber der Fachzeitschriften Signifikanztest verlangten. Durch Kommentare wie die von Melton (1962) oder durch Aussagen wie im APA Publication Manual (1974) besteht tatsächlich die Gefahr, dass Wissenschaft gegängelt, ja geradezu manipuliert wird durch das alleinige Ziel, signifikante Ergebnisse zu erzielen. Eine solche Haltung erschwert es, nach anderen, geeigneteren Methoden Ausschau zu halten (oft kann zum Beispiel eine gute Graphik einen Signifikanztest ersetzen[8]). Die Schuldzuweisung der Doktoranden an die Professoren, und der Professoren an die Herausgeber ist aber nicht die ganze Geschichte. G.R. Loftus (1993a), damals Herausgeber von *Memory and Cognition*, berichtet, dass selbst nach seinem „editorial comment", in dem er verkündete, er wolle in seiner Zeitschrift keine unnötigen Signifikanztests mehr sehen, die meisten Autoren unbeeindruckt weiterhin ihre Artikel mit Signifikanztests einreichten und sich weigerten, diese durch adäquatere Methoden (wie z.B. Abbildungen mit „error bars") zu ersetzen. Der Ausdruck „zwanghaftes Ritual" scheint also durchaus keine Übertreibung zu sein.

[8] Siehe hierzu Loftus (1993b): A picture is more worth than thousand p-values: On the irrelevance of hypothesis testing in the computer age.

3. Back to school: Verbesserungsvorschläge für den Stochastikunterricht

Welche der genannten Probleme treffen auch auf den Schulunterricht zu? Und: Welche Änderungsvorschläge ergeben sich daraus für den Stochastikunterricht?

Auch wenn in Schulbüchern mathematisch meist etwas sorgfältiger gearbeitet wird als in Statistik-Kompendien für Psychologie- und Pädagogikstudenten, denken wir, dass im Unterricht sehr ähnliche Probleme vorliegen. Schulbücher vermeiden zwar explizite Fehldeutungen eines signifikanten Testergebnisses, aber sie tragen auch nicht zur Ausrottung dieser Fehldeutungen bei. Die Geschichte der Stochastik spielt sich meist nur in der Form von Namensnennungen ab, Konflikte zwischen widersprüchlichen Ansichten werden auch hier meist totgeschwiegen[9].

In der didaktischen Literatur wird schon seit einiger Zeit auf Probleme mit den Signifikanztests aufmerksam gemacht. M. Buth (1991) führt in der Zeitschrift *Stochastik in der Schule* aus, dass man doch eigentlich gerne etwas über die Wahrscheinlichkeit der Hypothese erfahren würde, wenn man sie denn schon testet. Die Frage ist laut Buth, ob Signifikanztests nicht gar den gesunden Menschenverstand behindern, da sie etwas liefern, was man eigentlch gar nicht wissen möchte. In der psychologischen Literatur wurde diese „Behinderung des gesunden Menschenverstandes" durch ein (mittlerweile berühmt gewordenes) Zitat von J. Cohen (1994, S. 997) zum Ausdruck gebracht:

> *What's wrong with significance testing? Well, among many other things, it does not tell us what we want to know, and we so much want to know what we want to know that, out of desperation, we nevertheless believe in that it does! What we want to know is "Given these data, what is the probability that H_0 is true?" But as most of us know, what it tells us is "Given that H_0 is true, what is the probability of these (or more extreme) data?"*

[9] Ein gutes Buch, allerdings für Studenten der Hochschulmathematik, ist das von U. Krengel (1990). Krengel entwickelt die gesamte Testtheorie am Streit zwischen Fisher und Neyman/Pearson.

Auch an der Vormachtstellung der Signifikanztests in der Schule wird in der didaktischen Literatur gerüttelt: Wickmann (1998, bzw. 2001a und b, in diesem Heft) fordert eine Neuorientierung, da die Signifikanztests in vielen Anwendungssituationen unzulänglich und inadäquat seien.

Selbst die Idee, Schülern Falschaussagen zur Bedeutung eines signifikanten Ergebnisses vorzulegen, findet sich bereits in der Didaktik der Stochastik (siehe R. Diepgen[10], 1985a). Diepgens Artikel sowie die von I. Birnbaum (1982) berichteten Schülerzitate zum Begriff „Signifikanz" zeigen, dass Probleme beim Verständnis des Signifikanzkonzepts durchaus auch im Schulunterricht zu finden sind. Ein weiterer Artikel von Diepgen (1985b) in *mathematik lehren* mit dem Titel „Signifikanz - Na und?" zeigt auf sehr unterhaltsame Weise, warum ein signifikantes Testergebnis an sich erst einmal noch nichts bedeuten muss. Diepgen gibt in diesem Artikel kritische Variablen an, die bei der *Bewertung* eines signifikanten Ergebnisses unbedingt beachtet werden müssen. So ist zum Beispiel den meisten Schülern und Studenten unbekannt, dass ab einer gewissen Stichprobengröße sehr leicht „alles signifikant wird". Obwohl Diepgens Artikel durchaus auf Schulniveau geschrieben ist und auch interessante Übungsaufgaben bereitstellt, werden solche Einsichten dennoch nur selten in den Schulunterricht einbezogen.

Eine Tatsache ist weiterhin, dass Schüler (ebenso wie Studenten) Schwierigkeiten haben, bei einer Statistik-Textaufgabe das angemessene statistische Verfahren zu finden. Das heißt, dass Schüler genau in den beiden Bereichen, in denen statistisches Denken gefordert wäre, nämlich bei der *Wahl* des Verfahrens und bei der *Interpretation* des Ergebnisses, Schwierigkeiten haben. Was Schüler lernen, ist Einsetzen und Ausrechnen. Das kann und darf aber nicht das Ziel des Stochastikunterrichts sein.

Eine Aufhebung dieser Missstände im Schulunterricht wäre übrigens alles andere als Selbstzweck. Diepgen (1985) fordert, dass den Missständen in den Sozialwissenschaften so früh wie möglich, also bereits im Stochastikunterricht, entgegengewirkt werden müsse: „Den gängigen Irrationalismen statistischer Alltagspraxis vorzubeugen sollte daher meines Erachtens wichtige Aufgabe eines wissenschaftspropädeutischen und kritischen Statistikunterrichtes auf der Schule sein." (S.32).

Derzeit haben jedoch auch in der Schule statistische Rituale und Rezeptbuchstatistik den Vorrang vor dem Versuch, Schülern statistisches Denken und den Umgang mit Unsicherheit beizubringen. Die Fähigkeit, statistisch zu denken, ist aber sowohl im Alltag als auch in der Wissenschaftspraxis dringend von Nöten.

Was kann man im Unterricht besser machen?

- *Von Spannungen und Kontroversen berichten:* Der heutige Unterricht gibt ein schiefes Bild von der Stellung und der Aussagekraft der Signifikanztests. Man sollte Signifikanztests im Unterricht zur Diskussion stellen. Das bedeutet, dass man Schülern von dem spannenden und aufschlussreichen Streit um Signifikanztests erzählen sollte, und zwar beginnend von Fisher gegen Neyman/Pearson bis zum heutigen Diskussionsstand (siehe Gigerenzer et al., 1999, Kap. 3). Eine Unterrichtsdiskussion über den Streit um Signifikanztests ist auch aus motivationalen Gründen zu empfehlen, da Lernende bei dieser Thematik erfahrungsgemäß mit großem Interesse dabei sind.
- *Von Alternativen berichten:* Man sollte Alternativen zum Signifikanztesten im Unterricht ansprechen, denn nur dadurch kann frühzeitig ein ritualisierter Gebrauch der Signifikanztests verhindert werden. Die Einführung verschiedenartiger Verfahren gewährleistet nicht nur einen Einblick in den „statistischen Werkzeugkasten", sondern durch die Gegenüberstellung dieser Verfahren wird auch klar, welche Aussagekraft die einzelnen Verfahren haben und was sie *nicht* leisten können (siehe dazu Krauss & Wassner, 2001).
- *Statt Kochbuchrezepte statistisches Denken lehren:* Dies würde bedeuten, den Schwerpunkt des Unterrichts auf *Auswahl des Verfahrens* und *Interpretation der Ergebnisse* zu legen. Das mechanische Rechnen zwischendrin sollte eher selbstverständlich sein. Diese Forderung wird nachdrücklich sowohl durch die Folgerungen aus der TIMS-Studie als auch durch die NCTM-Standards 2000 unterstützt. In den NCTM-Standards ist zum Beispiel zu lesen: „Die Schüler sollen lernen, ihr *eigenes*

[10] Diepgen testete seine Schüler mit 2 Fragen (nämlich Aussage 2 und 5 von Oakes).

mathematisches Denken beim Problemlösen selbst zu kontrollieren und zu reflektieren." Die TIMS-Studie beklagt, dass die deutschen Schüler zwar gut im Reproduzieren von Routineverfahren sind, aber schlecht bei der Anwendung des Gelernten auf neue Probleme, beim mathematischen Modellieren, beim Herstellen von Querverbindungen und beim aktiven, nicht-standardisierten Argumentieren und zeigte dadurch ebenfalls die Notwendigkeit einer Wende vom Unterrichten von statistischen Ritualen zum Unterrichten von statistischem Denken auf.

- *Anbindung an die reale Forschungspraxis*: Folgender Punkt wurde bislang noch nicht angesprochen: Signifikanztests werden in der Schule gewöhnlich als „pure Mathematik" gelehrt. Die wirklichen Probleme bei der Anwendung von Signifikanztests bei einer realen wissenschaftlichen Problemstellung werden nicht angesprochen. Buth (1993) gibt in seinem Artikel „Was man aus der Forschungpraxis für die Schule lernen kann" ein Anwendungsbeispiel, das man den Schülern nicht vorenthalten sollte. Realistische Beispiele eignen sich nicht nur zur Verwirklichung von fächerübergreifendem Unterricht (z.B. Mathematik-Biologie, Mathematik-Wirtschaft), sondern können auch zur Motivation von Schülern beitragen. Die überall geforderte Anbindung von Mathematik an die Realität (NCTM-Standards, TIMS-Studie) könnte insbesondere in Stochastik bestens verwirklicht werden!

3 Schlussbemerkung

W. Krämer (1991) vermutet in seinem Buch „So lügt man mit Statistik", dass in einer durchschnittlichen Zeitung mehr Statistiken zu finden sind, als Goethe oder Schiller in ihrem ganzen Leben sahen, da heutzutage „ein Trommelfeuer von Daten, Zahlen, Fakten, Tabellen, Kurven, Trends und Tests den modernen Medienkonsumenten eindeckt". Krämer stellt sogar die These auf, dass „Prozent" mittlerweile das häufigste Substantiv der deutschen Sprache ist (Krämer, 1991, S. 51). Wie dem auch sei, wir denken, dass eine Wende hin zum Unterrichten von „statistischem Denken" nicht nur Durchblick im „Alltagsdschungel der Prozente" gewährleisten würde, sondern auch die Voraussetzung für ein Überleben im „Wissenschaftsdschungel statistischer Schlussfolgerungen" ist. Dass dies nicht nur für Schüler von Bedeutung ist, die später die wissenschaftliche Laufbahn einschlagen wollen, sondern für Bildung und Praxis überhaupt, wird leider oft verkannt. Viele Psychologiestudenten stellen sich z.B. auf den Standpunkt: „Ich möchte sowieso klinischer Psychologe werden. Wo brauche ich später bei der Durchführung von Psychotherapien schon statistisches Denken?" Aber welche Psychotherapiemethoden werden in der Ausbildung gelehrt? Falls sich Therapeuten auf die Methoden verlassen, die die aktuelle klinische Forschung vorschlägt, sollten sie zumindest sehr vorsichtig sein, denn diese Forscher haben zur Überprüfung ihrer Hypothesen wahrscheinlich Signifikanztests verwendet. Würden Studenten den Artikel „Misuse of statistical tests in three decades of psychotherapy research" (Dar, Serlin, & Omer, 1994) kennen, würden sie einen solchen Standpunkt wahrscheinlich noch einmal überdenken.

Der Stochastikunterricht sollte deshalb die Chance nicht verpassen, die Signifikanztests auch in einen Zusammenhang mit wissenschaftlichem Schlussfolgern zu bringen. Auf der Tagung des AK Stochastik 2000 wurde gelegentlich davon gesprochen, dass der Stochastikunterricht des neuen Jahrhunderts eine Wende hin zum statistischen Denken erfahren wird. Da außerdem die nächste Tagung 2001 den Begriff „Statistisches Denken" im Titel trägt, besteht also durchaus Anlass zu Hoffnung. Vielleicht werden Schüler zukünftiger Abiturjahrgänge in der Tat statt der mechanischen Reproduktion von Signifikanztests in der Schule statistisches Denken und den Umgang mit Unsicherheit lernen.

Literatur:

Beck-Bornholdt, H.P. & Dubben, H.H. (1997). *Der Hund, der Eier legt. Erkennen von Fehlinformationen durch Querdenken.* Rowohlt Taschenbuch Verlag Reinbek bei Hamburg.

Birnbaum, I. (1982). *Die Interpretation statistischer Signifikanz* (übersetzt von Günter Fillbrunn), Stochastik in der Schule, 2, 3.

Buth, M. (1991). Zur Behinderung des gesunden Menschenverstandes durch Stochastik, *Stochastik in der Schule*, 3, S. 12-22.

Buth, M. (1993). Zum Thema Testen von Hypothesen: Was man aus der Forschungspraxis für die Schule lernen kann, *Stochastik in der Schule*, 2. S. 34-46.

Carver, R.P. (1978). The case against statistical significance testing. *Haward Educational Review, Vol 48, 3*, p.378-399.

Cohen, J. (1994). The earth is round (p < .05). *American Psychologist, 49*, 997-1003.

Dar, R., Serlin, D., & Omer, H. (1994). Misuse of statistical tests in three decades of psychotherapy research. *Journal of Consulting and Clinical Psychology, 62*, 75-82.

Diepgen, R. (1985a). Was Schüler zum Hypothesentesten wissen sollten, *Stochastik in der Schule*, 1, S. 32-36.

Diepgen, R. (1985b). Signifikanz – Na und? Politik und Computer im Statistikunterricht. *mathematik lehren, 13*, S. 54-57.

Engel, J. (2000). Die NCTM-Standards: Anstösse für den Mathematikunterricht nach TIMSS. *Zentralblatt für Didaktik der Mathematik* (3), 69-74.

Falk, R., & Greenbaum, W. (1995). Significance tests die hard. *Theory & Psychology, 5*, 75-98.

Fisher, R.A. (1925). *Statistical methods für research workers*. Edinburgh: Oliver and Boyd.

Fisher, R.A. (1935). *The design of experiments*. Edinburgh: Oliver and Boyd.

Fisher, R.A. (1956). *Statistical methods and scientific inference*. Edinburgh: Oliver and Boyd.

Gigerenzer, G. (1993a). The superego, the ego, and the id in statistical reasoning. In G. Keren & C. Lewis (Eds.), *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 E. Roth (Ed.), *Sozialwissenschaftliche Methoden*. München: R. Oldenbourg Verlag.

Gigerenzer, G. (1998). We need statistical thinking, not statistical rituals. *Behavioral and Brain Sciences, 21*, p. 199-200.

Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J., & Krüger, L. (1999). *Das Reich des Zufalls*. Heidelberg: Spektrum.

Krämer, W. (1991). *So lügt man mit Statistik*. Campus Verlag Frankfurt/Main.

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

Krengel, U. (1990). *Einführung in die Wahrscheinlichkeitstheorie und Statistik*. Vieweg.

Loftus, G.R. (1991). On the tyranny of hypothesis testing in the social sciences. *Contemporary psychology, 36*, Vol. 36, 2, p.102-105.

Loftus, G.R. (1993a). Editorial comment. *Memory and cognition, 21, 1*, p. 1-3.

Loftus, G.R. (1993b): A picture is more worth than thousand p-values: On the irrelevance of hypothesis testing in the computer age. *Behavior Research Methods, Instrumentation and Computers, 25*, 250-256.

Loftus, G.R. (1994). Why psychology will never be a real science until we change the way we analyze data. Paper presented at Psychonomic Society meetings, Washington DC.

Melton, A.W. (1962). Editorial. *Journal of Experimental Psychology, 64*, 553-557.

NCTM - National Council of Teachers of Mathematics (2000). *Principles and Standards for school mathematics*. Reston, Va.: National Council of Teachers of Mathematics.

Nickerson, R.S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy, *Psychological Methods, 5*, 241-301.

Nunally, J.C. (1975). *Introduction to statistics for psychology and education*. New York: McGraw-Hill.

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

Rosnow, R., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist, 44*, 1276-1284.

Sedlmeier, P. (1996). Jenseits des Signifikanztest-Rituals: Ergänzungen und Alternativen. *Methods of Psychological Research Online, Vol. 1, No. 4*, S. 41-63.

Wickmann, D. (1998). Zur Begriffsbildung im Stochastikunterricht, *JDM 19*, S. 78.

Wickmann, D. (2001a). Der Theorieneintopf ist zu beseitigen. *In diesem Heft*.

Wickmann, D. (2001b). Inferenzstatistik ohne Signifikanztest. *In diesem Heft*.