WAS SCHÜLER ZUM HYPOTHESENTESTEN WISSEN SOLLTEN

von R. DIEPGEN, Bochum

Der mathematischen Statistik wurde ein Platz im gymnasialen Mathematikcurriculum eingeräumt, weil diese Statistik - im Unterschied zu den meisten anderen Gebieten der Mathematik - Anwendung findet nicht nur in physikorientierter Naturwissenschaft und Technik, sondern gleichermaßen auch in Disziplinen wie Medizin, Biologie, Ökonomie, Pädagogik, Psychologie, Soziologie, Politikwissenschaften usw.. Deshalb sollte der Statistikunterricht auf der Schule Propädeutik für die Anwendung insbesondere in diesen (vor allem human-)wissenschaftlichen Disziplinen betreiben. Dies ganz besonders auch deshalb, weil die wissenschaftstheoretische und methodenkritische Diskussion der letzten Jahre, etwa die sog. Signifikanztestdebatte in Psychologie und Soziologie, deutlich gemacht hat, daß bislang die inferenzstatistischen Standardverfahren vor allem in den "mathematikfernen" Disziplinen häufig ohne Verständnis ihrer Logik, sondern lediglich orientiert an unverstandener Konvention ("Rezeptbuchstatistik") angewandt und dabei in ihren Ergebnissen oft miß-, zumeist überinterpretiert werden. Den gängigen Irrationalismen statistischer Alltagspraxis vorzubeugen sollte daher meines Erachtens wichtige Aufgabe eines wissenschaftspropädeutischen und kritischen Statistikunterrichtes auf der Schule sein.

Im folgenden sei anhand einiger Fragestellungen demonstriert, was in diesem Sinne Schüler zum Hypothesentesten wissen sollten, genauer zum Standardverfahren Signifikanztest. Diese Fragestellungen beziehen sich lediglich auf den Binomialtest (Vorzeichentest), d.h. den Test des Parameters einer Binomialverteilung; dieser von der mathematischen Struktur her einfachste Signifikanztest dürfte nämlich Bestandteil eines jeden Unterrichtes über Hypothesentesten sein. Mehr als auf die richtigen Antworten kommt es bei diesen Fragen an auf die stichhaltigen Begründungen für die Antworten. (Diese Fragen habe ich im Rahmen eines lehrzielorientierten Testes zur Evaluation einer Unterrichtsreihe über Hypothesentesten eingesetzt.)

Frage 1: a) Ein Signifikanzniveau von 5 % bedeutet, daß wir, wenn

wir die Nullhypothese zurückweisen, dies mit einer Wahrscheinlichkeit von 5 % zu Unrecht tun. Stimmt dies?

b) Wenn die Nullhypothese auf dem Signifikanzniveau von 1 % zurückgewiesen wird, so bedeutet dies, daß mit 99 % Wahrscheinlichkeit die Alternativhypothese richtig ist. Stimmt dies?

Kommentar: Beide Fragen sind zu verneinen. Sie gehen von einem weitverbreiteten Unverständnis über die Logik des Signifikanzkonzeptes aus, ein Unverständnis, das zu einer Überschätzung der Leistungsfähigkeit von Signifikanztests führen dürfte. Beide Formulierungen unterstellen nämlich, daß ein signifikantes Ergebnis etwas über die Wahrscheinlichkeit von Null- oder Alternativhypothese aussagt; hier wird verkannt, daß das Signifikanzkonzept lediglich die Wahrscheinlichkeit einer Fehlentscheidung unter der Bedingung der Geltung der Nullhypothese betrifft und damit keineswegs einen Rückschluß auf die Wahrscheinlichkeit der Nullhypothese selbst erlaubt. Ja, die herkömmliche Theorie der Signifikanztests kennt überhaupt nicht den Begriff der Wahrscheinlichkeit einer Hypothese, im Gegensatz etwa zur Inferenzstatistik in der Tradition von BAYES. Der logischen Struktur des Signifikanzkonzeptes und seiner Grenzen sollten sich die Schüler bewußt sein, um die Aussagekraft signifikanter Ergebnisse angemessen einschätzen zu können.

Frage 2: Es soll ein neu entwickeltes, sehr teures Medikament gegen eine bislang zu 90 % tödlich verlaufende Krankheit an einer Zufallsstichprobe von 20 Patienten getestet werden. (Bislang gab es keinerlei Therapie gegen diese Krankheit.)

Schädliche Wirkungen des Medikamentes sind so gut wie ausgeschlossen. Getestet wird also (einseitig) die Nullhypothese, daß die Überlebenswahrscheinlichkeit bei Einnahme des Medikamentes dieselbe ist wie bei Nichteinnahme des Medikamentes.

Welches Signifikanzniveau erscheint Ihnen eher angemessen: 5 % oder 10 %?

Kommentar: Hier geht es um die Begründung der Wahl eines Signifikanzniveaus, um die sich statistische Alltagspraxis häufig herum-

Stochastik in der Schule, Heft 5 Bd. 5 (1985)

drückt. Dazu gilt es, die Folgen eines Fehlers 1. Art, also der irrtümlichen Verwerfung der Nullhypothese, zu vergleichen mit den Konsequenzen eines Fehlers 2. Art, der fälschlichen Beibehaltung der Nullhypothese. Folge eines Fehlers 1. Art dürfte hier die Geldverschwendung für den Einsatz eines in Wahrheit wirkungslosen teuren Medikamentes sein. Ein Fehler 2. Art bedeutet, daß todkranken Patienten das lebensrettende Heilmittel vorenthalten wird, weil man es irrtümlich für wirkungslos hält. Verschwendetes Geld versus vermeidbaren Sterbens: hier hat offensichtlich der Fehler 2. Art besonderes Gewicht, wobei aber auch ein Fehler 1. Art ziemlich unerfreuliche Konsequenzen hat. Daher erscheint es sinnvoll, durch liberale Wahl des Signifikanzniveaus von 10 % ein höheres Risiko eines Fehler 1. Art in Kauf zu nehmen, um das Risiko eines Fehlers 2. Art möglichst zu mindern. Die begründete Wahl eines Signifikanzniveaus setzt also die Fähigkeit zur Reflexion der Konsequenzen beider möglichen Fehlentscheidungen voraus sowie natürlich das Wissen, daß die (bedingte) Wahrscheinlichkeit eines Fehlers der einen Art nur zu senken ist auf Kosten der (bedingten) Wahrscheinlichkeit eines Fehlers der anderen Art.

Frage 3: Eine Untersuchung an einer Zufallsstichprobe von rund 10.000 Personen, die regelmäßig eine Ausdauersportart wie Jogging, Radfahren usw. betreiben, hat ergeben, daß die Auftretenshäufigkeit von Erkrankungen des Herz-Kreislauf-Systems in dieser Personengruppe signifikant (bei einem Signifikanzniveau von 5 %) geringer ist als in der Gesamtbevölkerung. (Negative gesundheitliche Auswirkungen der Sportübung seien ausgeschlossen.)

Kommt Ihrer Meinung nach diesem Testergebnis eine praktische Bedeutung für die Gesundheitspolitik zu, so daß etwa eine aufwendige Kampagne für das Betreiben von Ausdauersportarten allein aufgrund dieses Befundes sinnvoll erscheinen könnte?

Kommentar: Hier geht es darum, daß statistische Signifikanz allein noch wenig bedeutet. Denn die extrem große Stichprobe und damit die extrem hohe Testgüte in der skizzierten Untersuchung würde zu einem signifikanten Ergebnis schon dann führen, wenn sich die Erkrankungshäufigkeiten zwischen Gesamtbevölkerung und Sportlern auch nur in

nahezu verschwindend geringem Ausmaß unterscheiden würden; die aufwendige Kampagne wäre aber wohl überhaupt erst dann zu erwägen, wenn es einen nennenswerten Unterschied gäbe. Es kommt hier also zusätzlich an auf die absolute Größe des Unterschiedes in der Erkrankungshäufigkeit, worüber nichts ausgesagt ist; die Signifikanz alleine besagt praktisch noch nichts. Diese Fragestellung zielt auf die weitverbreitete Unsitte in den Humanwissenschaften, sich lediglich für die statistische Signifikanz von Ergebnissen zu interessieren. Wo sich wissenschaftliche Anerkennung ausschließlich an statistischer Signifikanz orientiert, da degeneriert diese Signifikanz zu einem Maß für den empirischen Aufwand des Forschers, der ja so gut wie sicher sein kann, daß seine einfache Nullhypothese ganz exakt nicht stimmen wird. Legion sind die Untersuchungen, in denen an gigantischen Stichproben signifikante Ergebnisse erzielt wurden, die praktisch und wissenschaftlich völlig irrelevant sind.

Frage 4: Manche Herausgeber humanwissenschaftlicher Fachzeitschriften verfahren nach folgendem Kriterium für die Aufnahme empirischer Arbeiten: Veröffentlicht wird eine statistisch-empirische Untersuchung nur dann, wenn sie ein signifikantes Ergebnis erbracht hat, d.h. wenn die Nullhypothese eines Testes aufgrund der Daten verworfen werden konnte. Was halten Sie von dieser Vorgehensweise der Herausgeber?

Kommentar: Diese letzte Frage ist etwas anspruchsvoller, wird sie doch überhaupt erst brisant, wenn man den gedanklichen Rahmen der Signifikanz eines einzelnen Experimentes verläßt. Orientiert am Konzept der statistischen Überprüfung einer Hypothese in einem einzelnen Experiment nämlich erscheint die skizzierte, in der Realität übrigens tatsächlich vorherrschende Herausgeberphilosophie plausibel; denn Idee des Signifikanztestes ist es doch gerade, sich erst dann für eine Forschungshypothese zu entscheiden, und das bedeutet natürlich, sie als wissenschaftliche Aussage zu veröffentlichen, wenn die Ergebnisse des Prüfexperimentes signifikant, also "überzufällig" von dem abweichen, was man unter der zur Forschungshypothese konträren Nullhypothese erwarten würde. Zumeist beschäftigt sich aber mit einer inhaltlichen Fragestellung insbesondere in den Humanwissenschaften jeweils unabhängig voneinander eine ganze Reihe

von Forschern; zu einer Forschungshypothese werden daher häufig unabhängig voneinander mehrere Experimente gemacht. Nehmen wir nun an, daß eine in Wirklichkeit zutreffende Nullhypothese weltweit in 20 Experimenten jeweils zum Signifikanzniveau von 5 % überprüft wird. Erwartungsgemäß wird von diesen Experimenten im Rahmen der gewählten Wahrscheinlichkeit für einen Fehler 1. Art genau eines signifikant; in den anderen 19 Experimenten wird die Nullhypothese zu Recht beibehalten. Die skizzierte Herausgeberphilosophie führt aber in diesem Falle unsinnigerweise dazu, daß genau das eine "irrtümlich" signifikante Experiment veröffentlicht wird; die anderen Experimente werden von den Forschern, die sich ja auf die Herausgeberkriterien von vornherein einstellen, erst gar nicht zur Veröffentlichung eingereicht. Denkt man dies überspitzt zu Ende, so könnte die Fixierung auf Signifikanz als Kriterium die Fachzeitschriften zu einer Sammlung zufälliger Irrtümer machen. Signifikanz als Bewertungsmaßstab in der wissenschaftlichen Kommunikation ist also beileibe nicht ganz unproblematisch.

Ich hoffe, diese vier Fragen konnten exemplarisch andeuten, worum es in einem wissenschaftspropädeutischen Unterricht über statistisches Hypothesentesten gehen sollte. Abschließend seien ganz kurz als didaktische Konsequenz drei thematische Schwerpunkte für den Unterricht genannt: die entscheidungstheoretische Logik des Signifikanztestes, die Bedeutung des Fehlers 2. Art und der Testgüte einschließlich ihrer Abhängigkeit von Signifikanzniveau und Stichprobengröße, schließlich die wissenschaftstheoretische Bedeutung und wissenschaftspraktische Rolle des Signifikanztestes.