Probleme eines Statistikunterrichtes nach STRICK-Muster

von Raphael Diepgen, Bochum

Während ein Großteil mathematischer Verfahren bislang vor allem nur von physikorientierter Naturwissenschaft und Technik benötigt wird, verlangen nach statistischen Kenntnissen heute nahezu alle Disziplinen mit empirischem Selbstverständnis, also auch Disziplinen wie Pädagogik, Psychologie, Soziologie, Ökonomie, Medizin, Biologie. Gerade dieses breite Anwendungsspektrum hat der beurteilenden Statistik einen Platz im gymnasialen Mathematikcurriculum verschafft. Nach der zunächst euphorischen Übernahme inferenzstatistischer Verfahren in den Methodenkanon vor allem humanwissenschaftlicher Disziplinen hat indessen die wissenschaftstheoretische und methodenkritische Diskussion der letzten Jahre, etwa die sog. Signifikanztestdebatte in Soziologie und Psychologie - siehe zum Überblick etwa BREDENKAMP (1972) oder WITTE (1980) -, deutlich gemacht, daß die inferenzstatistischen Standardverfahren häufig ohne jedes Verständnis ihrer Logik, sondern lediglich in Orientierung an unverstandener Konvention angewandt und in ihren Ergebnissen miß- und in der Regel überinterpretiert werden. Aufgabe eines Statistikunterrichtes auf der Schule sollte es daher meiner Meinung nach sein, einer irrationalen statistischen Praxis im Forschungsalltag vorzubeugen, indem Logik und zugleich Problematik inferenzstatistischer Verfahren ausführlich thematisiert werden; die Vermittlung rechentechnischer Fertigkeiten im Umgang mit inferenzstatistischen Standardverfahren allein kann daher meines Erachtens nicht Ziel eines wissenschaftspropädeutischen Statistikunterrichtes sein, zumal diese Fertigkeiten durch die Verbreitung entsprechender Computerprogramme zunehmend überflüssig werden. Vielmehr sollte der Unterricht die argumentativen Grundlagen für einen aufgeklärten und kritischen Umgang mit dem statistischen Instrumentarium vermitteln.

Nun leidet aber der Unterricht über schließende Statistik unter einem gleich zweifachen Ausbildungsdefizit der Mathematiklehrer: Die meisten von ihnen haben sich während ihres Studiums nicht mit der Theorie mathematischer Statistik beschäftigt. Und noch viel weniger vertraut sind sie mit der inferenzstatistischen Praxis außerhalb der Mathematik, mit jenen Anwendungen also, auf die hin mathematische Statistik überhaupt formu-

liert ist. In einer solchen Situation bekommen Schulbücher hervorragende Bedeutung für den Unterricht. Denn es lernen aus ihnen die Theorie - und wozu sie gut sein soll - nicht nur die Schüler, sondern zunächst einmal die Lehrer. Erklärlich daher, daß ein Schulbuch Verbreitung findet besonders dann, wenn es beurteilende Statistik leicht verständlich und zugleich eingebettet in ihre Anwendungen darzustellen verspricht. Dies Versprechen bietet von der Aufmachung her insbesondere das Buch "Einführung in die Beurteilende Statistik" von STRICK (1980), dessen didaktische Konzeption vom Autor mehrfach dargestellt wurde (STRICK 1978, 1979, 1981a, 1981b). Es sei daher - stellvertretend auch für andere Schulbücher zur Statistik - näher betrachtet.

Intention von STRICK ist es, in einem einsemestrigen Stochastikkurs durch die radikale Beschränkung auf einfachste wahrscheinlichkeitstheoretische Voraussetzungen breiten Raum zu lassen für die beurteilende Statistik, d.h. für Konfidenzintervalle und Hypothesentesten. Dies gelingt STRICK, indem er Inferenzstatistik reduziert auf das Grundmuster der Bereichsschätzung für Binomialverteilungen bei großen Stichproben. Dabei nutzt er die Approximation der Binomialverteilung an die Normalverteilung - natürlich ohne dies auszuführen, sondern nur an Beispielen illustrierend - zur Gewinnung der zentralen Aussage, daß beim n-stufigen BERNOULLI-Versuch mit hinreichend großem n die relative Häufigkeit der Erfolge mit einer Wahrscheinlichkeit von ca. 95,5 %, der sog. "Sicherheitswahrscheinlichkeit", in der $2\sigma/n$-Umgebung um die Erfolgswahrscheinlichkeit p liegt (S. 57, 60). Mit dieser einfachen Aussage werden nun beim sog. "Schluß von der Gesamtheit auf die Stichprobe für eine bekannte Erfolgswahrscheinlichkeit p" alle Stichprobenergebnisse innerhalb der $2\sigma/n$-Umgebung von p als "verträglich mit der Erfolgswahrscheinlichkeit p" bezeichnet (S. 60). So weit, so gut: Denn offensichtlich hat die Rede von der Verträglichkeit eines Stichprobenergebnisses mit einer bekannten Erfolgswahrscheinlichkeit einen Sinn; ich kann auf lange Sicht bei Ziehungen aus der Gesamtheit in 95,5 % der Fälle mit solchen Stichprobenergebnissen rechnen, d.h. mein Verhalten darauf einstellen, ich kann auf das Auftreten solcher Ergebnisse mit bestimmten Einsätzen wetten usw.. Der Begriff der Sicherheitswahrscheinlichkeit ist hier ohne Problem interpretierbar.

Der Einstieg in die schließende Statistik liest sich dann aber folgendermaßen: "In den meisten Fällen kennt man jedoch die Erfolgswahrscheinlich-

keit nicht, sondern kennt das Ergebnis einer Stichprobe. Aufgaben von diesem Typ heißen: Schluß von der Stichprobe auf die Gesamtheit. Die Grundfrage dieser Aufgaben lautet: Welche Erfolgswahrscheinlichkeit liegt dem Zufallsversuch zugrunde? Wir suchen alle Erfolgswahrscheinlichkeiten p, die mit dem Stichprobenergebnis verträglich sind. In unserem Lösungsansatz gehen wir also davon aus, daß das Stichprobenergebnis innerhalb der $2\sigma/n$-Umgebung um p liegt. Dies gilt mit einer Sicherheitswahrscheinlichkeit von 95,5 %." (S. 64). Die Menge aller dieser mit dem Stichprobenergebnis verträglichen Parameter p bekommt dann auf der folgenden Seite den Namen Konfidenzintervall. Nur der nachdenkliche Leser wird bei diesen glatten Formulierungen stutzig: Denn bislang kennt er nur den Begriff der Verträglichkeit eines Stichprobenergebnisses mit einer (bekannten) Erfolgswahrscheinlichkeit p. Was heißt aber die Verträglichkeit einer (unbekannten) Erfolgswahrscheinlichkeit p mit einem beobachteten Stichprobenergebnis, von dem hier plötzlich geredet wird? Kann ich, in Umkehrung des Schlusses von der Gesamtheit auf die Stichprobe, nun bei Vorliegen eines Stichprobenergebnisses mit bestimmten Einsätzen darauf wetten, daß die Erfolgswahrscheinlichkeit im Konfidenzintervall liegt? Oder kann etwa ein Forscher damit rechnen, daß auf lange Sicht - etwa über sein wissenschaftliches Leben - die von ihm ermittelten Konfidenzintervalle zu 95,5 % jeweils die wahren Parameter enthalten? Bedeutet die Sicherheitswahrscheinlichkeit von 95,5 % dafür, daß "das Stichprobenergebnis innerhalb der $2\sigma/n$-Umgebung um p liegt", wenn ich gar nicht weiß, um welches p es sich handelt, nicht dasselbe wie die Wahrscheinlichkeit, daß p innerhalb der $2\sigma/n$-Umgebung um das beobachtete Stichprobenergebnis liegt? Wie gesagt: Nur der aufmerksame Leser wird sich diese Fragen stellen, aber er wird sie aufgrund seiner Lektüre kaum beantworten können und daher schließlich nicht wissen, was denn der Sinn von Konfidenzintervallen ist, warum man ihnen vertrauen sollte. Der flüchtige und weniger nachdenkliche Leser dürfte es hier besser haben: Er wird einem Konfidenzintervall vertrauen, weil er glaubt, es enthalte mit 95,5 %-iger Wahrscheinlichkeit den wahren Parameter. Daß zu dieser zweiten Lesergruppe durchaus auch Mathematiklehrer gehören, die schon jahrelang mit diesem Schulbuch arbeiten, zeigt die Fruchtlosigkeit der Warnung vor dieser Fehlinterpretation im Lehrerbegleitheft zum Schulbuch (S. 49); denn diese Warnung ist nicht verbunden mit einer positiven Interpretation, die das Vertrauen in ein Konfidenzintervall

rechtfertigen würde. STRICKs suggestiver Übergang vom wohldefinierten und wohlinterpretierten Begriff der Verträglichkeit eines Stichprobenergebnisses mit einer Erfolgswahrscheinlichkeit p zum lediglich rein formal rekonstruierbaren, die ganze inferenzstatistische Problematik aber berührenden Begriff der Verträglichkeit einer Erfolgswahrscheinlichkeit p mit einem Stichprobenergebnis legt das skizzierte Mißverständnis natürlich nahe. Wer beim Inferenzschluß von einer Sicherheitswahrscheinlichkeit redet, ohne ausdrücklich hervorzuheben, daß diese Wahrscheinlichkeit unter der Bedingung steht, daß in Wahrheit p eine bestimmte Größe hat - was man leider nicht weiß -, der braucht sich nicht darüber zu wundern, daß der Leser zwar rechentechnische Fertigkeit in der Bestimmung von Konfidenzintervallen erwirbt, im Grunde aber nicht versteht, welchen Sinn und welche Begründung dieses Tun hat.

Ähnliche Unklarheiten begleiten die Darstellung des Hypothesentestens bei STRICK (S. 71). Der Test einer Nullhypothese H_0 erscheint als Regel, H_0 zu verwerfen, wenn das Stichprobenergebnis nicht mit der hypothetischen Erfolgswahrscheinlichkeit unter H_0 verträglich ist. Warum indessen die fehlende Verträglichkeit die Verwerfung von H_0 rechtfertigt, dies bleibt im Dunkeln. Mißverständlich auch hier (S. 70) die Bezeichnung der "Wahrscheinlichkeit für eine falsche Entscheidung (hier 4,5 %) als Irrtumswahrscheinlichkeit", so als stände diese Wahrscheinlichkeit nicht unter der Bedingung der Geltung von H_0 und als gäbe es nur falsche Entscheidungen in Form des Fehlers 1. Art, nicht aber in Form des Fehlers 2. Art. Rein formal werden zwar später (S. 74) beide Fehlerarten unterschieden, dies aber ohne irgendwelche Folgen. Weder wird jemals von der (bedingten) Wahrscheinlichkeit für einen Fehler 2. Art gesprochen - es fehlt also gänzlich das Konzept der Testgüte -, noch wird die gegenläufige Beziehung zwischen beiden Fehlerwahrscheinlichkeiten und damit zwischen Signifikanzniveau und Testgüte erörtert, eine Beziehung, die erst ein sinnvolles Abwägen bei der Wahl des Signifikanzniveaus erlauben würde. Der Charakter des Hypothesentestens als eines Entscheidungsverfahrens, dessen sinnvoller Einsatz ein Abwägen zwischen beiden Fehlentscheidungsmöglichkeiten erfordert, bleibt undeutlich. Dem Leser wird die Logik vorenthalten, die dem Hypothesentesten überhaupt eine Begründung als sinnvolles Entscheidungsverfahren verleiht. So dürfte er kaum zu einem begründeten Umgang mit dem Testinstrumentarium geführt werden, eher zu den vielen Un-

sinnigkeiten falschverstandener statistischer Alltagspraxis: Er wird die Wahl eines Signifikanzniveaus nicht legimitieren können, denn er weiß ja gar nicht, was sie ihm bringt und was sie ihn kostet. Er wird wahrscheinlich der Unsitte verfallen, erst nach der Inspektion der Daten das Signifikanzniveau festzulegen, denn er hat den Charakter des Tests als eines im Vorhinein definierten Entscheidungsverfahrens kaum verstanden. Er wird mit Tests, bei denen die Nullhypothese mit der Forschungshypothese übereinstimmt, mit Anpassungstests also, nichts anfangen können, oder er wird sie wie normale Tests behandeln. Er wird signifikante Ergebnisse für wichtig halten und hochsignifikante Ergebnisse für noch wichtiger, seien die Stichproben auch noch so groß und die Abweichungen von der Nullhypothese auch noch so klein und wissenschaftlich wie praktisch völlig irrelevant. Er wird nicht die Gefahr verstehen, daß Signifikanz in der Forschungspraxis degenerieren kann zu einem Maß für den empirischen Aufwand, den der Forscher bereit ist zu betreiben, der ja so gut wie sicher sein kann, daß seine Nullhypothese exakt ohnehin nicht gelten wird. Denn dem Leser fehlt das entscheidende Konzept der Fehlerwahrscheinlichkeit 2. Art und Testgüte.

Fazit: Die didaktisch zunächst reizvolle Idee von STRICK, die formale und rechnerische Äquivalenz zwischen Konfidenzintervallen und Hypothesentests einerseits, der Bereichsschätzung bei bekanntem Parameter andererseits zu nutzen in Form eines einfachen Lösungsschemas für alle statistischen Fragestellungen, hat den Nachteil, daß die spezifische Logik der Inferenzstatistik - und damit zugleich auch ihre Problematik - verschleiert wird. Der Schüler lernt die rezeptartige Berechnung von Konfidenzintervallen und Hypothesentests, hier ja immer nur die Ermittlung von σ-Umgebungen, er wird aber meines Erachtens kaum dazu in die Lage versetzt, dieses Tun überzeugend zu begründen. Er erlernt sein simples "Strick"-Muster, bestenfalls vielleicht eine gewisse Skepsis gegenüber veröffentlichten Untersuchungsergebnissen, eine Skepsis, die er kaum wird in klare Worte fassen können. Allein, diese sicher sinnvolle Skepsis könnte der Unterricht - so würde ich vermuten - viel einfacher erzeugen durch einige Demonstrationsbeispiele mithilfe von Simulation.

Stellvertretend wird hier das Buch von STRICK kritisiert für all jene Schulbücher, die auf das Konzept der Fehlerwahrscheinlichkeit 2. Art oder Testgüte verzichten. Diese Schulbücher verharren damit auf

dem allenfalls heuristischen Argumentationsniveau des ursprünglichen Signifikanzgedankens, klassisch formuliert etwa in FISHERs "Problem of the lady tasting tea", der Idee nämlich, eine isolierte einfache Hypothese als wenig plausibel zu verwerfen allein deshalb, weil in einem Experiment Ereignisse auftreten, die bei Geltung der Hypothese unwahrscheinlich sind. Das argumentative Ungenügen dieser allzu simplen Idee wurde erst behoben durch die Entwicklung der Testtheorie, etwa von NEYMAN und PEARSON, also durch die Betrachtung der Fehlerwahrscheinlichkeit 2. Art. Wenn Schulbücher darauf verzichten, so bleiben sie dem Leser eine ausreichende Begründung schuldig, warum denn bloß das als "Hypothesentesten" vorgeschlagene und eintrainierte Entscheidungsverhalten vernünftig sein soll. Ohne Kenntnis des Konzepts der Testgüte scheint mir schließlich eine aufgeklärte und kritische statistische Praxis schlechterdings unmöglich. Auf eine solche Praxis aber sollte der schulische Statistikunterricht propädeutisch abzielen.

Warum ich - abgesehen von der Möglichkeit zum Wortspiel - unter vergleichbaren Schulbüchern gerade jenes von STRICK ausgewählt habe? Nun, das Einzigartige beim Konzept von STRICK ist, daß hier der Lehrer auch nicht mehr im Nachhinein die Möglichkeit hat, durch zusätzliche Betrachtung von Fehlerwahrscheinlichkeiten 2. Art die Argumentationslücken zu schließen. Denn dank der Reduzierung der Wahrscheinlichkeitsrechnung einzig auf die Berechnung von σ-Umgebungen hin ist es dem Schüler im Statistikunterricht nach STRICK gänzlich unmöglich, die Wahrscheinlichkeit zu berechnen, daß ein Stichprobenergebnis in die $2\sigma/n$-Umgebung um die hypothetische Erfolgswahrscheinlichkeit p_0 fällt, wenn in Wahrheit eine andere Erfolgswahrscheinlichkeit p_1 (mit $p_1 \neq p_0$) vorliegt. Letztlich erweist es sich an dieser Stelle als folgenreiche Fehlentscheidung von STRICK (1981a, S. 10), dem Konfidenzintervall bloß seiner rechnerischen und formalen Universalität wegen didaktische Priorität vor dem Hypothesentesten zugesprochen zu haben. Ich würde die Prioritäten andersherum setzen, im Hinblick sowohl auf die Interpretierbarkeit, als auch auf die Bedeutung im Wissenschaftsalltag.

Ich weiß nicht, ob es in einem nur einsemestrigen Stochastikgrundkurs möglich ist, angemessen beurteilende Statistik darzustellen. Sollte dazu die Zeit fehlen, dann erschiene es mir allerdings sinnvoller, die Wahrscheinlichkeitsrechnung wenigstens so ausführlich zu betreiben, daß

darauf aufbauend die Universität Statistik angemessen darstellen könnte.

LITERATUR

BREDENKAMP, J., 1972: Der Signifikanztest in der psychologischen Forschung. Frankfurt

STRICK, H. K., 1978: "Mathematische Statistik. Vorschlag für einen einsemestrigen Grundkurs." Lernzielorientierter Unterricht 4, 28-37

STRICK, H. K., 1979: "Parameterschätzung und Hypothesentesten im einsemestrigen Grundkurs Wahrscheinlichkeitsrechnung/Statistik. In: Beiträge zum Mathematikunterricht 1979. Vorträge. Hannover, S. 356-359

STRICK, H. K., 1980: Einführung in die Beurteilende Statistik. Hannover

STRICK, H. K., 1981a: "Die Bestimmung von Konfidenzintervallen im Grundkurs Wahrscheinlichkeitsrechnung/Statistik." MNU 34, 7-11

STRICK, H. K., 1981b: "Methoden der beurteilenden Statistik im Grundkurs Stochastik." In: DÖRFLER, W., FISCHER, R. (Hrsg.): Stochastik im Schulunterricht. Wien, S. 245-248

WITTE, E. H., 1980: Signifikanztest und statistische Inferenz. Analysen, Probleme, Alternativen. Stuttgart

Anmerkung zu R. DIEPGEN: Probleme eines Statistikunterrichtes nach STRICK-Muster

von H. K. Strick, Leverkusen

Schon vor einigen Heften hatten wir dazu aufgerufen, in kritischer Form zu unseren Beiträgen Stellung zu nehmen. Dies ist bisher leider nicht geschehen. Der vorstehend abgedruckte Aufsatz setzt sich in kritischer Weise mit der Konzeption von Stochastikkursen auseinander - er weist auf gefährliche Stellen und Schwachpunkte insbesondere meines Grundkursbuches hin. Ich sehe diese gefährlichen Stellen in meiner Konzeption ebenfalls; wenn ich auch meine, daß bei konsequenter Unterrichtsführung Mißverständnisse nicht auftreten sollten. Daß in einem einsemestrigen Grundkurs nicht alle Fragen der Beurteilenden Statistik erarbeitet werden können, ist offensichtlich. Ob man aus diesem Grunde auf eine "Einführung" (und mehr biete ich nicht an) verzichten sollte, darüber könnte man streiten. Bei der Konzeption meines Kurses erschien es mir als wichtigste Idee, die "Gesetzmäßigkeiten" des Zufalls - bei BERNOULLI-Versuchen - zu vermitteln, um damit dem Schüler einen Einblick in Schätz- und Testverfahren zu gewähren.