Die Behinderung des gesunden Menschenverstandes durch Stochastik

von *Manfred Buth*, Hamburg

Kurzfassung: Beim Testen von Hypothesen stellt sich eine Diskrepanz ein zwischen den offiziellen Verfahren der Stochastik und den Fragen, die dem gesunden Menschenverstand einleuchten. Dieser Unterschied wird an einem überschaubaren Beispiel ausführlich erörtert. Schließlich werden Folgerungen für den Unterricht gezogen.

Problemlage

Im Stochastikunterricht sollen die Schüler in die Eigenart stochastischen Denkens eingeführt und damit vor den vielfältigen Irrtümern und Fehlern bewahrt werden, die dem Laien im Umgang mit dem Zufall leicht unterlaufen. Der Stochastikunterricht verfolgt daher ausgesprochen emanzipatorische Ziele. Aber er bringt auch unbeabsichtigte Nebenwirkungen mit sich, die praktisch auf eine Behinderung des gesunden Menschenverstandes hinauslaufen. Wie das gemeint ist, soll an einem Beispiel aus dem Bereich des Testens von Hypothesen erläutert werden.

Den Ausgangspunkt bilde eine reale Situation, deren praktische Bedeutung unmittelbar einleuchtet: Zwei Fußballmannschaften entscheiden durch das Werfen einer Münze, welche Partei die Seitenwahl haben soll. Da alle Spieler an einem gerechten Verfahren interessiert sind, vereinbaren sie, daß die Münze vorher zehnmal hochgeworfen und genau dann für gut befunden werden soll, wenn sie mindestens dreimal und höchstens siebenmal Wappen zeigt. Dabei wird die Münze als gut definiert, wenn die Wahrscheinlichkeit für das Werfen von Wappen eben so groß ist wie für das Werfen von Zahl (vgl. Athen u.Griesel, 1979).

Stochastik in der Schule 11 (1991), Nr. 3, 12-22.

Aus dem ganzen inhaltlichen Zusammenhang geht hervor, daß einzig und allein das Problem interessiert, *ob* die Münze gut ist und mit welcher Wahrscheinlichkeit sich die Entscheidung für oder gegen die Münze aufgrund des Testergebnisses absichern läßt. Die Lehrer aber fragen im Unterricht umgekehrt und lassen die Schüler ausrechnen, mit welcher Wahrscheinlichkeit die Münze den Test besteht, *wenn* sie gut ist. Also braucht man sich nicht zu wundern, wenn die Schüler die einzig sinnvolle Frage bearbeiten und die gestellte Aufgabe beiseite lassen. Antworten der folgenden Art sollten deshalb nicht überraschen: "Wir haben die Hypothese auf dem 5% Signifikanzniveau verworfen, also gilt die Alternative mit 95% iger Sicherheit" (Riemer, 1986).

Didaktiker sind über derartige Ergebnisse entrüstet: "Solche vom Lernenden immer wieder ausgespochenen Sätze dokumentieren ein tiefes Unverständnis der Testtheorie, welches auf Versäumnisse in früher liegenden Stufen des Stochastikunterrichts liegen muß" (Riemer, 1986, vgl. auch Wiedling, 1979). Aber ist das wirklich zutreffend? Liegt das tiefe Unverständnis nicht vielmehr auf Seiten der Didaktiker? Denn sie verschweigen den Schülern, daß die einzig interessante Frage, *ob* die Münze gut ist, keineswegs so einfach zu beantworten ist, wie die Schüler denken, und sie sind dennoch darüber erstaunt, daß die belanglose Frage, was geschieht, *wenn* die Münze gut ist, von den Schülern beiseite geschoben wird.

Welche Schwierigkeiten in der Aufgabenstellung enthalten sind, zeigt bereits eine einfache und noch relativ grobe Überlegung: Wenn ich in mein Portemonnaise greife und eine Münze hervorhole, dann kann ich mit ziemlicher Sicherheit behaupten, die Münze sei schlecht. Denn der Begriff der guten Münze wurde oben durch die Gleichheit zweier Wahrscheinlichkeiten definiert. Da man die Verteilung der p-Werte, d.h. der Wahrscheinlichkeiten für das Auftreten von Wappen, praktisch als kontinuierlich ansehen darf, kann es kaum vorkommen, daß die Münze exakt den Wert $p = 0.5$ besitzt. Wenn die Münze aber mit Sicherheit nichts taugt, dann sind die Ergebnisse der Testtheorie irrelevant, weil sie auf der Voraussetzung beruhen, daß die Münze gut sei.

Auch der Einwand, es sei unvernünftig, bei einer annähernd kontinuierlichen Verteilung der p-Werte einen einzelnen Wert herauszugreifen, zählt nicht. Denn so lange die Verteilung der p-Werte unbekannt ist, hat man kein vernünftiges Kriterium für die Auswahl eines Intervalls, innerhalb dessen die Münze noch als gut gelten kann.

Analyse eines Beispiels

Um das Beispiel mit dem Münzwurf noch konkreter zu fassen und einer gründlichen Untersuchung zugänglich zu machen, seien vier Münzen mit den p-Werten 20%, 50%, 50% und 80% gegeben, wobei p die Wahrscheinlichkeit für das Auftreten von Wappen bezeichne. Die Münzen werden durch Glücksräder simuliert, deren gefärbte Flächen jedoch anders als in Fig. 1 nach unten gekehrt und somit verdeckt sein sollen (vgl. auch Birnbaum u. Fillbrunn, 1984).

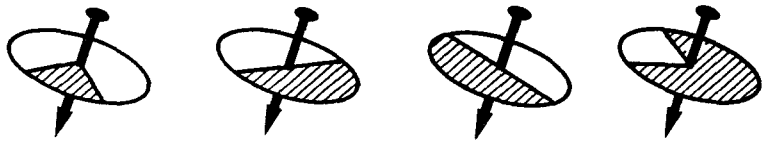

Fig. 1

Die Wahrscheinlichkeitsverteilung der p-Werte ist in Fig. 2 graphisch dargestellt, sie lautet

$$v(p) = \begin{cases} 0.25 & \text{für } p = 0.2 \\ 0.50 & \text{für } p = 0.5 \\ 0.25 & \text{für } p = 0.8 \end{cases}$$

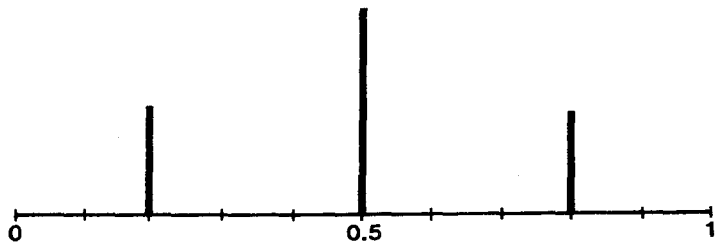

Fig. 2

Der ganze Komplex der Seitenwahl beim Fußballspiel einschließlich der Überprüfung der Münze läßt sich als ein umfangreiches Zufallsexperiment beschreiben: Wähle eine der vier Münzen aus und wirf sie zehnmal hoch. Der Wahrscheinlichkeitsraum wird so gewählt, daß die Elementarereignisse durch den p-Wert der herausgegriffenen Münze und die Zahl k für das Auftreten von Wappen gekennzeichnet sind. Der Wahrscheinlichkeitsraum enthält demnach 33 Elemente und ist in Fig. 3 symbolisch dargestellt.

$p\backslash k$	0	1	2	3	4	5	6	7	8	9	10
0.2											
0.5											
0.8											

Fig. 3

Die Wahrscheinlichkeit $w(p,k)$ für das Auftreten des Elementarereignisses mit den Parametern p und k beträgt

$$w(p,k) = \binom{10}{k} p^k (1-p)^{10-k} v(p)$$

Die Funktionswerte sind in Tabelle 1 wiedergegeben. Da die Funktion w von zwei Parametern abhängt, muß man sich die Stäbe in Fig. 4 über den Mittelpunkten der Rechtecke von Fig. 3 errichtet denken.

Tabelle 1: Die Wahrscheinlichkeit w(p,k) (in Promille)

p/k	0	1	2	3	4	5	6	7	8	9	10
0.2	27	67	76	50	22	7	1	0	0	0	0
0.5	0	5	22	59	103	122	103	59	22	5	0
0.8	0	0	0	0	1	7	22	50	76	67	27

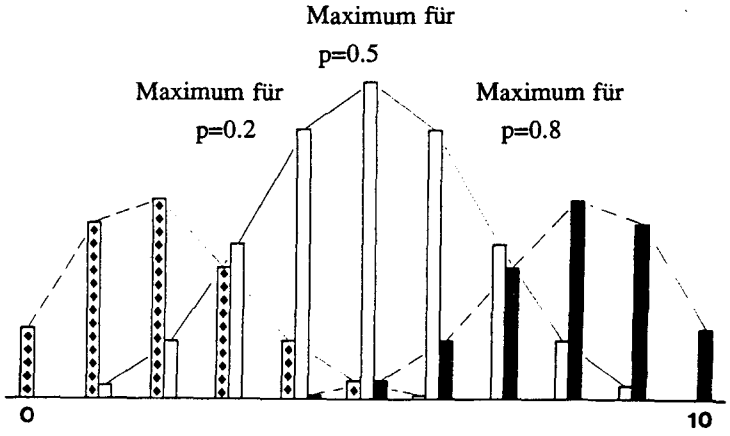

Fig.4

In diesem sehr konkreten Modell kann man einige Rechnungen durchführen, deren Ergebnisse dazu beitragen, die Schwierigkeiten beim Testen von Hypothesen besser zu verstehen.

Die erste Aufgabe lautet: Wenn der Test positiv ausgefallen ist, d.h. wenn er $3 \le k \le 7$ Köpfe ergeben hat, und man sich folglich dazu entschlossen hat, der gewählten Münze den Wert p zuzuordnen, wie sicher kann man dann sein, die richtige Entscheidung getroffen zu haben? Die Lösung der Aufgabe führt auf die bedingte Wahrscheinlichkeit w(p|+), bei der + für den positiven Ausfall des Münztests steht. Um w(p|+) zu bestimmen, muß man das Ereignis 'Die Münze hat den Wert p' innerhalb des eingeschränkten Wahrscheinlichkeitsraums berechnen, der zu $3 \le k \le 7$ gehört und in Fig. 5 schraffiert dargestellt ist.

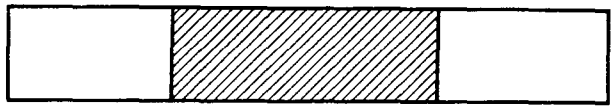

Fig. 5

Die Rechnung ergibt

$$\begin{aligned} w(p=0.2\,|\,+) &= 13\% \\ w(p=0.5\,|\,+) &= 74\% \\ w(p=0.8\,|\,+) &= 13\% \end{aligned} \qquad (1)$$

Für das Interesse der Fußballspieler an einer guten Münze und damit an einem gerechten Verfahren bedeutet der Test eine Erhöhung der Sicherheit. Denn, während man bei der blinden Wahl einer Münze nur eine Chance von 50% für das Greifen einer guten Münze hat, erhöht sich die Wahrscheinlichkeit nach dem Test auf 74%.

Die Frage nach den Chancen für ein positives Testergebnis bei vorgegebenem p-Wert führt auf die bedingte Wahrscheinlichkeit w(+|p). Um sie zu bestimmen, muß man die Wahrscheinlichkeit für einen positiven Testausgang in dem eingeschränkten Wahrscheinlichkeitsraum berechnen, der in Fig. 6 für $p = 0.5$ schraffiert dargestellt ist.

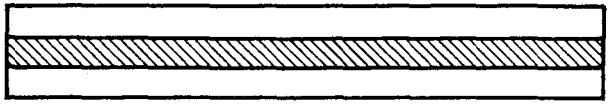

Fig. 6

Die Rechnung ergibt

$$\begin{aligned} w(+\,|\,p=0.2) &= 32\% \\ w(+\,|\,p=0.5) &= 89\% \\ w(+\,|\,p=0.8) &= 32\% \end{aligned} \qquad (2)$$

Die bedingten Wahrscheinlichkeiten in (2) unterscheiden sich von denen in (1) dadurch, daß die Bedingung und das Bedingte miteinander vertauscht sind. Die

Werte in (1) und die Werte in (2) stehen über den Satz von Bayes zueinander in Beziehung. Es gilt

$$w(p \mid +) = \frac{w(+ \mid p)\, v(p)}{w(+ \mid 0.2)\, v(0.2) + w(+ \mid 0.5)\, v(0.5) + w(+ \mid 0.8)\, v(0.8)} \qquad (3)$$

Aus den drei Werten von (2) ergeben sich auch die üblichen Ergebnisse der Testtheorie. Der Fehler erster Art, also die Wahrscheinlichkeit, die Münze abzulehnen, obwohl sie in Wirklichkeit gut ist, beträgt

$$a = 1 - w(+ \mid p{=}0.5) = 11\% \qquad (4)$$

Der Fehler zweiter Art gibt an, mit welcher Wahrscheinlichkeit man die Münze akzeptiert, obwohl sie in Wirklichkeit schlecht ist. Dieser Fehler beträgt

$$b = \frac{w(+ \mid 0.2)\, v(0.2) + w(+ \mid 0.8)\, v(0.8)}{v(0.2) + v(0.8)} = 32\% \qquad (5)$$

Testtheorie und Bayes-Formel

Bei den Rechnungen des vorigen Abschnitts wurde die Verteilung der p-Werte als bekannt vorausgesetzt. Was bleibt von den Ergebnissen übrig, wenn man von dieser Kenntnis absieht? Der Vergleich von Fragestellungen, die mit einem statistischen Test behandelbar sind, und jenen, die im Zusammenhang mit der Bayes-Formel möglich werden, wirft ein klares Licht auf die Beschränktheit der Testtheorie. Im Rahmen des statistischen Testens stellt man sich auf den Standpunkt, daß die Verteilung der Werte von p (das ist $v(p)$) nicht bekannt ist und auch nicht wie üblich mittels eines Experimentes über die beobachteten relativen Häufigkeiten gemessen werden kann.

Dann aber läßt sich nur ein Fehler 1. Art berechnen. Schon der Fehler 2. Art entartet nun zu einer sogenannten Gütefunktion. Die Wahrscheinlichkeit eines Fehlers 2. Art hängt vom jeweiligen Wert von p ab. Ein globaler Fehler 2. Art, gemittelt über alle Werte von p aus der Alternativhypothese (wie in (5)) kann nicht berechnet werden. Schließlich kann die Wahrscheinlichkeit einer falschen Zuordnung, das ist $w(p{\neq}0.5 \mid +)$ nicht mehr berechnet werden, weil mangels der Kenntnis der Verteilung der Werte von p die Bayes-Formel (3) nicht mehr

verwendet werden kann. Fazit: Die interessanten Fragen beim Testen einer Münze können nicht beantwortet werden, und die berechenbaren Ergebnisse sind belanglos.

Lassen sich die Probleme beim Verständnis der Testtheorie durch den Satz von Bayes lösen?

Als ein Ausweg aus der Schwierigkeit, die Verteilung der Werte von p nicht wirklich zu kennen, bleibt noch eine Methode, die sehr an die Art und Weise erinnert, wie sich der Baron von Münchhausen am eigenen Schopf aus dem Sumpf gezogen hat.

Man kennt die Verteilung der Werte von p nicht und wählt deshalb willkürlich eine Schätzung der Werte $v(p{=}0.2)$, $v(p{=}0.5)$ und $v(p{=}0.8)$, von der man meint, daß sie ganz gut paßt. Um die Willkür einer solchen Wahl zu minimieren, korrigiert man diese Verteilung durch die Bayes-Formel (3), indem man nacheinander Daten einbringt. Also, man wirft eine Münze und ermittelt, je nach dem, welches Ergebnis M der Wurf gebracht hat, mit Hilfe von (3) und anhand der Schätzwerte von $v(p)$ die bedingten Wahrscheinlichkeiten $w(p{=}0.2 \mid M)$, $w(p{=}0.5 \mid M)$ und $w(p{=}0.8 \mid M)$. Die Ergebnisse werden durch die Festlegung

$$v(p) = w(p \mid M)$$

als eine verbesserte Schätzung für die Verteilung der p-Werte herangezogen. Nun wirft man die Münze nochmals und verbessert damit wiederum die Schätzung für die Verteilung der p-Werte. Das Ergebnis, das man nach zehn Würfen ermittelt hat, wird dann als Grundlage gewählt, um über die Annahme oder Ablehnung der Münze zu entscheiden. In Tabelle 2 sind die Ergebnisse einer Rechnung zusammengestellt, die nach dieser Methode durchgeführt wurde. Die anfängliche Schätzung

50 25 25

für die Verteilung der p-Werte ist willkürlich festgesetzt und wurde so ausgewählt, daß ihre Korrektur durch den nachfolgenden Test besonders deutlich ins Auge fällt. Um die bedingten Wahrscheinlichkeiten $w(p \mid W)$ zu bestimmen, beginnt man mit $w(W \mid p)$. Das ist die Wahrscheinlichkeit für das Auftreten von

Wappen unter der Vorausssetzung, daß Wappen mit der Wahrscheinlichkeit p auftritt. Also gilt $w(W|p) = p$. Der Satz von Bayes liefert

$$w(p|W) = \frac{w(W|p)\, v(p)}{w(W|0.2)\, v(0.2) + w(W|0.5)\, v(0.5) + w(W|0.8)\, v(0.8)}$$

Daraus ergeben sich mit Hilfe der geschätzten Verteilung die gesuchten Werte

$$w(p=0.2 \mid W) = 24\%$$
$$w(p=0.5 \mid W) = 29\%$$
$$w(p=0.8 \mid W) = 47\%$$

Sie führen über die Gleichung $v(p) = w(p|W)$ zu der verbesserten Schätzung, die in Zeile 2 der Tabelle aufgeführt ist und die wiederum den Ausgangspunkt für weitere Korrekturen bildet.

Die am Beispiel noch einmal erläuterte Methode, über die der Leser bei Riemer (1985) nähere Ausführungen findet, bildet eine Mathematisierung dessen, was man als das Lernen aus Erfahrung bezeichnet. Sie kann zugleich als ein sinnvoller Vorschlag für den Unterricht gelten. Aber sie hat mit der üblichen Testtheorie wenig zu tun, weil sie andere Fragen behandelt. Wenn man indirekt den Vergleich zwischen dem Testen und dem Lernen aus Erfahrung methodologisch diskutiert, kann man daraus neue Einsichten für das Testen gewinnen. Allerdings ist die Methode insofern anfechtbar, als man nicht weiß, wann man genug Daten einbezogen hat, damit man eine ausreichend genaue Information über die Verteilung der Werte von p erhält. Mathematisch gesehen greift

Indiz	Schätzwerte in %		
	50	25	25
W	24	29	47
Z	44	34	22
W	20	39	40
W	7	35	58
Z	17	50	33
W	6	46	48
Z	13	61	26
Z	22	67	11
W	10	71	19
W	4	67	29

Tab.2: 'Lernen aus Erfahrung'

man bei der Stabilisierung der korrigierten Verteilung auf das Gesetz der großen Zahlen zurück.

Zusammenfassung

Vom Testen einer Hypothese verspricht sich der gesunde Menschenverstand eine Hilfestellung bei der Entscheidung, ob man eine Vermutung annehmen oder zurückweisen soll. Dazu kann die Testtheorie aber wenig beitragen, weil die erforderliche Wahrscheinlichkeitsverteilung der p-Werte im allgemeinen nicht bekannt ist. Es kann also nur die ziemlich belanglose Wahrscheinlichkeit berechnet werden, mit der man einen Irrtum riskiert, wenn man sich gegen eine Hypothese ausspricht, die in Wirklichkeit zutrifft. Nicht einmal der Fehler zweiter Art läßt sich berechnen.

Mein Vorschlag zur Vermeidung der üblichen Verständnisschwierigkeiten beim Testen von Hypothesen besteht darin, den Schülern genau diesen Sachverhalt zu erklären, der hier auf der Ebene der didaktischen Diskussion erörtert wurde und der zu Unterrichtszwecken noch methodisch aufbereitet werden müßte. Dabei kann der Computer gute Dienste leisten. Man sollte ihn einsetzen, um durch vielfältige Variation einiger geeigneter Beispiele möglichst viel Anschauungsmaterial bereitzustellen.

Die Schätzverfahren mit Hilfe der Regel von Bayes mögen zwar eine interessante und zugleich didaktisch kluge Alternative für den Umgang mit Hypothesen bilden. Ob man damit auch schon die übliche Testtheorie besser versteht, mag bezweifelt werden. Denn der Kalkül der Bayes-Formel wirkt sehr formal und behindert daher tiefere Einsichten, wie Formel (3) die korrigierte Verteilung der Werte von p berechnet. Da würde eine andere Darstellung des Kalküls (siehe Borovcnik, 1986) weitgehend Abhilfe schaffen.

Für Anregungen und Hinweise danke ich vor allem Herrn Dr. M. Borovcnik.

Literatur

Athen, H. u. H. Griesel (Hrsg.): 1979, *Mathematik heute, Grundkurs Stochastik*, Hannover: Schroedel / Schöningh.

Birnbaum, I. u. G. Fillbrunn: 1984, Die Interpretation statistischer Signifikanz. In: *PM* **26.**

Borovcnik, M.: 1986, Anwendungen der Bayes-Formel. In: *Didaktik der Mathematik* **14**, 183-203.

Riemer, W.: *1985, Neue Ideen zur Stochastik*, Mannheim: Bibliograph. Institut.

Riemer, W.: 1986, Die Bayessche Regel - wie man den Schwierigkeiten beim Verständnis der Testtheorie vorbeugen kann. In: *Bericht über die 9. Fachleitertagung für Mathematik.*

Wickmann, D.: 1990, *Bayes-Statistik*, Mannheim: Bibliographisches Institut.

Wiedling, H.: 1979, Konsequenzen des Satzes von Bayes bei der Interpretation von Stichprobenergebnissen, in: *MNU* **32**, 335-339.

Brosamen für den rauhen Alltag

Man müßte eigentlich jeden Statistik-Neuling, der ein Statistikpaket "fahren" möchte, warnen, daß die Kombination Statistikpaket-Computer in verschiedener Hinsicht einem Rennwagen gleicht; man kommt zwar unter Umständen extrem schnell irgendwohin, wenn man aber die Maschine ungenügend beherrscht und ihr nicht Zügel anlegt, so ist das meist der Straßengraben oder eine Hausmauer.

* * *

Du bist auf dem Pfad zur statistischen Erkenntnis nie allein. Kollege "Zufall" ist stets im Spiel, auch wenn es deine wissenschaftliche Aufgabe ist, zu beweisen, daß er diesmal die Karten nicht verteilt hat.