RETROSPEKTIVER EINSATZ VON KLASSISCHEN TESTVERFAHREN

Eine erkenntnistheoretische Fragestellung beim Testen von Hypothesen

von Jörn Bruhn, Elmshorn

Zusammenfassung: Die Methodologie klassischer Testverfahren und die der explorativen Datenanalyse sind grundsätzlich verschieden. Werden im Unterricht Testverfahren zur Datenanalyse eingesetzt, wie dies häufig geschieht, können Schüler die Grundlagen der Testverfahren nicht verstehen, sondern nur lernen, Rezepte schematisch anzuwenden. An zwei Beispielen wird der Unterschied zwischen statistischer und inhaltlicher Absicherung von Hypothesen aufgezeigt.

1. Vorbemerkungen

Viele Mathematiklehrpläne der Sekundarstufe II enthalten den Hinweis, daß an geeigneten Stellen erkenntnistheoretische Fragestellungen behandelt werden sollten. Die Realisierung in der Schulpraxis setzt aber voraus, daß die Schüler die auftretenden Probleme als relevant erkennen und selbständig Lösungen finden können. Eine solche Stelle liegt im Stochastikunterricht beispielsweise beim retrospektiven Einsatz klassischer Testverfahren vor, bei dem schon vorliegende Daten nachträglich einem Test unterworfen werden.

2. Wie statistische Aussagen zustande kommen - Die übliche Vorgangsweise

In vielen Schulbüchern findet man mit geringen Abwandlungen das folgende Beispiel

Beispiel 1 - Lottozahlen: Die nachfolgende Übersicht zeigt die Ziehungshäufigkeit der Lottozahlen nach 996 Ziehungen (November 1974):

Zahl	1	2	3	4	...	13	...	48	49
Häufigkeit	124	125	123	118	...	96	...	137	142

Dieser Datensatz zeigt, daß 13 die Zahl mit der kleinsten Ziehungshäufigkeit war. Kann man aufgrund der vorliegenden Ziehungshäufigkeit mit einer statistischen Sicherheit von 95 % ausschließen, daß 13 eine Unglückszahl ist?

Bei der Lösung wird in den Schulbüchern im allgemeinen etwa folgendermaßen argumentiert: Wir gehen davon aus, daß die Lottoziehung nicht manipuliert war, und wählen daher als Nullhypothese "Die Wahrscheinlichkeit p für das Auftreten von 13 bei einer Ziehung ist 6/49". Da 13 die kleinste Ziehungshäufigkeit hat, ist es naheliegend, als Gegenhypothese "p < 6/49" zu wählen, also einseitig zu testen.

Stochastik in der Schule, Jg. (1986), Heft 2,

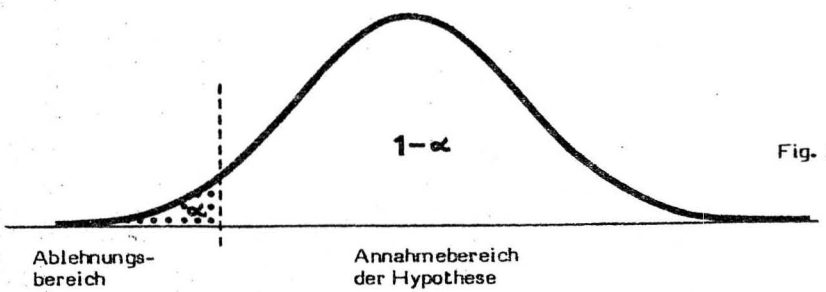

Fig. 1

Die weitere Lösung wird dann etwa so durchgeführt: Bei 996 Ziehungen wird man jede der Zahlen von 1 bis 49 etwa

$$\mu = n * p = 996 * 6/49 = 121{,}99,$$

also ca. 122-mal erwarten.
Die Standardabweichung beträgt

$$\sigma = \sqrt{n * p * (1-p)} = \sqrt{996 * 6/49 * 43/49} \approx 10{,}3.$$

Beim einseitigen Test mit $\alpha = 5\,\%$ wird die Grenze des Ablehnungsbereichs bestimmt durch

$$\mu - 1{,}645 * \sigma.$$

Mit den oben bestimmten Werten ergibt sich somit für die Grenze des Ablehnungsbereichs

$$122 - 1{,}645 * 10{,}3 = 104{,}9\,.$$

Das bedeutet: Der Ablehnungsbereich reicht von 0 bis 104. Die von 13 erreichte Häufigkeit, nämlich 96, liegt daher im Ablehnungsbereich. Also, so wird dann argumentiert, kann man mit 95 % Sicherheit nicht ausschließen, daß 13 eine Unglückszahl ist.

Anmerkung: Bei einem Signifikanzniveau von 5 % würde man auch mittels eines zweiseitigen Tests, d.h. bei der Gegenhypothese "$p \neq 6/49$", zu demselben Schluß kommen, denn der Ablehnungsbereich würde dann linksseitig bis 101 reichen.

Die Lotto-Aufgabe zeigt exemplarisch, auf welche Art viele statistische Testaussagen zustande kommen: Ausgangsbasis ist ein vorhandener Datensatz (hier: Häufigkeit der Lottozahlen). Dieser Datensatz zeigt gewisse Auffälligkeiten (hier: geringe Ziehungshäufigkeit der Zahl 13). Man setzt einen Test an, um diese Auffälligkeit an demselben Datensatz auf Signifikanz zu prüfen. So zeigt auch die folgende Anwendungsaufgabe diese Struktur:

Beispiel 2 - Maschinenstillstände: In einem Werk werden einen Monat lang die Maschinenstillstände einer Schicht notiert, um zu prüfen, ob die Wahrscheinlichkeit hierfür in gewissen Stunden der Schicht besonders groß ist.

Schichtstunde	1	2	3	4	5	6	7	8
Anzahl der Maschinenstillstände	27	16	19	24	23	18	16	17

Weicht die Anzahl der Maschinenstillstände in der ersten Stunde (Auffälligkeit: größte Häufigkeit) signifikant von der mittleren Anzahl der Maschinenstillstände ab?

Ähnlich arbeitet man in vielen Bereichen, in denen die Statistik eingesetzt wird: Ein Datensatz mit gewissen Auffälligkeiten liegt vor, beispielsweise erhöhte Krebserkrankungen in gewissen Gebieten, erhöhtes Waldsterben in gewissen Regionen, verringerte Unfallzahlen zu gewissen Zeiten, Einfluß von Chemikalien auf gewisse Umweltfaktoren, Arzneimittelwirkung auf gewisse Organe, Ergebnisse von psychologischen Tests in gewissen Gruppen usw. Diese Auffälligkeiten prüft man dann mit demselben Datenmaterial in einem Testverfahren auf statistische Signifikanz.

3. Ist die übliche Vorgangsweise einwandfrei? - Gewinnung versus Testen von Hypothesen

Ist die oben beschriebene Vorgehensweise einwandfrei? Um eine Antwort auf diese Frage zu erhalten, betrachten wir wieder das obige Lottozahlen-Beispiel.

Beispiel 3 - Lottozahlen (Fortsetzung): Auch wenn alle Zahlen die gleiche Ziehungswahrscheinlichkeit haben, wird man nicht erwarten, daß alle mit der gleichen Häufigkeit gezogen werden. Eine der gezogenen Zahlen wird die geringste Ziehungshäufigkeit haben. Bis zur 996-ten Ziehung beim Samstags-Lotto war dies die Zahl 13. Die zu testende Vermutung "13 ist eine Unglückszahl" ist nicht unabhängig von den Daten entstanden, die zum Test herangezogen werden, sondern aus diesen gewonnen worden. Entscheidend ist: Der Test wurde also nicht mit irgendeiner der Zahlen von 1 bis 49 durchgeführt, sondern mit der Zahl, von der man weiß, daß sie die geringste Ziehungshäufigkeit hatte.

Wir bestimmen näherungsweise die Wahrscheinlichkeit dafür, daß mindestens eine Zahl mit ihrer Ziehungshäufigkeit im Ablehnungsbereich der Hypothese liegt, also "selten" ist. Eine einfache Modellvorstellung hilft: Kugeln mit den Zahlen von 1 bis

49 werden auf zwei Fächer "selten" und "nicht selten" mit den Wahrscheinlichkeiten α bzw. $1 - \alpha$ verteilt:

Fig. 2

Kugel	Wahrscheinlichkeit für das Hineinlegen in das Fach "nicht selten"
1.	$1 - \alpha$
1. und 2.	$(1 - \alpha)^2$
1. und 2. und 3.	$(1 - \alpha)^3$
...	...
1. und 2. und 3. und ... und 49.	$(1 - \alpha)^{49}$

Die Wahrscheinlichkeit, daß alle 49 Kugeln "nicht selten" sind, beträgt demnach:

$$(1 - \alpha)^{49}.$$

Für die Wahrscheinlichkeit W, daß mindestens eine Kugel "selten" ist, ergibt sich:

$$W = 1 - (1 - \alpha)^{49}.$$

Für $\alpha = 5\,\%$ folgt somit $W = 0{,}919$. Das bedeutet: Die Wahrscheinlichkeit, daß mindestens eine Zahl "selten" ist, beträgt bei dieser Wahl von α ungefähr 92 %. Wenn aber eine Zahl "selten" ist, dann ist es mit Sicherheit insbesondere auch die mit der kleinsten Ziehungshäufigkeit. Also:

Das Ergebnis, daß die Zahl der kleinsten Ziehungshäufigkeit (in dem Lottozahlen-Beispiel die Zahl 13) "selten" ist, also in den Ablehnungsbereich des Tests mit dem Signifikanzniveau $\alpha = 5\,\%$ fällt, ist überhaupt nicht verwunderlich, sondern mit großer Wahrscheinlichkeit, nämlich mit 92 %, zu erwarten.

Anmerkung: Die Wahrscheinlichkeit dafür, Kugeln im Ablehnungsbereich zu finden, kann gut durch Simulation bestimmt werden. Oft finden Schüler gerade dadurch ein "Gefühl" für das Problem. Der zugehörige Algorithmus ist einfach; er zeigt darüber hinaus den Hintergrund der o.a. Näherung auf: Die Schritte sind nicht unabhängig voneinander. Wenn nämlich eine Zahl "selten" ist, müssen andere entsprechend häufiger auftreten.

Die Behandlung des Beispiels zeigt: Ein einzelner gegebener Datensatz kann entweder zur Gewinnung oder zum Testen von Hypothesen dienen.

4. Verbesserte Vorgangsweise beim Testen - statistische und inhaltliche Prüfung von Hypothesen

Wenn ein Datensatz für das Aufstellen einer Hypothese benutzt wird, dann muß die Hypothese entweder statistisch durch einen anderen davon unabhängigen Datensatz oder inhaltlich überprüft werden. Wir betrachten zunächst eine Möglichkeit der statistischen Überprüfung. Dazu wählen wir wieder das Lottozahlen-Beispiel.

Beispiel 4 - Lottozahlen (Fortsetzung): Im November 1974 nach 996 Ziehungen war 13 die Zahl mit der geringsten Ziehungshäufigkeit (96-mal). Die Abweichung vom erwarteten Mittelwert $\mu = n * p \approx 122$ war signifikant. Zu diesem Zeitpunkt hätte man die Hypothese "13 ist eine Unglückszahl" aufstellen können, d.h. die Ziehungshäufigkeit von 13 ist signifikant kleiner als die der anderen Lottozahlen.

Dann hätte man beispielsweise die nächsten 520 Ziehungen bis zum 27. Oktober 1984 abwarten können. Das neugewonnene (vom vorigen Datensatz unabhängige) Datenmaterial zeigt dann: In den 520 Ziehungen ist die Zahl 13 insgesamt 54-mal gezogen worden. Die Abweichung vom erwarteten Mittelwert

$$\mu = 520 * 6/49 = 63{,}7$$

ist nicht signifikant: Wegen

$$\sigma = \sqrt{n*p*(1-p)} \approx 7{,}5$$

liegt die Ziehungshäufigkeit der Zahl 13 im Annahmebereich der Hypothese, denn die Grenze zwischen Annahme- und Ablehnungsbereich liegt zwischen 51 und 52. Die an dem im November 1974 vorliegenden Datenmaterial aufgestellte Hypothese "13 ist eine Unglückszahl" wird also durch das neue, unabhängig gewonnene Datenmaterial nicht gestützt.

Anmerkungen: 1) Betrachtet man alle 1516 Ziehungen des Samstags-Lottos 6 aus 49 (Stand 27. Oktober 1984), so ist die Zahl 13 genau 150-mal gezogen worden. Die Abweichung vom Mittelwert ist signifikant. Dieser Datensatz ist aber nicht unabhängig vom ersten (996 Ziehungen), sondern enthält diesen als Untermenge.

2) Man könnte im Prinzip auch eine inhaltliche Überprüfung der Hypothese vornehmen, z.B. nachprüfen, ob der Zufallsmechanismus beim Lotto nicht einwandfrei war, oder, ob sich die Beschaffenheit der Kugel 13 von der der anderen Kugeln unterscheidet. Natürlich wird man das in diesem Fall mit ziemlicher Sicherheit ausschließen können.

Die an einem Datensatz gewonnene Hypothese kann auch inhaltlich gestützt oder widerlegt werden. Betrachten wir dazu das Beispiel mit den Maschinenstillständen.

Beispiel 5 - Maschinenstillstände (Fortsetzung): Das vorliegende Datenmaterial zeigt, daß die Anzahl der Maschinenstillstände in der ersten Stunde am größten ist. Die Firma kann nun beispielsweise einen Arbeitswissenschaftler beauftragen, zu untersuchen, ob dieser Befund eine Erklärung hat, z.B., ob die Maschinen in einem schlechten Zustand auf die nächste Schicht übergeben werden. Die Arbeiter der nächsten Schicht müßten dann in der ersten Stunde die Maschinen relativ oft stillegen, um diese zu warten. Es könnte sein, daß die Maschinenstillstände in der ersten Stunde arbeitspsychologisch zu erklären sind, z.B. durch Anlaufschwierigkeiten o.ä.

Anmerkung: Auch im Beispiel mit den Maschinenstillständen könnte man die Vermutung durch einen zweiten, unabhängigen Datensatz statistisch zu stützen oder zu widerlegen versuchen. In erster Linie wird diese statistische Prüfung an einem weiteren Datensatz dadurch nahegelegt, daß eine tragfähige inhaltliche Erklärung und Absicherung der aufgestellten Hypothese nicht gefunden werden konnte.

Verwendet man weder die Information über schlechte Schichtstunden aus denselben Daten noch andere Informationen beispielsweise aus früheren Daten, so kann man nur einen undifferenzierten, auf keine spezifische Hypothese eingehenden χ^2-Test auf Gleichverteilung ($\chi^2 = 6$, df = 7) ansetzen. Sein Ausfall zeigt, daß das vorliegende empirische Ergebnis sich nicht als signifikant im statistischen Sinne einstufen läßt, d.h. daß die Schwankungen der beobachteten Daten noch als zufällig angesehen werden müssen.

Die mit dem χ^2-Test verbundene Annahme ("fehlendes inhaltliches Wissen") und die aus dem Ausfall des χ^2-Tests resultierende Einsicht ("keine Signifikanz") legen nahe, einen neuen Datensatz für eine statistische Prüfung zu erheben. Läßt man umgekehrt die dem χ^2-Test zugrundeliegende Annahme ("kein inhaltliches Wissen zur Absicherung der spezifischen Hypothese") fallen, so liegt es gerade dadurch nahe, sich zunächst fehlendes inhaltliches Wissen zu beschaffen, beispielsweise durch neues Datenmaterial. Neben einer weiteren statistischen Untersuchung ist es sinnvoll, alle Anstrengungen auf eine bessere inhaltliche Durchdringung der anstehenden Frage zu lenken!

Eine Analyse der Bedeutung statistischer Aussagen, insbesondere im Zusammenhang mit dem χ^2-Test, findet man in Borovcnik (1984).

Am Beispiel mit den Maschinenstillständen kann man leicht klar machen: Auffälligkeiten im Datenmaterial können inhaltlich erklärt werden, wenn man die Einflußfaktoren kennt. Dies gilt z.B. auch für die Ursachen von Krebserkrankungen, Waldsterben sowie Verkehrsunfällen. Auffälligkeiten in einem Datensatz können etwa die Richtung einer wissenschaftlichen Untersuchung bestimmen.

Anmerkung: Auffälligkeiten im Datenmaterial aufzuspüren und zu untersuchen wird bisher in der Schule zu wenig behandelt. Retrospektives Testen kann als Hilfsmittel der sogenannten Explorativen Datenanalyse eingesetzt werden. In dieser Verwendung ist retrospektives Testen durchaus sinnvoll (siehe etwa P. Ihm, 1980). Darüber hinaus bietet die Explorative Datenanalyse anschauliche graphische Methoden: Diese bieten eine große Chance für den Stochastikunterricht.

5. Die Problematik von "vorher" und "nachher"

Eine immer wieder beklagte methodische Schwierigkeit beim Unterrichten von Testverfahren in der Schule beruht gerade darauf, daß zwischen "vorher" und "nachher" nicht sorgfältig unterschieden wird. Dies soll an einem weiteren bekannten Beispiel erläutert werden.

Beispiel 6 - Arzneimittel: Ein Hersteller behauptet, daß ein Medikament in 70 % der Fälle bei einer bestimmten Krankheit hilft. Ein Test im Krankenhaus ergab, daß 15 von 25 Patienten gesund wurden. Kann die Hypothese "Heilungschance beträgt 70 %" aufrechterhalten werden?

Bei dieser Aufgabe, so wird bemerkt, bestimmen die Schüler die Bernoulli-Wahrscheinlichkeit $B_{25;0,7}(15)$ und argumentieren, daß dieses Ergebnis sehr selten sei; es sei nicht leicht, die Schüler dazu zu bringen, die kumulierten Wahrscheinlichkeiten $B_{25;0,7}(X \leq 15)$ zu betrachten. Diese Schwierigkeit sollte m.E. nicht durch automatisierte Verfahren übergangen werden, sondern sollte vielmehr durch geeignete Aufgabenstellung und Analyse behoben werden.

Beispiel 7 - Arzneimittel (Fortsetzung): Ein Hersteller vermutet aufgrund seiner Laboruntersuchungen, daß ein Medikament in 70 % der Fälle bei einer bestimmten Krankheit hilft.

a) Welches Testverfahren einseitig/zweiseitig und welches Signifikanzniveau ist angemessen?

b) Bestimme den Verwerfungsbereich der Hypothese für das gewählte Testverfahren und das angenommene Signifikanzniveau.

c) Entscheide, ob die Hypothese "Heilungschance beträgt 70 %" abgelehnt werden muß, wenn eine Testdurchführung in einem Krankenhaus ergab, daß 15 von 25 Patienten gesund wurden.

Der Verwerfungsbereich wird bestimmt, bevor das Ergebnis der Testdurchführung vorliegt.

Anmerkung: Es ist ein elegantes und in der Praxis häufig verwendetes Verfahren, das nominelle Signifikanzniveau, den sogenannten p-Wert, zu ermitteln, das durch ein schon vorliegendes Testergebnis bestimmt wird. Man nimmt dieses Ergebnis als Grenze zwischen Verwerfungs- und Annahmebereich und bestimmt die Summe der Wahrscheinlichkeiten für dieses und alle "extremeren" Ergebnisse. Die Formulierung deutet schon an, daß man sich vorstellt, das Ergebnis sei noch nicht eingetreten.

6. Abschließende Bemerkungen

Die angesprochene Problematik beim Testen von Hypothesen ist Schülern nicht ohne weiteres einsichtig. Zu Beginn der Problemstellung wird im Klassengespräch etwa geäußert: "Entweder ist die Abweichung gewisser Merkmalsausprägungen signifikant oder nicht. Darum spielt es keine Rolle, ob Hypothese und Test vor oder nach Kenntnis der Daten festgelegt werden." Es kommt also darauf an, daß Schüler den Unterschied zwischen inhaltlicher und statistischer Absicherung von Hypothesen erkennen. Soll eine Auffälligkeit statistisch abgesichert werden, so darf dies nicht an den Daten geschehen, an denen die Auffälligkeit erkannt worden ist. Diese klare und einfache Abgrenzung muß im Unterricht deutlich werden, wenn Schüler die "Philosophie" der klassischen Testverfahren und ihrer Anwendungen durchschauen sollen.

Literatur

BRUHN, J.: Statistische Verfahren. Braunschweig-Wiesbaden: Vieweg 1986.

BOROVCNIK, M.: Was bedeuten statistische Aussagen. Wien-Stuttgart: Hölder-Pichler-Tempsky / Teubner 1984.

ENGEL, A.: Statistik auf der Schule: Ideen und Beispiele aus neuerer Zeit. In: Der Mathematik-Unterricht 28 (1982), 57-85.

IHM, P.: Explorative und konfirmatorische Datenanalyse - Gegensatz oder Ergänzung? In: N. VICTOR e.a. (Hrsg.): Explorative Datenanalyse. Berlin-Heidelberg-New York: Springer 1980, 38-53.

UNKELBACH, H.D. u. T. WOLF: Ehrlichkeit beim statistischen Testen. In: VOLLMAR, J. (Hrsg.): Biometrie in der chemisch-pharmazeutischen Industrie. Stuttgart-New York: G. Fischer 1983, 7-12.

Prof. Jörn Bruhn
Institut für Didaktik der
Mathematik und Naturwissenschaften
Universität Hamburg
D-2000 Hamburg (BRD)