Hypothesentest und bedingte Wahrscheinlichkeit

RENATE MOTZER, AUGSBURG

Zusammenfassung: *Schülerinnen und Schüler werden im Lauf des Stochastikunterrichts meist mit 2 Arten von Tests konfrontiert: Test, in denen ein Einzelner sein Risiko abschätzt (z. B. eine bestimmte Krankheit in sich zu tragen) und Tests, bei denen eine Hypothese über eine bestimmte Wahrscheinlichkeit (z. B. Wirksamkeit eines Medikaments) beurteilt werden soll.*
Bei beiden Arten von Tests können wie bei jedem Test Fehler auftreten. Doch gibt es noch mehr Gemeinsamkeiten bei diesen Arten von Tests? Können sie daher im Unterricht analog behandelt werden?

1 Die Gemeinsamkeiten: Es gibt zwei Arten von Fehlern

Da ich im vergangenen Schuljahr auf das Thema bedingte Wahrscheinlichkeit viel Wert gelegt habe und die Schülerinnen und Schüler sich in diesem Zusammenhang einige Aufgaben zur Deutung eines medizinischen Tests bzw. von Zeugenaussagen u.ä. erarbeitet haben, stellte sich für einige die Frage, ob es sich nicht bei Hypothesentests um ähnliche Phänomene handelt. In beiden Fällen werden Tests gemacht und es kann zwei Sorten von Fehlern geben. Die Situation kann jeweils durch eine Vierfeldertafel oder durch ein Baumdiagramm dargestellt werden.

Soweit die Gemeinsamkeiten. Vor ein paar Jahren wurde diskutiert, wie weit man diese Analogie im Unterricht fruchtbar machen kann und soll (vgl. Krauss und Wassner 2001 und Diepgen 2002). Krauss und Wassner sprechen sich vor allem deswegen dafür aus, weil den Schülerinnen und Schülern klar gemacht werden soll, dass der Fehler 1. Art und der Fehler 2.

Art beim Signifikanztest nur bedingte Wahrscheinlichkeiten sind. Dieses wird in der Deutung nicht nur von Schülerinnen und Schülern, sondern auch von Erwachsenen, die in ihrem Beruf mit Tests zu tun haben, oft missverstanden. Diepgen spricht sich eher gegen eine analoge Behandlung aus, weil die Fragestellungen und die nötigen Berechnungen doch sehr unterschiedlich sind.

Da auch meine Schülerinnen und Schüler große Schwierigkeiten in der Deutung der Fehler beim Hypothesentest haben, habe ich mich dafür entschieden, die Parallelen und die Unterschiede der beiden Testarten bewusst zu thematisieren. Wenn dann hängen bleibt, dass der Fehler 1. und 2. Art bedingte Wahrscheinlichkeiten sind, ist aus meiner Sicht durchaus etwas gewonnen.

Wenn man andererseits bei Wikipedia schaut, wird der Fehler 1. Art analog zu einem „falsch positiven" AIDS-Test eingeführt (auch wenn das ein zugehöriger Kommentar zu Recht kritisch sieht). Eine Schülerin hat dies in einem Referat im Unterricht eingebracht, so dass die Frage der Mitschüler nach dem Zusammenhang der Testarten noch berechtigter wurde.

Man findet bei wikipedia unter „Fehler 1. Art"[1]:

> In der Statistik besteht beim Testen von Hypothesen ein **Fehler 1. Art** darin, eine Nullhypothese zurückzuweisen, obwohl sie wahr ist (beruhend auf falsch-positiven Ergebnissen). Man nennt diesen Fehler auch ***α*-Fehler**. Mathematisch formuliert bezeichnet er die Wahrscheinlichkeit, dass die so genannte Null- bzw. Ausgangshypothese „H_0" abgelehnt wird, obwohl sie richtig ist.
>
> Die Ausgangshypothese H_0 ist hierbei die Annahme, die Testsituation befinde sich im „Normalzustand", das kann zum Beispiel heißen: *der Patient ist gesund, der Angeklagte ist unschuldig oder die Person hat Zugangsberechtigung*. Wird also dieser „Normalzustand" nicht erkannt, obwohl er tatsächlich vorliegt, handelt es sich um einen Fehler 1. Art. Beispielsweise wird eine Person zu Unrecht als krank bezeichnet, obwohl sie tatsächlich gesund ist. Falsch Positive (englisch: *false positives*) sind in diesem Fall *zu Unrecht als krank bezeichnete Gesunde*.

Ich plädiere hier nicht dafür, Wikipedia-Artikel zur Grundlage des Unterrichts zu machen. Wenn sich Schülerinnen und Schüler jedoch mit einigen Aspekten der Schulmathematik selbstständig beschäftigen, können sie leicht auf diese Artikel stoßen.

Auch kann eine Diskussion solch eines Wikipedia-Artikels im Unterricht dazu beitragen, bei den Schülerinnen und Schülern die Einsicht zu fördern, dass im Internet fachliche Begriffe manchmal falsch oder zumindest missverständlich dargestellt sein können.

2 Unterschiede in der Aufgabenstellung

Trotz der Analogie der möglichen Fehler gibt es wesentliche Unterschiede bzgl. der Verwendung der beiden Testarten.

Macht ein Einzelner einen medizinischen Test, so geht es ihm um seine persönliche Gesundheit. Die Fehler 1. und 2. Art sind schon bekannt (bzw. die Spezifität $1 - \alpha$ und Sensitivität $1 - \beta$). Sie wurden in statistischen Tests (bei anderen Patienten) ermittelt und dürfen als gegeben vorausgesetzt werden. Nun wird ermittelt, mit welcher Wahrscheinlichkeit ein positiver Test bei dieser einen Person eine wirklich vorhandene Krankheit anzeigt.

Beim Hypothesentest wird i. A. die relative Häufigkeit für eine größere Gruppe untersucht.

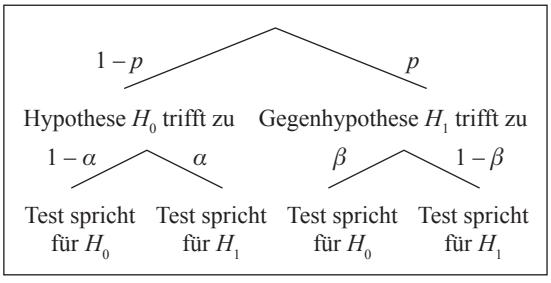
Um wiederum ein Beispiel aus dem medizinischen Bereich heranzuziehen: Welchem Anteil der Behandelten hilft ein bestimmtes Medikament/eine bestimmte Therapie?

Es wurde ein neues Medikament/eine neue Therapie entwickelt. Hilft sie mehr Menschen als die bisherige Standardtherapie?

Ob sie einem Einzelnen hilft, kann man dabei relativ leicht feststellen. Ob im anderen Fall jemand den Krankheitserreger in sich trägt oder nicht, weiß man (noch) nicht und will dies durch den Test in Erfahrung bringen, wobei der Test aber fehlerhaft sein kann. Das Risiko liegt also hier in einem falschen Testergebnis, während beim Hypothesentest das Risiko in einer falschen Interpretation eines (eindeutigen) Testergebnisses (in der untersuchten Gruppe hat es … Teilnehmern geholfen) liegt.

3 Gegenüberüberstellung anhand eines Baumdiagramms

Als Baumdiagramm stellen sich die beiden Situationen so dar:



Beim medizinischen Bayes-Test eines Einzelnen (H_0: Der Patient trägt die Krankheit nicht in sich) hängt

die Bewertung eines positiven Testergebnisses nicht nur von α und β ab, sondern ganz wesentlich von p, d. h. der Prävalenz, dem Prozentsatz der (untersuchten) Personen, die die Krankheit wirklich in sich tragen.

Will man wissen, mit welcher Wahrscheinlichkeit jemand fälschlicherweise ein positives Testergebnis erhält, so berechnet man den Anteil aller, die ein positives Testergebnis bekommen: $(1 - p) \cdot \alpha$ (jemand hat die Krankheit nicht und erhält trotzdem ein positives Ergebnis) $+ p \cdot (1 - \beta)$ (jemand hat die Krankheit und erhält ein positives Ergebnis).

Von dieser Gesamtmenge bestimmt man nun den Anteil derer, die fälschlicherweise ein positives Ergebnis bekommen:

p(„gesund trotz positivem Test“) =

$$\frac{\alpha \cdot (1 - p)}{\alpha \cdot (1 - p) + p \cdot (1 - \beta)}$$

Dass es didaktisch vorteilhaft ist, hier mit natürlichen Häufigkeiten statt nur mit Wahrscheinlichkeiten bzw. relativen Häufigkeiten zu arbeiten, ist vielfach herausgestellt worden (u. a. auch von Krauss & Wassner 2001). Auf diesen Aspekt der unterrichtlichen Behandlung will ich hier nicht weiter eingehen, sondern ihn voraussetzen.

Bei einem Hypothesentest gibt es kein p, sondern man geht davon aus, dass in der Situation prinzipiell nur H_0 oder H_1 zutreffen kann, aber nicht für manche H_0 und für andere H_1: entweder hilft das neue Medikament mehr Menschen oder nicht.

α und β sind jedenfalls nur **bedingte** Wahrscheinlichkeiten und geben Auskunft über die Wahrscheinlichkeiten von Daten unter den Bedingungen H_0 bzw. H_1. Sie geben aber keine Aussagen über die Wahrscheinlichkeiten von Hypothesen. Dazu müsste man z. B. etwas über die Prävalenz von H_1 wissen.

Beck-Bornholdt und Dubben 2003 versuchen in dem Buch „Der Schein der Weisen“ die beiden Testarten zusammenzubringen. Sie unterstellen eine Gesamtheit von Studien und geben für p den Anteil der Studien, die sich mit neuen Medikamenten/Therapien beschäftigen, die wirklich besser sind als die bisherigen.

Mit $\alpha = 5$ % und $\beta = 20$ % können sie für den Fall von $p = 10$ % errechnen, dass die Wahrscheinlichkeit, dass bei einem positiven Studienergebnis die neue Behandlung tatsächlich besser ist, bei 64 % liegt. (vgl. S. 197, wobei dort ein zweiseitiger Test angewendet wird und daher bezogen auf das obige Baumdiagramm mit $\alpha = 2{,}5$ % gerechnet wird, was am Schluss zu einer Wahrscheinlichkeit von 78 % führt).

H. Wirths (2005) beschreibt einen Weg vom Baumdiagramm, das Bayes-Probleme löst, zu einem Baumdiagramm, das den Alternativtest behandelt, wobei er für beide Alternativen $p = 0{,}5$ ansetzt.

Eine meiner Schülerinnen fragte, ob denn nicht H_1 das Gegenereignis zu H_0 sei und man daher nicht bei den Fehlern $\beta = 1 - \alpha$ rechnen dürfe. Ich konnte ihr bzgl. der Feststellung „Gegenereignis“ durchaus recht geben, aber da kein p bzgl. der oberen Verzweigung da ist, kann man auch nicht $1 - p$ für die andere Verzweigung rechnen. Die Fehlerwahrscheinlichkeiten haben damit gar nichts zu tun und müssen anders berechnet werden.

Vor allem beziehen sich die beiden Fehler auf unterschiedliche Ergebnisse des Tests. Liegt ein Testergebnis vor, kann nur noch einer der beiden Fehler gemacht werden.

4 Bestimmung des Annahme- bzw. Ablehnungsbereichs beim Signifikanztest

Die Aufgabe, den Annahme- bzw. Ablehnungsbereich zu bestimmen, wenn man insgesamt jeweils n Patienten untersucht, ist damit noch nicht angegangen.

Diese (in der Schule) übliche Frage zum Signifikanztest hat nicht direkt mit den Fragen zur bedingten Wahrscheinlichkeit zu tun, wie sie im Baumdiagramm dargestellt ist.

Sie bestimmt allerdings, wie der Punkt „Test spricht für H_0/H_1“ konkret durchzuführen ist.

Beim medizinischen Bayes-Test hat diese Frage nichts mit einer (vom Schüler durchzuführenden) Rechnung zu tun, sondern wird von dem Labor beantwortet, das die Blutprobe o. ä. untersucht.

Beim Hypothesentest haben die Schülerinnen und Schüler meist das Finden der Entscheidungsregel im Unterricht zu leisten (häufig unter Zuhilfenahme des Tabellenwerks).

Man könnte freilich, wenn denn ein konkretes Messergebnis vorliegt, daraus auch auf α (und wenn eine quantifizierbare Gegenhypothese vorliegt auch auf β) schließen. Krauss und Wassner nennen diesen Wert von α den P-Wert, während das Signifikanzniveau α ein vorgegebener Wert ist, so dass gelten soll: der P-Wert sollte kleiner als α sein (häufig wird für α 5 % gewählt).

Bsp.: $n = 100$. Die Hypothese heißt: Die neue Methode hilft **höchstens 40 %**, Gegenhypothese: Sie hilft mehr als 40 % (mindestens 50 %).

Die neue Methode hat beim Test **47** Patienten geholfen.

Ist sie wirklich besser als die alte Methode (die etwa 40 % geholfen hat)?

Wählt man den Annahmebereich ganz knapp so, dass 47 gerade noch dazugehört (und man daher schließen würde: „Nein, die neue Methode ist nicht besser."), so gilt: $\alpha = 1 - 0{,}9362 = 6{,}38\ \%$ und $\beta = 30{,}87\ \%$, wobei man mit dieser Entscheidung für H_0 nur den β-Fehler machen kann.

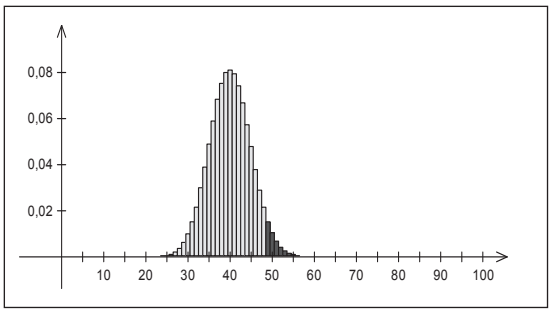
Wählt man den Annahmebereich ganz knapp so, dass 47 nicht mehr dazu gehört, so gilt:
$\alpha = 1 - 0{,}9070 = 9{,}3\ \%$ und $\beta = 24{,}21\ \%$, wobei man mit dieser Entscheidung für H_1 nur den α-Fehler machen kann.

Jedenfalls ist das Ergebnis nicht signifikant, wenn man als signifikant nur die Untersuchungen bezeichnet, bei denen das Ergebnis zum Ablehnungsbereich gehört, wobei α höchstens 5 % ist.

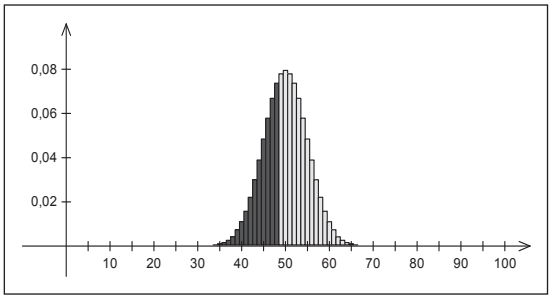
Wäre die Methode bei 49 Personen erfolgreich gewesen, ergäbe sich $\alpha = 1 - 0{,}9577 = 4{,}23\ \%$.

Ab 49 Personen wäre das Ergebnis also signifikant.

Hier das Histogramm für H_0 (40 %). Dunkel gekennzeichnet ist der Ablehnungsbereich, wenn man das Signifikanzniveau 5 % vorgibt (exakter Wert 4,23 %). Der Ablehnungsbereich beginnt bei 49.



Das folgende Histogramm gibt H_1 (mit 50 %) an. Der β-Fehler beträgt max. 38,22 %.



5 Fazit

Der Vergleich der Fragestellungen, die bei Bayes-Aufgaben bzw. bei Aufgaben zum Hypothesentest eine Rolle spielen, anhand eines Baumdiagramms kann Schülerinnen und Schüler helfen, die Zusammenhänge der beiden Testarten zu erkennen (dass jeweils 2 Arten von Fehlern möglich sind), aber auch die Unterschiede in den Fragestellungen. Einsichtig kann dabei werden, dass man beim Hypothesentest nichts über die Wahrscheinlichkeit der Hypothesen sagen kann, sondern nur über die Wahrscheinlichkeit, dass bestimmte Daten auftreten, wenn die eine oder andere Hypothese zutrifft.

Beide Testarten sind für Schülerinnen und Schüler nicht einfach zu verstehen. Ihnen sollte aber im Unterricht große Aufmerksamkeit zukommen, da die dort behandelten Fragen durchaus alltagsrelevant sein können. Um eine Ahnung davon zu haben, was es heißt, dass ein Untersuchungsergebnis signifikant ist, ist es gut zu wissen, dass dem eine bedingte Fehlerwahrscheinlichkeit von höchstens 5 % zugrunde liegt. Dass solch eine bedingte Wahrscheinlichkeit aber nicht die in der Situation wirklich gefragte Wahrscheinlichkeit ist, kann man wiederum den Bayes-Aufgaben entnehmen. Daher lohnt sich auch beide Testverfahren anhand eines Baumdiagramms zu vergleichen.

Anmerkung

1 http://de.wikipedia.org/wiki/Fehler_1._Art (besucht am 6.8.2009)

Literatur

Beck-Bornholdt, Hans-Peter; Dubben, Hans-Hermann (2003): Der Schein der Weisen – Irrtümer und Fehlurteile im alltäglichen Denken. Reinbeck: Rowohlt Taschenbuch Verlag.

Wirths, Helmut (2005): Vom Rückwärtsschließen im Baumdiagramm zum Testen von Hypothesen. In: *Stochastik in der Schule* 25(2), S. 4–10.

Krauss, Stefan; Wassner, Christoph (2001): Wie man das Testen von Hypothesen einführen sollte. In: *Stochastik in der Schule* 21(1), S. 29–34.

Diepgen, Raphael (2002): $P(H|D)$ versus $P(D|H_0)$? Wie man das Testen von Hypothesen – doch nicht – einführen sollte. In: *Stochastik in der Schule* 22(3), S. 34–38.

Anschrift des Verfassers

Renate Motzer
Didaktik der Mathematik
Universität Augsburg
Universitätsstr. 10
86135 Augsburg
Renate.Motzer@math.uni-augsburg.de