Verteilungen und Beurteilende Statistik -
über Probleme bei der Einführung des Testens von Hypothesen

von Heinz Klaus Strick, 5090 Leverkusen

Im folgenden Aufsatz geht es um Probleme bei der Einführung des Hypothesentests im Stochastikunterricht der Sekundarstufe II. Es soll anhand von Beispielen dargestellt werden, welche Art von Fragestellungen sich hierfür eignen. Gleichzeitig soll dabei untersucht werden, ob ein Einstieg auch über nicht-binomialverteilte Zufallsgrößen günstig ist. Außer Frage steht, daß durch eine mehrfache Behandlung der Thematik anhand verschiedener Verteilungen die Idee des Verfahrens gefestigt werden kann.

1. Einstieg: Binomialtest für große Stichprobenumfänge

Es ist möglich, das Testen von Hypothesen am Beispiel von binomialverteilten Zufallsgrößen einzuführen. Eine Einführung in den Problemkreis ist m.E. besonders einfach, wenn eine Situation gegeben ist, in der sich ein Binomialmodell mit großem Stichprobenumfang anwenden läßt. Das Konzept zu diesem Zugang habe ich an verschiedenen Stellen erläutert; deshalb soll hier ein Beispiel genügen:

Beispiel 1:

Immer mehr Menschen haben unter Allergien zu leiden. Bisher auf dem Markt angebotene Salben gegen eine bestimmte Hautallergie führten in 70 % aller Fälle zu einer Heilung innerhalb von 10 Tagen. Von einem neuen Medikament behauptet der Hersteller, daß es eine größere Wirksamkeit zeige als die bisher vorhandenen Salben. Der Hersteller stützt seine Behauptung auf die Ergebnisse eines Großversuchs, bei dem 357 Patienten mit dieser Hautkrankheit behandelt wurden.

Welches Ergebnis hätte der Großversuch haben müssen, damit man das neue Medikament für wirksamer halten könnte?

Bei der Aufstellung der geforderten Entscheidungsregel benutzt man die Approximation der Binomialverteilung durch die Normalverteilung. Dies ist bei großen Stichprobenumfängen zulässig und es erspart mühsame Rechnungen.

Schwierigkeiten bei der Einführung des Hypothesentest-Verfahrens können verschiedener Art sein. Auf eine Schwierigkeit stößt man bei Schülern immer wieder - sie ist auch nicht unbedingt typisch für den Binomialtest:

Welche Hypothese ist überhaupt zu testen?

Für die Problemstellung kommen zwei Hypothesen in Frage, die sich aufgrund verschiedener Standpunkte ergeben. Es ist üblich, diese als Nullhypothese und Alternativhypothese zu bezeichnen: da jedoch je nach Standpunkt jeweils die Rolle von Null- und Alternativhypothese vertauscht wird, erscheint es ratsam, nur von "der zu testenden Hypothese" zu sprechen.

Welche der beiden in Frage kommenden Hypothesen getestet werden muß, ergibt sich oft aus der Formulierung des Aufgabentextes. Erfahrungsgemäß ist es für Schüler (z.B. in Prüfungssituationen) hilfreich, wenn sie sich im ersten Teil der Aufgabenlösung mit der Frage auseinandersetzen, welche Standpunkte hier eine Rolle spielen können, und unter welchen Bedingungen man bereit ist, von diesen

Standpunkten abzugehen.

In der o.a. Aufgabe könnte z.B. der Hersteller folgenden Standpunkt vertreten: Mein Produkt ist besser, da es nach neuesten Erkenntnissen hergestellt wurde und alle bisherigen Kenntnisse bei der Entwicklung mitverwertet wurden. Von diesem Standpunkt lasse ich mich erst abbringen, wenn im Großversuch besonders wenige Heilungen innerhalb von 10 Tagen eintreten. Bei diesem Standpunkt würde also die Hypothese $p > 0.7$ untersucht.
Extrem wenige Heilungen bedeutet auf dem 95 % - Niveau (gemeint ist: Irrtumswahrscheinlichkeit 1. Art = 5 %), daß auf jeden Fall weniger als 236 Patienten geheilt werden:

Rechnung für $p = 0.7$: (vgl. Abb. 1)

$\mu - 1.64\sigma = 249.9 - 1.64 * 8.66 = 235.7$

Die Rechnung für $p > 0.7$ kann exemplarisch durchgeführt werden, z.B. für $p = 0.71$, $p = 0.72$ usw.: wir begnügen uns hier mit der Feststellung: Für $p > 0.7$ ist auch $\mu - 1.64\sigma > 235.7$.

Wenn man als Verbraucher diesen Standpunkt übernehmen würde, wäre offensichtlich die Gefahr sehr groß, daß man ein besseres Medikament gegen ein schlechteres eintauscht. (236 von 357 Personen - das sind nur 66 %!). Das kann nicht im Interesse der Verbraucher sein.

Für die Verbraucherseite ist der gegenteilige Standpunkt angemessener: Man läßt sich erst durch eine extrem große Zahl von geheilten Patienten von der größeren Wirksamkeit des neuen Medikaments überzeugen. Man stellt sich also auf den Standpunkt, daß das neue Medikament

höchstens so gut ist wie die alten, d.h., daß die Heilungschancen p gleich 70 % sind oder weniger als 70 % betragen. Diesen Ansatz verwerfen wir (sehen wir als falsch an), wenn ungewöhnlich viele Patienten in dem Zeitraum geheilt würden.

Für das 95 %-Niveau ist folgende Rechnung notwendig:

Für $p = 0.7$ ist: $\mu + 1.64\sigma = 264.1$. (vgl.Abb. 1)

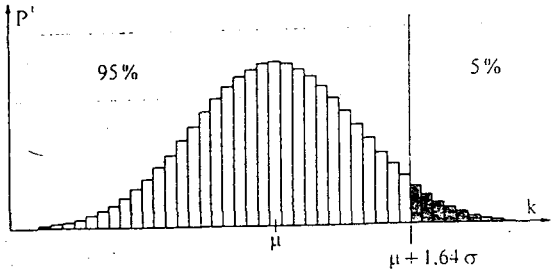
Für kleinere Werte von p ist der Wert von $\mu + 1.64\sigma$ kleiner als 264.1 (was man wiederum durch beispielhafte Rechnungen zeigen könnte).

Es ergibt sich daher die Entscheidungsregel:

Verwirf die Annahme (Hypothese) $p \leq 0.7$, falls die Anzahl der innerhalb von 10 Tagen geheilten Patienten größer als 264 ist.

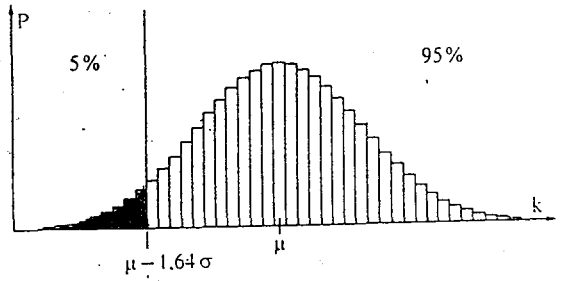
Natürlich können auch bei diesem Ansatz Fehler zu ungunsten der Patienten auftreten:

Die Beschränkung "95 % - Niveau" besagt: Falls die Heilungschancen bei der neuen Salbe tatsächlich höchstens 70 % betragen, dann kommt es zufällig in 5 % von Stichproben vom Umfang 357 zu über 264 Heilungen innerhalb von 10 Tagen, und man geht wegen der vereinbarten Entscheigungsregel davon aus, daß die neue Salbe besser ist als die bisherigen, obwohl dies nicht der Fall ist. Zu einer solchen falschen Entscheidung kommt es aber nur in 5 % der Fälle, da das Entscheidungsverfahren so angelegt ist (Fehler 1. Art).



$P(X \leq \mu + 1{,}64\,\sigma) = 0{,}95$

Abb. 1:
Annahme- und Verwerfungsbereich bei Binomialverteilungen mit großem n



$P(X \geq \mu - 1{,}64\,\sigma) = 0{,}95$

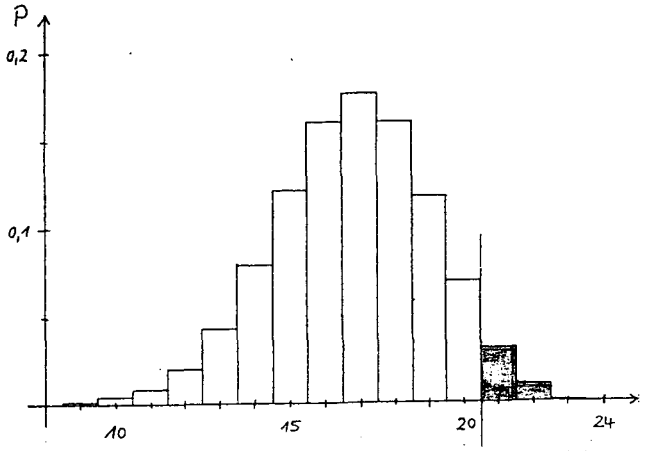


Abb. 2:
Annahme- und Verwerfungsbereich bei Binomialverteilungen für n = 24 und p = 0,7

Ein anderer Fehler geht auch zu Lasten der zukünftigen Patienten: Es kann vorkommen, daß das Stichprobenergebnis nicht jenseits des kritischen Werts von 264 liegt, obwohl die neue Salbe tatsächlich besser ist als die bisherigen (Fehler 2. Art).

Die Wahrscheinlichkeit für einen solchen Fehler könnte man allerdings nur bestimmen, wenn man die tatsächlichen Heilungschancen der neuen Salbe kennen würde. (Diesen Mangel behebt man durch systematische Untersuchung für verschiedene Werte von p: Der Graph der entstehenden Funktion ist die Operationscharakteristik bzw. die Gütefunktion des Tests.)

2. Binomialtest bei kleinem Stichprobenumfang - welcher Einstieg ist günstiger?

Je nach Schwere der Krankheit würde man das Risiko eines Großversuchs nicht auf sich nehmen wollen. Wie ist zu verfahren, wenn der Stichprobenumfang so klein ist, daß die Approximation durch die Normalverteilung nicht mehr zulässig ist?

Beispiel 2:

Der im Beispiel 1 beschriebene Versuch werde mit 24 Patienten durchgeführt.

Gib hier eine Entscheidungsregel für das 95 %-Niveau an.

Es gibt viele Lehrer, die mit einem solchen Beispiel in den Problemkreis des Testens von Hypothesen einsteigen. Das Beispiel unterscheidet sich scheinbar nur unwesentlich von Beispiel 1. Dennoch sind bei Schülern in Beispiel 2 mehr Schwierigkeiten zu beobachten als bei Beispiel 1.

Die Probleme liegen meist in der Vorgeschichte des Einstieges:

Vorher stehen über Stunden Berechnungen von Binomial-Wahrscheinlichkeiten auf dem Programm; dies geschieht bei kleinem n mit dem Taschenrechner, bei größerem n vielleicht mit dem programmierbaren Taschenrechner oder Kleincomputer oder aber auch für bestimmte Werte von n mit Hilfe von Tabellen. Das kostet erfahrungsgemäß alles Zeit; Schüler haben an allen möglichen und unmöglichen Stellen ihre Probleme und manche sind froh, wenn sie schließlich alle geforderten Berechnungen glatt schaffen. Und dann kommt eine Aufgabenstellung, bei der auch wieder Wahrscheinlichkeiten zu berechnen sind; und auch noch für n = 24.

Daß es beim Hypothesentest weniger auf Rechnung als auf Argumentation ankommt, wird manchen Schülern erst spät klar. Mit dieser Hypothek sollte der Unterricht nicht belastet sein.

Worin liegen die Probleme für manche Schüler, wenn man als Einstieg Beispiel 2 wählt?

Man kann für n = 24 und p = 0.7 alle Wahrscheinlichkeiten P (X=k) berechnen:

P(X=24) = 0.0002
P(X=23) = 0.0020
P(X=22) = 0.0097
P(X=21) = 0.0305
P(X=20) = 0.0687
P(X=19) = 0.1177
P(X=18) = 0.1598

usw. (vgl. Abb. 2)

Aus der Tabelle können wir Beispiele ablesen: Wenn p=0.7 ist, dann kommt dem Ergebnis '24 geheilte Patienten' nur eine Wahrscheinlichkeit von 0.02 % zu, dem Ergebnis '23 geheilte Patienten' nur die Wahrscheinlichkeit 0.2 %, usw. Das sind alles kleine Wahrscheinlichkeiten: jedes Ergebnis ist für sich genommen 'selten'.

Der Schritt zu den kumulierten Wahrscheinlichkeiten

$P(X \leq 24) = 1.0000$
$P(X \leq 23) = 0.9998$
$P(X \leq 22) = 0.9978$
$P(X \leq 21) = 0.9881$
$P(X \leq 20) = 0.9576$
$P(X \leq 19) = 0.8889$

usw.
und deren Interpretation bereitet die genannten Schwierigkeiten.

Was den Schülern oft fehlt, ist die Einsicht, daß es auf Bereiche ankommt: Beim Hypothesentest betrachtet man einen Bereich, in dem mit großer Wahrscheinlichkeit (z.B. 95 %) ein Stichprobenergebnis liegt, wenn der Ansatz (die hypothetische Erfolgswahrscheinlichkeit) richtig ist. Fällt das Stichprobenergebnis nicht in diesen Bereich, dann liegt es in der Strategie des Verfahrens, den Ansatz als falsch anzusehen. Auf einzelne Wahrscheinlichkeiten kommt es beim Hypothesentest nicht an.

Diese Einsicht muß beim Einstieg über Binomialverteilungen bei kleinem Stichprobenumfang erst herausgearbeitet werden, d.h., man muß eine Vereinbarung darüber treffen, bei welchen Ergebnissen des Versuchs man die neue Methode als besser ansehen würde - es geht hier also um die Fest-

legung des Verwerfungsbereichs der Hypothese $p \leq 0.7$.

Zusätzliche Probleme treten erfahrungsgemäß auf, wenn man in Beispiel 2 die Aufgabenstellung durch eine Information erweitert und scheinbar präzisiert: 'Bei der Versuchsdurchführung waren innerhalb von 10 Tagen 20 Patienten geheilt.'

Ein ungeübter Schüler liest hieraus nur die Aufgabe ab: Wie groß ist die Wahrscheinlichkeit, daß von 24 Patienten 20 geheilt werden, wenn die Heilungschancen im Einzelfall 70 % sind?
Diese gewohnte Fragestellung behindert das Verständnis für das Vorgehen beim Hypothesentest.

Nach meiner Erfahrung treten diese Schwierigkeiten bei Schülern nicht auf, wenn sie in den Unterrichtsstunden vor der Einführung mit Wahrscheinlichkeitsberechnungen für Bereiche befaßt wurden, wie dies bei der Beschäftigung mit ein- und zweiseitigen σ-Umgebungen (vgl. Beispiel 1) auf 'natürlich' Weise geschieht.

Einen weiteren Vorteil des Zugangs zum Problemkreis des Hypothesentests sehe ich im folgendem:

In der Bearbeitung von Beispiel 2 wurden nur Rechnungen für $p=0.7$ durchgeführt. Eigentlich müßten Rechnungen auch für $p<0.7$ folgen, da ja die Hypothese p 0.7 getestet wird. Natürlich wird man sich später (ähnlich wie oben) mit dem Kommentar begnügen:

'Für $p=0.7$ ist $P(X\leq 20) = 0.9576$, für $p<0.7$ ist dann auch $P(X\leq 20) < 0.9576$.'

Das will man aber an einem Beispiel erläutern: Wie umständlich und aufwendig ist die Berechnung von $P(X\leq 20)$ für z.B. $p=0.69$ im Vergleich zur Berechnung des kritischen Werts $\mu + 1.64\sigma$ für $p=0.69$ bei großem Stichprobenumfang:

Zusammenfassend zum Binomialtest sei also die Empfehlung wiederholt, den Binomialtest mit Problemen bei großem Stichprobenumfang voranzustellen, um dann, wenn das Verfahren, die Argumentationsweise, deutlich geworden ist, auf Probleme mit kleinem n einzugehen. Die Einführung in das Verfahren des Hypothesentests sollte nicht durch Probleme mit dem Berechnen von Wahrscheinlichkeiten belastet werden - die Probleme mit der Denkweise sind für die Schüler oft groß genug.

3. Kombinatorik und Exakter Test von FISHER

Es ist gelegentlich der Vorschlag gemacht worden, unmittelbar nach Behandlung der Kombinatorik in den Problemkreis des Hypothesentests einzusteigen. Die rechnerischen Voraussetzungen für dieses Testverfahren stehen scheinbar nach der Kombinatorik bereit. Dieser Zugang soll am folgenden Beispiel erläutert werden:

Beispiel 3:

Es handelt sich wie in Beispiel 1 und 2 um die Frage, wie man entscheiden soll, ob ein neues Medikament besser ist als die bisherigen. Psychologische Nebeneffekte im Bereich der Medizin sind nicht selten, und so könnte eine Information an die Patienten 'Dieses Medikament ist eine Weiterentwicklung der bisherigen.' dazu beitragen, daß bei einer größeren Anzahl schnelle Heilungserfolge sichtbar werden (oder auch genau den gegenteiligen Effekt

haben). Diese Wirkung der Information könnte man dadurch ausschließen, daß man diese Information bekanntgibt, bei jedem zweiten Patienten jedoch ein herkömmliches Medikament anwendet. Das Versuchsergebnis kann dann in Form einer Vierfeldertafel dargestellt werden.

Läßt sich z.B. aus der folgenden Tafel ablesen, daß die neue Salbe signifikant besser ist als die herkömmlichen Salben?

	Heilung innerhalb von 10 Tagen	Keine Heilung	
neue Salbe	11	4	15
herkömml. Salbe	9	5	14
	20	9	29

Wahrscheinlichkeitsberechnungen erfolgen hier nach hypergeometrischem Ansatz für die Zufallsgröße

Z: Anzahl der geheilten Patienten in der Gruppe der Patienten, die mit der neuen Salbe behandelt wurden (genauer: bei einer Gesamtzahl von 20 geheilten Patienten)

Dann ist:

$$P(Z=k) = \frac{\binom{15}{k} \cdot \binom{14}{20-k}}{\binom{29}{20}}, \quad k=6, \ldots, 15$$

Wie in Beispiel 2 verstellt die Angabe des Versuchsergebnisses den Zugang zur Lösung des Problems. Für Schüler liegt es nahe, P(Z=11) zu berechnen und mit dem Signifikanzniveau zu vergleichen. Dabei kommt es auch hier auf einen ganzen Bereich an, auf das 'extreme Ende'

einer Verteilung. Dazu sind alle möglichen Vierfeldertafeln zu betrachten, bei denen die Randwerte vorgegeben sind. Bezeichnen wir z.B. die Anzahl der geheilten Patienten, die mit neuer Salbe behandelt wurden mit k, dann ergibt sich die folgende 'allgemeine' Vierfeldertafel:

	Heilung innerhalb von 10 Tagen	Keine Heilung	
neue Salbe	k	15-k	15
herkömml. Salbe	20-k	k-6	14
	20	9	29

Auch bei dieser Verteilung müssen nacheinander P(Z=15), P(Z=14), usw. berechnet werden. Erst nach Bestimmung des 5 %-Endes der Verteilung kann man die Entscheidungsregel aufstellen und das konkrete Versuchsergebnis beurteilen:

P(z=15) = 0.0002
P(Z=14) = 0.0045
P(Z=13) = 0.0360
P(Z=12) = 0.1364

usw.

Wegen $P(Z \leq 12) = 0.9593$ verhalten wir uns wie folgt:

Wir verwerfen den Ansatz 'Die neue Salbe ist höchstens so gut wie die bisherigen', falls von den 20 geheilten mehr als 12 mit der neuen Salbe behandelt wurden.

Wenn das Verfahren des Exakten Tests von FISHER das erste Testverfahren ist, das die Schüler kennenlernen, treten die Schwierigkeiten auf, wie sie im Zusammenhang mit Beispiel 2 erwähnt wurden. Hinzu kommen jedoch erfahrungsgemäß Verständnisprobleme zu der Frage, warum die Wahrschein-

lichkeiten so berechnet werden, wie oben angegeben.

Berechnungen von 'hypergeometrischen' Wahrscheinlichkeiten im Rahmen der Kombinatorik sind zwar mit der Kenntnis der Binomialkoeffizienten in der Kombinatorik möglich und durchaus üblich, z.B. gehört es hier zu den Standardaufgaben, Wahrscheinlichkeiten für Lottogewinne auszurechnen. Bei der Übertragung der Modellbildung 'Gewinn in einem Rang beim Lottospiel' auf die 'Anzahl der geheilten Personen' im o.a. Beispiel treten bei den Schülern Schwierigkeiten auf, die nur durch genauere Analyse des Sachverhalts behoben werden können. Erst danach zeigte sich, daß Schüler das Verfahren des Exakten Tests von FISHER mit Verständnis anwenden können.

Wir betrachten dazu folgende Zufallsgrößen:

X: Anzahl der geheilten Patienten in der Gruppe der 15 Patienten, die mit der neuen Salbe behandelt wurden

Y: Anzahl der geheilten Patienten in der Gruppe der 14 Patienten, die mit der herkömmlichen Salbe behandelt wurden

S: Anzahl der geheilten Patienten insgesamt (d.h. in der Gruppe aller 29 Patienten) - also S = X + Y

Beim Test gehen wir von der Annahme aus, daß die neue Salbe höchstens so gut ist wie die herkömmlichen Salben. Die unbekannte Wahrscheinlichkeit p für die Heilung mit der neuen Salbe ist also bestenfalls genauso groß wie die (ebenfalls) unbekannte Wahrscheinlichkeit p' für die Heilung mit herkömmlichen Salben.
Für die Wahrscheinlichkeiten p und p' gilt also bestenfalls p=p':

Die Zufallsgrößen X, Y und S sind also binomialverteilt mit Erfolgswahrscheinlichkeit p:

$$P(X=k) = \binom{n}{k} p^k \cdot (1-p)^{n-k} \quad \text{(hier: } n=15\text{)}$$

$$P(Y=r) = \binom{m}{r} p^r \cdot (1-p)^{m-r} \quad \text{(hier: } m=14\text{)}$$

$$P(S=s) = \binom{n+m}{s} p^s \cdot (1-p)^{n+m-s} \quad \text{(hier: } n+m=29\text{)}$$

Wegen der angenommenen Unabhängigkeit der Zufallsgrößen X und Y berechnet sich die Wahrscheinlichkeit für k Heilungen mit neuer Salbe bei s Heilungen insgesamt wie folgt:

$$P(X=k,\ S=s) = P(X=k,\ Y=s-k)$$

$$= \binom{n}{k} p^k (1-p)^{n-k} \binom{m}{s-k} p^{s-k} (1-p)^{m-s+k}$$

$$= \binom{n}{k}\binom{m}{s-k} p^s (1-p)^{m+n-s}$$

Hieraus ergibt sich die bedingte Wahrscheinlichkeit $P_{S=s}(X=k)$:

$$P_{S=s}(X=k) = \frac{\binom{n}{k}\binom{m}{s-k}}{\binom{n+m}{s}}$$

Dies ist die Wahrscheinlichkeit dafür, daß k von den mit neuer Salbe behandelten Patienten geheilt werden unter der Voraussetzung, daß insgesamt s geheilt werden.

Diese Interpretation und Herleitung des FISHERschen Verfahrens geht erheblich über die bloßen Vorkenntnisse der Kombinatorik hinaus. Eine Behandlung kann - wegen des dargelegten Sachverhalts - erst nach der Binomialverteilung erfolgen.

4. Verteilungen zum Abzählen: Rangsummen

Nachdem sich die Methode des Exakten Tests von FISHER als schwieriger herausgestellt hat als man zunächst vermuten konnte, fällt es schwer, eine weniger bekannte Verteilung und das zugehörige Testverfahren als elementarer anzukündigen. Betrachten wir wieder ein Beispiel:

Beispiel 4:

Auch hier geht es wieder um eine Situation wie in den vorangegangenen Beispielen: Zwei medizinische Behandlungsmethoden sollen verglichen werden.

Die Situation muß so beschaffen sein, daß man die Wirkungen an den verschiedenen Patienten in eine Rangfolge bringen kann. (Denkbar wäre z.B. eine Rangfolge gemäß der Reihenfolge, in der die Heilung sichtbar wird:)

Betrachten wir einen Versuch, bei dem jeweils 6 Patienten nach zwei Methoden (A bzw. B) behandelt werden. Aus dem Versuch ergeben sich Meßwerte (z.B. für die Dauer des Heilungsprozesses), die miteinander verglichen werden können. Bei dem zu entwickelnden Testverfahren kommt es jedoch nicht auf die Meßwerte sondern auf die sich daraus ergebende Reihenfolge/Rangfolge an, z.B. ABAABABBAABB (auf dem ersten Rang soll die beste Heilung stehen). Hieraus sollen dann Schlüsse gezogen werden.

Bei welcher Anordnung wird man die Hypothese 'Methode A ist höchstens so gut wie Methode B' verwerfen?

Es ist vielleicht etwas schwierig, nachdem soviel vom Hypothesentest die Rede war, sich vorzustellen, wie Schüler reagieren, wenn man zu einem frühen Zeitpunkt des Stochastikunterrichts, etwa nach der Einführung von Zufallsgrößen, dieses Problem vorlegt.

Unterrichtserfahrungen zeigen immer wieder einen ähnlichen Verlauf der Diskussion:

Zunächst wird vorgeschlagen, die betrachteten Meßwerte zu addieren (oder bei ungleicher Anzahl den Mittelwert zu bilden).
Was soll aber ein Maß für den Vergleich der beiden Summen/Mittelwerte sein? Wird man eine Methode als 'wesentlich besser' bezeichnen, wenn sich bei der einen der Mittelwert 6.4 (Tage) ergibt, bei der anderen 5.9 (Tage)?

Hier fehlt ein Maß für die Beurteilung des Unterschieds, wie es etwa die Standardabweichung bei binomial- oder normalverteilten Zufallsgrößen darstellt.

Die Vernachlässigung der einzelnen Meßwerte und die Betrachtung nur der Rangfolge der Meßwerte führt trotzdem zu einem brauchbaren Verfahren.

Auf die Frage, bei welcher Konstellation man Methode A als besser ansehen würde als Methode B, nennen die Schüler natürlich sofort die Anordnungen AAAAAABBBBBB, AAAAABABBBBB, AAAABAABBBBB, aber auch AAAABBAABBBB, AAAABABABBBB, AAABAABABBBB. Falls bei der Versuchsdurchführung eine der erstgenannten Anordnungen auftritt, gibt es keinen Streit darüber, diese Ergebnisse als so 'extrem' zu bezeichnen, daß man also Methode A als besser ansieht.

Bei der Suche nach einer Ordnung für die Rang-Konstellationen wird erfahrungsgemäß immer der Vorschlag gemacht, die Summe der Ränge zu betrachten. Mit Hilfe der Zufallsgröße 'Rangsumme' lassen sich Versuchsergebnisse bewerten:

Wir betrachten z.B. nur die Summe der Ränge, die Meßwerte nach Methode A einnehmen:

Demnach hat z.B. die Rang-Konstellation AABABAABBBAB die Rangsumme 31, da die A's auf den Rängen 1, 2, 4, 6, 7 und 11 stehen. Der kleinstmögliche Wert ist: 1+2+3+4+5+6 = 21, der größtmögliche Wert ist: 7+8+9+10+11+12 = 57.

Allgemein läßt sich sagen: Je mehr A's vorne (auf den ersten Rängen) stehen, desto kleiner ist die Rangsumme. Kleine Rangsummen sprechen also dafür, daß Methode A besser ist als Methode B.

Für die Rangsumme 21 gibt es nur die genannte Möglichkeit, für die Rangsumme 22 nur die Zerlegung 1+2+3+4+5+7, dagegen für die Rangsumme 23 schon zwei Möglichkeiten: 1+2+3+4+5+8 bzw. 1+2+3+4+6+7, usw.

Was hat dies mit Wahrscheinlichkeiten zu tun?

Sehen wir beide Methoden als gleich gut an, dann haben alle möglichen Rang-Konstellationen die gleiche Wahrscheinlichkeit, da es nur vom Zufall abhängt, welche Reihenfolge sich ergibt.

Es gibt $\binom{12}{6} = 924$ Möglichkeiten, 6 Ränge von 12 möglichen auszusuchen (Anzahl der 6-elementigen Teilmengen einer 12-elementigen Menge): Jeder Rang-Konstellation kommt daher - unter der Annahme der Gleichwertigkeit der

Rangkonstellationen	Rangsumme	kumulierte Wahrscheinl.
1 2 3 4 5 6	21	1 / 924
1 2 3 4 5 7	22	2 / 924
1 2 3 4 5 8 1 2 3 4 6 7	23	4 / 924
1 2 3 4 5 9 1 2 3 4 6 8 1 2 3 5 6 7	24	7 / 924
1 2 3 4 5 10 1 2 3 4 6 9 1 2 3 4 7 8 1 2 3 5 6 8 1 2 4 5 6 7	25	12 / 924
1 2 3 4 5 11 1 2 3 4 6 10 1 2 3 4 7 9 1 2 3 5 6 9 1 2 3 5 7 8 1 2 4 5 6 8 1 3 4 5 6 7	26	19 / 924
1 2 3 4 5 12 1 2 3 4 6 11 1 2 3 4 7 10 1 2 3 4 8 9 1 2 3 5 6 10 1 2 3 5 7 9 1 2 3 6 7 8 1 2 4 5 6 9 1 2 4 5 7 8 1 3 4 5 6 8 2 3 4 5 6 7	27	30 / 924
1 2 3 4 6 12 1 2 3 4 7 11 1 2 3 4 8 10 1 2 3 5 6 11 1 2 3 5 7 10 1 2 3 5 8 9 1 2 3 6 7 9 1 2 4 5 6 10 1 2 4 5 7 9 1 2 4 6 7 8 1 3 4 5 6 9 1 3 4 5 7 8 2 3 4 5 6 8	28	43 / 924
1 2 3 4 7 12 1 2 3 4 8 11 1 2 3 4 9 10 1 2 3 5 6 12 1 2 3 5 7 11 1 2 3 5 8 10		

Methoden - die Wahrscheinlichkeit 1/924 zu.

Was jetzt noch fehlt, ist die Einigung darüber, bis zu welcher Rangsumme man Methode A als besser ansieht als Methode B. Genannt werden hier von Schülern meist irgendwelche Grenz-Rangsummen, manchmal auch Prozentsätze. Hier muß dann nach einer gewissen Zeit der Diskussion der Lehrer mitteilen, daß es **üblich** ist, solche Bereiche am Ende der Verteilung zu betrachten, denen ein bestimmter Prozentsatz zukommt, z.B. 5 %.

Gesucht werden jetzt nur noch alle die Rang-Konstellationen mit möglichst kleiner Rangsumme, die zum unteren 5 %-Ende der Verteilung gehören: dies sind ca. 46 (= 5% von 924 gleichwahrscheinlichen) Konstellationen.

Man kann nicht genau 5 % des unteren Endes der Verteilung vernünftig abschneiden, da von der 44. Rang-Konstellation an weitere 18 zur Rangsumme 29 gehören. Fest steht:

Falls beide Methoden gleichwertig sind, wird es zufällig in etwas weniger als 5 % der Fälle zu einer Rang-Konstellation kommen, deren Rangsumme kleiner als 29 ist. Wenn wir das 5 %-Ende als 'extrem' ansehen, dann verhalten wir uns bei der Auswertung des Versuchs nach der folgenden Regel: Wir halten Methode A für besser als Methode B, wenn die Rangsumme kleiner als 29 ist.

Das ist nichts anderes als die gesuchte Entscheidungsregel!

Vergleichen wir den Aufwand bei Rangsummenverteilungen mit dem bei anderen Verteilungen, dann wird deutlich, daß man bei keinem der vorher angesprochenen Beispiele

mit solch geringem Aufwand zur Lösung des Problems gelangt waren, eine vernünftige Strategie zur Beurteilung von Stichprobenergebnissen zu finden.

Beim zuletzt angesprochenen Rangsummentest - das Verfahren geht auf WILCOXON zurück und gehört zu den sogenannten parameterfreien Testverfahren - braucht man (bei einseitigen Fragestellungen) nur alle Rang-Konstellationen am oberen (oder unteren) Ende der Verteilung aufzuschreiben, um den Bereich zu bestimmen, den man in der Terminologie des Hypothesentests als Verwerfungsbereich bezeichnet.

Wählt man die Untersuchung von Rangsummen für den Einstieg in die Beurteilende Statistik, dann sind weitergehende Betrachtungen nicht ausgeschlossen: Zur Anwendung des Verfahrens gehört auch die Kenntnis der Fehlermöglichkeiten:

Wenn beide Methoden gleichwertig sind, dann kommt es zufällig in z.B. 5 % der Fälle zu Rang-Konstellationen, die zum so definierten Verwerfungsbereich gehören (Fehler 1. Art); andererseits ist denkbar, daß ein Versuch keine extreme Rang-Konstellation hat, obwohl der o.a. Ansatz nicht richtig ist (Fehler 2. Art).

5. Vergleiche und Anregungen zum Unterricht

Abschließend sollen die vier hier angedeuteten Zugänge zum Hypothesentest noch einmal im Vergleich betrachtet werden. Gleichzeitig ergeben sich hierbei Empfehlungen für den Unterricht über dieses Thema.

In allen hier vorgestellten Beispielen (und den meisten denkbaren zu diesem Thema) geht es in einer einseitigen Fragestellung um den Vergleich zweier Methoden.

Die Behauptung (oder auch wünschenswerte Aussage) ist in allen Fällen: Eine bestimmte (neue) Methode ist besser als die andere (alte).

Die Untersuchung erfolgt immer durch einen indirekten Ansatz, die Annahme: Die eine (neue) Methode ist höchstens ebenso gut wie die andere (bisherige).

Die Untersuchungen nach den verschiedenen Testverfahren haben Gemeinsamkeiten, zeigen aber auch unterschiedliche Aspekte auf:

<table>
<tr><th></th><th>Art des Vergleichs</th><th>Art des Versuchs-ergebnisses</th><th>Verwerfungs-bereich</th></tr>
<tr><td>Binomial-test</td><td>Vorkenntnisse über alte Methode (Vert.)</td><td rowspan="2">Anzahl der Ausprägungen von zwei Merkmalen z.B. positiv / negativ</td><td>oberes Ende einer Binomialvert.</td></tr>
<tr><td>FISHER-Test</td><td rowspan="2">direkter Vergleich zwischen den Methoden (Verteilungen)</td><td>oberes Ende einer Hypergeom.Vert.</td></tr>
<tr><td>Rang-summen-test</td><td>Meßreihe, die eine Anordnung zuläßt</td><td>unteres Ende einer Rang-summenvert.</td></tr>
</table>

(Anmerkung zum Verwerfungsbereich des Rangsummentests: Ordnet man dem schlechtesten Ergebnis der Meßreihe den 1. Rang zu, liegt der Verwerfungsbereich entsprechend am oberen Ende der Rangsummenverteilung.)

Wie oben dargestellt, eignen sich verschiedene Testverfahren zum Einstieg in den Themenkreis des Hypothesen-

tests innerhalb der Beurteilenden Statistik. Am elementarsten erscheint der Zugang über die Rangsummenverteilung, da hier der rechnerische Aufwand nicht von der Fragestellung ablenkt. Eine ähnliche Voraussetzung ist beim Binomialtest mit großem Stichprobenumfang gegeben: Hier müssen Rechnungen nur in geringem Umfang durchgeführt werden, so daß der Hauptaugenmerk auf den argumentativen Teil gerichtet werden kann. Wegen der (eigentlich) vorauszusetzenden theoretischen Kenntnisse erscheint der Exakte Test von FISHER am wenigsten für den Einstieg geeignet. Gleichwohl erscheint es sinnvoll, sich nicht nur auf ein Verfahren zu beschränken; durch Behandlung mehrerer Verfahren wird eine deutlich spürbare Festigung bei den Schülern erreicht.

Daß der Abschnitt über das Testen von Hypothesen im Unterricht vorbereitet werden muß, wurde bereits oben deutlich gemacht. Erprobungen haben allerdings gezeigt, daß die Merkregel 'Auf Bereiche kommt es an, nicht auf Einzelwahrscheinlichkeiten' nur bei den Binomialtests und dem Exakten Test von FISHER eine Rolle spielt. Dies ist wohl damit zu erklären, daß bei den zuletzt genannten Verfahren die Bestimmung der Wahrscheinlichkeitsverteilungen notwendig zur Vorbereitung gehört, während sich die Frage nach dem Bereich beim Rangsummentest automatisch aus der Frage 'Bei welchen Rang-Konstellationen wird man ... als besser ansehen?' ergibt.

Bezüglich der Art der Fragestellung sollte man im Rahmen des Unterrichts die mögliche Spannbreite ausnutzen: Die Frage nach einer Entscheidungsregel (Ab welcher Anzahl/ Rangsumme soll man ... verwerfen / ... für besser halten?) ist die grundsätzliche Fragestellung - sie steht eigentlich immer im Zentrum eines Hypothesentests. Gleichwohl

ist die Vorgabe einer konkreten Situation (Wie soll man ein Stichprobenergebnis ... beurteilen?) für den Unterricht attraktiver, da erfahrungsgemäß lebhaftere Diskussionen unter den Schülern entstehen. (Natürlich darf im Unterricht nicht der Hinweis fehlen, daß eine Hypothese nicht erst nachträglich und passend zum Stichprobenergebnis formuliert werden darf.)

Zu den wesentlichen Zielen des Unterrichts gehört es auch, die Methode des Testens von seiner logischen Seite her zu untersuchen. Durch die (naive) Fragestellung 'Bei welchem Stichprobenergebnis sehen wir ... als besser an?' kommt man zwar auf natürliche Weise zu den extremen Enden einer Verteilung. Es bleibt aber die Analyse des Vorgehens, die Erläuterung des skeptischen Standpunkts in der indirekten Argumentation.

In den Ausführungen beschäftigen wir uns nur mit einseitigen Fragestellungen. Solche sind in der Praxis von größerer Bedeutung als die zweiseitigen. Manche Schüler haben jedoch erfahrungsgemäß einen leichteren Zugang zum Thema, wenn man mit zweiseitigen Tests beginnt. Aus der Fragestellung 'Ist die eine Methode besser als die andere?' wird dann 'Ist die eine Methode genauso gut wie die andere?'. Der Unterschied in der Behandlung liegt nicht nur in der Tatsache, daß nunmehr beide Enden der Verteilung betrachtet werden müssen; Schwierigkeiten bei einseitigen Fragestellungen können bei Schülern auch in der Tatsache liegen, daß man sich bei einseitigen Fragestellungen oft auf die Untersuchung einer Verteilung (p=p) beschränkt, obwohl die Hypothese z.B. $p < p'$ lautet, man also eigentlich (unendlich) viele Verteilungen betrachten müßte. Dies ist für den Exakten Test von FISHER und den Rangsummentest prinzipiell gar nicht

möglich (die Verteilungen können nur für den Fall p=p' aufgestellt werden, s.o.) ; beim Binomialtest sind selbst beispielhafte Rechnungen lästig. Man wird sich also meistens argumentativ mit der Einseitigkeit auseinandersetzen (... ist bestenfalls so gut wie ...).

Literatur:

1. A.ENGEL: Wahrscheinlichkeitsrechnung und Statistik, Band 1, Klett-Verlag, Stuttgart, 1973

2. H.WEGMANN / J.LEHN: Einführung in die Stochastik, Vandenhoeck & Ruprecht, Göttingen, 1984

3. J.BRUHN / H.K.STRICK: Mathematik heute - Leistungskurs Stochastik, Schroedel Schulbuchverlag, Hannover, Schülerband 1984, Lösungsband 1986

4. H.K.STRICK: Einführung in die Beurteilende Statistik, Schroedel Schulbuchverlag, Hannover, bearbeitete Ausgabe 1986

5. H.K.STRICK: Parameterfreie Verfahren im Stochastikunterricht, MNU, Heft 3/1982

6. H.K.STRICK: Kombinatorik auf dem elektronischen Rechner, MU, Heft 1/1984

Anschrift des Verfassers:
Heinz Klaus Strick, Pastor-Scheibler-Str. 10,
5090 Leverkusen 3.