Übungen zum subjektiven Zugang zu Wahrscheinlichkeiten

von **T. Nemetz** und **N. Kusolitsch**
(bearbeitet von **G. Ihorst**, Dortmund)

Zusammenfassung: In dieser Arbeit werden einige Übungen vorgestellt, die den Schülern ein Gefühl für subjektive Wahrscheinlichkeiten geben sollen. Diese Übungen basieren auf Methoden, die Shannon (1951) entwickelt hat, um den Informationsgehalt einer Sprache (die Entropie) zu schätzen. Die Lernziele dieser Übungen sind die folgenden: Erstens soll den Schülern gezeigt werden, wie statistisch abhängige bzw. unabhängige Daten aussehen. Zum zweiten soll der Unterschied geklärt werden, wie man zu einer statistischen Entscheidung gelangt im Fall von hochgradig abhängigen Daten und im Fall von nahezu unabhängigen Daten. Daneben wollen wir Übungen vorstellen, für die statistische Daten leicht beschafft werden können.

Übung 1: (Shannons Ratespiel)

Der Lehrer wählt zufällig einen Text vorgegebener Länge (z.B. 30 oder 50 Buchstaben) aus einem kaum bekannten Buch oder einer Zeitung aus. Dieser Text - außer dem letzten Buchstaben - wird dem Schüler mitgeteilt. Der Schüler muß nun diesen letzten Buchstaben raten, wobei er immer genau einen Buchstaben vorschlagen kann, etwa a, e, x, oder Leerzeichen. Der Lehrer sagt ihm, ob er richtig geraten hat oder nicht. Wenn nicht, muß der Schüler noch einmal raten, bis er den richtigen Buchstaben gefunden hat. Die Anzahl der erforderlichen Rateversuche wird gezählt. Für diese Übung schlagen wir etwa 100 unabhängig ausgewählte Texte vor. Die relativen Häufigkeiten der Versuche, die mit genau i Rateversuchen beendet waren, werden mit $f_i, i = 1, ..., 27$, bezeichnet. Die Größe

$$H(f_1, ..., f_{27}) := -\sum_{i=1}^{27} f_i \log f_i$$

gibt dann eine gute Approximation der Entropie an. Die Logarithmen werden zur Basis 2 berechnet, und es ist $\log 0 = 0$. Natürlich soll man den letzten Buchstaben jedes Textes so schnell wie möglich herausfinden. In jeder europäischen Sprache erreichen Spieler, die ein gutes Sprachgefühl haben, einen Wert um 1.3. Bei guten Schülern

kann man einen Wert bis zu 1.5 erwarten; Ergebnisse unter 2 sind ganz akzeptabel.

Das Lernziel dieser Übung ist, daß Schüler die möglichen Ausgänge eines Experiments nach der Größe ihrer Wahrscheinlichkeiten in sehr ähnlichen, aber verschiedenen Situationen ordnen. Tabelle 1 enthält die Ergebnisse von 5 Schülern, die den letzten Buchstaben in 135 Texten raten mußten, die aus jeweils 30 Buchstaben bestanden.

Übung 1a: (Modifikation von Übung 1)

Der Lehrer wählt wieder einen Text von vielleicht 150 Buchstaben aus. Dieser Text wird dem Schüler bis zu einem festgelegten Punkt mitgeteilt, z.B. dem 30. Buchstaben. Dann wird der Schüler gebeten, den 31. Buchstaben zu raten. Er rät so lange, bis er ihn gefunden hat. Anschließend muß er dann den 32. Buchstaben raten, und so weiter.

Übung 1b:

Der Basis-Text, der erraten werden soll, besteht jetzt nur noch aus jedem 10. Buchstaben eines sinnvollen Textes. Ansonsten wird diese Übung genau wie Übung 1a durchgeführt. Tabelle 2 zeigt ein konkretes Beispiel für diese Übung. Die erste Reihe enthält dabei die ausgewählten Buchstaben, und unter jedem Buchstaben stehen die Rateversuche eines Schülers. Diese Übung stellt eine Methode dar, die Entropie zu schätzen, und sie beschreibt die Redundanz einer Sprache.

Schüler	H	f_1	f_2	f_3	f_4	$f_{>4}$
S1	1.73	96	14	5	3	17
S2	1.79	95	13	6	2	19
S3	1.84	94	13	6	4	18
S4	1.85	93	13	8	5	16
S5	2.08	83	18	11	7	16

Tabelle 1

ausgew. Buchstaben	H	I	E	A	A	L	S	D
Rateversuche	A ETO NISR	ETA ON		T	ET	TANS EOI	.	HRSI ALEN L
Anzahl der erforderlichen Versuche	10	7	2	3	4	11	1	11

Tabelle 2

Diese Übungen sind relativ einfach. Die Schüler müssen lediglich die Buchstaben nach der Größe der Wahrscheinlichkeit ihres Auftretens ordnen, sie brauchen diese Wahrscheinlichkeiten selbst aber nicht zu schätzen. Wir glauben daher, daß die Übungen sich gut zur Vorbereitung von numerischen Bewertungsspielen eignen.

Ein spielerischer Zugang zur Schätzung der Entropie

Von Cover und King (1978) wurde eine proportionale Spiel-Methode vorgeschlagen, von der die Autoren zeigen konnten, daß sie - auf lange Sicht - optimal ist. Wir zitieren aus Cover und King: "Das Wichtigste in Hinsicht auf den Spiel-Schätzer ist es, ein optimales Spiel-Schema herauszufinden. Anstatt Symbole zu raten und die Versuche bis zum richtigen Raten zu zählen (wie in Shannons Verfahren), setzt der Spieler einen Anteil seines momentanen Kapitals. Dieser Anteil ist proportional zur bedingten Wahrscheinlichkeit des nächsten Symbols im Alphabet, gegeben die vorangegangenen Symbole. Dieser Prozeß wird mit aufeinanderfolgenden Symbolen im Text wiederholt, wobei der Spieler nach n Einsätzen S_n Dollar angesammelt hat. Wenn wir einen idealen Spieler haben, der sein Kapital bei jedem Einsatz gemäß der wahren Wahrscheinlichkeitsverteilung des nächsten Symbols aufteilt, dann gilt:

$$\left(1 - \frac{1}{n}\log_{27} S_n\right) \cdot \log_2 27$$

konvergiert fast sicher gegen die Entropie.

Die Ideen von Cover und King werden für die folgenden Übungen verwendet:

Übung2:

Wie in Übung 1a wird ein Text ausgewählt, und man gibt den Schülern die ersten etwa 30 Buchstaben bekannt. Dann sollen sie Einsätze festlegen auf die Trichotomie: Vokal - Konsonant - Leerzeichen. Der Buchstabe wird dann verraten, und die Prozedur wird mit den folgenden Buchstaben wiederholt. Das Kapital auf die Einsätze für die Trichotomie aufzuteilen ist dabei äquivalent dazu, Anteile für die drei Fälle festzulegen. Auf diese Weise bewerten die Schüler die drei Wahrscheinlichkeiten. Die Beurteilung dieser Übung kann dann gemäß Dowie (1984) erfolgen. Dowies Kommentare treffen auch hier zu, und das Ziel dieses Spieles ist ebenfalls im ersten Abschnitt von Dowies Arbeit beschrieben. Nemetz und Simon (1978) haben dieses Spiel in seiner Originalform mit verschiedenen Altersgruppen durchgeführt. Sie haben herausgefunden, daß die jüngeren Schüler kaum mit der Größe des Alphabets zurechtkommen, ihre Einsätze zeigten große Variationen, und die Aufgabe war sehr zeitraubend. Deshalb schlagen sie vor, die Verteilung der Einsätze in zwei Schritten durchzuführen, wie im folgenden beschrieben wird:

Übung 2a:

Wie in Übung 1a wird ein Text ausgewählt, und den Schülern werden die ersten 30 Buchstaben genannt. Dann werden sie gefragt, welche Spalte im Tabelle 3 den nächsten Buchstaben enthält.

Die Schüler müssen nun die Einsätze ihrem Gefühl nach auf die Spalten verteilen, d.h. sie müssen subjektive Wahrscheinlichkeiten für die einzelnen Spalten festlegen. Ihnen wird dann die korrekte Spalte mitgeteilt. Danach, wenn sie also die richtige Spalte kennen, müssen sie ihre subjektiven Wahrscheinlichkeiten festlegen für die Buchstaben dieser Spalte. Die Bewertung erfolgt wieder wie oben. Tabelle 3 wurde konstruiert gemäß den Häufigkeiten der Buchstaben im englischen Alphabet (vgl. Gaines (1956, S. 218)). Dabei ist nicht unbedingt klar, ob Tabelle 3 die beste Aufteilung des englischen Alphabets darstellt. Möglicherweise wäre eine Anordnung, in der die Vokale eine Gruppe bilden, psychologisch besser geeignet.

Übung 2b:

Wie in Übung 1b wird nur jeder zehnte Buchstabe eines Textes verwendet anstatt eines sinnvollen Textes. Ansonsten verläuft die Übung analog zu Übung2.

Übung 2c:

Wieder wird nur jeder zehnte Buchstabe eines Textes verwendet, und die Übung wird analog zu Übung 2a durchgeführt.

	I	*II*	*III*	*IV*	*V*
1		N	L	F	G
2	E	I	D	M	V
3	T	S	C	W	K
4	A	R	U	Y	Q, X
5	O	H	P	B	J, Z

Tabelle 3

Lernziele der Übungen 2b, 2c

Mit diesen Übungen wollen wir zeigen, daß man die Kenntnis der vorhergehenden Buchstaben nicht in Betracht ziehen kann, wenn man die subjektiven Wahrscheinlichkeiten festlegt. Außerdem beabsichtigen wir zu zeigen, daß in diesem Fall die subjektiven Wahrscheinlichkeiten übereinstimmen sollten mit den entsprechenden relativen Häufigkeiten der Buchstaben in einem langen Text. Wir schlagen daher vor, einen Text von etwa 2000 Buchstaben in einem Computer abzuspeichern und daraus die Häufigkeitsverteilung der Buchstaben zu berechnen.

Abschließende Bemerkungen

Neben den bereits erwähnten Lernzielen möchten wir die folgenden Punkte betonen:

(1) Eine Quelle für die statistischen Daten für diese Übungen ist in fast jedem Land der Welt schnell verfügbar.

(2) Die Übungen erlauben den Einsatz von Computer-Programmen. Die Zeit, die den Schülern für ihre Entscheidungen zur Verfügung steht, kann dabei als Teil des computergestützten Lehrprogrammes ihren Fähigkeiten angepaßt werden.

(3) Die Übungen können mit relativ jungen Personen durchgeführt werden. Schüler zwischen 11 und 14 haben sehr viel Spaß daran, aber die Aufgaben erregen auch das Interesse der Älteren.

Literaturverzeichnis

Cover, T.M. und King, R.C. (1978). A convergent Gambling Estimate of the Entropy of English, *IEEE, Trans. Inf. Th., IT-24*.

Dowie, J. (1984). Anyone for a Bayesian Wimbledon? *Teaching Statistics, Vol.6, No. 3.*

Gaines, H.G. (1956). *Cryptanalysis*, Dover.

Nemetz, T. und Simon, J. (1978). On Estimating the Entropy of Written Hungarian by Gambling Technique, *Trans. of 8th Prague Conf. on Inf. Th. Vol. B.*.

Shannon, C.E. (1951). Prediction and Entropy of Printed English, *Bell Syst. Tech. J.*

Original unter dem Titel "Exercises to improve the subjective approach to probability" in *Teaching Statistics, 8(3), 1986, S. 78-82.*