DIE ABHÄNGIGKEIT ZWISCHEN ZWEI GRÖSSEN GEMESSEN ALS RELATIVE VERLUSTVERRINGERUNG

nach R. EISENBACH & R. FALK, Tel-Aviv University, Hebrew University of Jerusalem
Originaltitel in 'Teaching Statistics' Vol. 6 (1984), Nr. 2:
Association Between Two Variables Measured as Proportion of Loss-Reduction
Übersetzung: J. Feuerpfeil

Ein Interpretationsproblem

Eine der Schwierigkeiten in der Statistik ist für Schüler die Interpretation von Abhängigkeitsmaßen. Die naheliegende Frage "Was bedeutet r = 0,7?" ist ein Beispiel. Der Korrelationskoeffizient wird meist als "Maß für die lineare Abhängigkeit" eingeführt; es wird erklärt, daß r als maximalen Wert 1 im Fall vollständiger linearer Abhängigkeit bzw. als minimalen Wert 0 bei Fehlen jeglicher derartiger Abhängigkeit annehmen kann; Werte zwischen diesen beiden Extremen sind als Zwischenstufen der Abhängigkeit zu verstehen.

Dennoch geben sich die meisten Schüler mit dieser Interpreation nicht zufrieden. Einige ersetzen sie durch ihre eigene i.a. falsche, aber "sinnvollere" Interpretation. Da eine Maßzahl zwischen 0 und 1 an eine relative Häufigkeit (oder auch an eine Wahrscheinlichkeit) erinnert, glauben manche Schüler irrigerweise, es sei der Anteil der identischen Paare (x;y) oder auch die Wahrscheinlichkeit für richtige Vorhersagen gemeint.

In diesem Aufsatz empfehlen wir eine klare und sinnvolle Interpretation verschiedener Abhängigkeitsmaße. Wir schlagen die Konstruktion einiger auf demselben Prinzip beruhender Abhängigkeitsmaße vor, nämlich Maße für die Vorhersagbarkeit, ausgedrückt mit Hilfe der Häufigkeit von Fehlern, die sich durch ein bestimmtes Vorhersageverfahren vermeiden lassen. Die allgemeine Definition, die auf diesem Prinzip beruht, führt zu verschiedenen Abhängigkeitsmaßen. So konstruierte Maße unterscheiden sich dadurch voneinander, daß sie von der besonderen Art der betreffenden Größen und den verwendeten Vorhersageverfahren abhängen.

Vorhersagbarkeit, ausgedrückt mit Hilfe der Verlustverringerungen

Die Abhängigkeit zwischen zwei Größen x und y kann als Vorhersagbarkeit der y-Werte bei Kenntnis der x-Werte angesehen werden. Die Unsicherheit bezüglich der y-Werte wird umso mehr verringert je besser die Möglichkeit der Vorhersage ist, wenn man sich auf die x-Werte stützt. Man spricht von statistischer Vorhersage (von y durch x), wenn y zufällige Werte annehmen kann und daher eine wahllose Schätzung dieser Werte zu Fehlern und Ungenauigkeiten führt. Wir schlagen für die Abweichung zwischen Schätzung und Realität allgemein den Begriff "Verlust" vor. Die Bezeichnung für den Verlust bei der Größe y sei V(y). Er nimmt entsprechend der Definition der Verlustfunktion (und entsprechend dem Wert der besten Schätzung) verschiedene Formen an. Möglichst gute Vorhersage der y-Werte unter Verwendung der Kenntnis der x-Werte führt im allgemeinen noch immer zu einem gewissen, wenn auch verringerten Verlust. Der bedingte Verlust aufgrund der Abweichung zwischen den vorhergesagten und dem tatsächlichen Werten wird mit V(y/x) bezeichnet. Es gilt $V(y/x) \leqq V(y)$. Die Differenz $V(y) - V(y/x)$ ist ein Maß für die Verlustverringerung. Der Quotient Θ aus Verlustverringerung und zu erwartendem Verlust ohne Vorhersageverfahren ist als Maß für die Vorhersagbarkeit von y durch x geeignet:

$$\Theta_{x/y} = \frac{V(y) - V(y/x)}{V(y)} \qquad (1)$$

Sowohl die Verlustfunktion als auch das Vorhersageverfahren müssen natürlich genau festgelegt sein. Die drei für die Definition des Verlusts üblichsten Funktionen sind: Anzahl der Fehler, Abweichung in absoluten Beträgen und Quadrate der Abweichungen. Ein wichtiger Faktor bei der Wahl einer Verlustfunktion ist der Grad der Meßbarkeit der Größe. Diese Wahl beeinflußt auch die "beste Schätzung", die den Verlust ohne Vorhersageverfahren minimiert (Falk, 1980). Handelt es sich z.B. um eine Merkmalsgröße, so wird man bestrebt sein, die "Fehleranzahl" zu minimieren, daher wird man den häufigsten Wert (Modalwert) als Schätzwert nehmen. Der optimale Schätzwert bringt den kleinstmöglichen Verlust.

Das gleiche Vorgehen läßt sich bei der Wahl eines Verfahrens zur

Vorhersage von y bei gegebenen x anwenden. So wird man bei einer zweidimensionalen Kontingenztafel mit dem Merkmalsgrößen x und y versuchen, die "Anzahl der Fehler" zu minimieren, das bedeutet: Als Vorhersage dient der häufigste Wert y bei gegebenem x. Auf diese Weise wird die Fehleranzahl für jeden einzelnen Merkmalswert von x minimiert und der Gesamtverlust wird im allgemeinen unter den unbedingten Verlust gedrückt.

Tabelle 1 zeigt ein Beispiel einer 2 x 2 - Kontingenztafel.

Tabelle 1. Zweidimensionale Häufigkeitsverteilung von Merkmalsgrößen

y \ x	A	B	Summe
C	30	15	45
D	5	50	55
Summe	35	65	100

Das Maß für die Vorhersagbarkeit von y durch x berechnet sich nach (1) folgendermaßen:

Verlustfunktion ist die "Fehleranzahl". Ohne Kenntnis von x ist die beste Schätzung für y das häufigste Merkmal D mit der Häufigkeit 55; der Verlust ist daher V(y) = N(Fehler/x unbekannt) = 45. Ist x bekannt, so wird C vorhergesagt, falls x zu A gehört, dagegen wird D vorhergesagt, falls x zu B gehört. Für beide Vorhersagemöglichkeiten zusammen erhalten wir:

$$V(y/x) = N(\text{Fehler/x bekannt}) = 5 + 15 = 20.$$

Das Maß θ in Formel (1) wird zum Goodman-Kruskal-Lambda (Hays, 1973, S. 745-749), das zum ersten Mal von Louis Guttman 1941 vorgeschlagen wurde.

$$\lambda_{y/x} = \frac{N(\text{Fehler/x unbekannt}) - N(\text{Fehler/x bekannt})}{N(\text{Fehler/x unbekannt})}$$

Für das Beispiel in Tabelle 1 erhalten wir:

$$\lambda_{y/x} = \frac{45 - 20}{45} = 0{,}556.$$

Die Interpretation des Ergebnisses $\lambda_{y/x} = 56\,\%$ ist:

Es ist uns gelungen, 56 % der Fehler beim Schätzen der y-Werte zu vermeiden, indem wir die Vorhersage gestützt auf die gegebenen x-Werte machen.

Tabelle 2 zeigt fünf Vorhersagbarkeitsmaße, die sämtlich nach demselben Prinzip konstruiert sind. Wir beginnen mit Nominalgrößen (sowohl x als auch y) und fahren der Reihe nach mit Skalen höheren Meßbarkeitsgrads fort.

Die Grundlagen für das Vorhersagbarkeitsmaß, V(y) und V(y/x) sind in Tabelle 2 jeweils mit den Verlustfunktionen und den optimalen Vorhersageregeln aufgelistet. Wegen der zunehmenden "Meßbarkeit" der Skalen von oben nach unten kann man die Maße von weiter oben stehenden Zeilen auf weiter unten stehende anwenden, aber nicht umgekehrt. Eine Ausnahme bildet der Fall, daß die Ordnung zwischen zwei Größenpaaren nicht eindeutig klar ist (Siehe Fußnote (a) zu Tabelle 2).

Wird die abhängige Größe mit einer Intervallskala gemessen, so ist die Summe der quadratischen Abweichungen die typische Verlustfunktion, und die beste Schätzung ist das arithmetische Mittel, bezüglich dessen die Verlustfunktion V(y) berechnet wird. Nimmt man bei gegebener unabhängiger Größe x jeweils den bedingten y-Mittelwert, so führt unser Maß θ zu dem vertrauten $\eta^2_{y/x}$. Jedoch wird bei Anwendung einer linearen Vorhersageregel der bedingte Verlust bezüglich der Regressionsgeraden der kleinsten Fehlerquadrate berechnet, und θ wird zum Quadrat des wohlbekannten Korrelationskoeffizienten r.

Die Verlustfunktion auf der Grundlage der absoluten Abweichungen wird üblicherweise bezüglich des Medians als der besten Schätzung berechnet. Unser Maß für Vorhersagbarkeit $\Delta_{y/x}$ (siehe zweite Reihe

Tabelle 2

Grad der Meßbarkeit x - y	Verlust-Funktion	V(y) Gesamtverlust ("beste Schätzung")	Optimale Vorhersage (y durch x)	V(y/x) bedingter Verlust	Maß für Vorhersagbarkeit $\Theta_{x/y} = \frac{(y) - (y/x)}{V(y)}$
Nominal-nominal	Anzahl der Fehler	N(Fehler/x unbek.) (Modus v. y)	Bedingter Modus	N(Fehler/x bek.)	Goodman-Kruskal-$\lambda_{y/x}$
Nominal-ordinal	Summe der absoluten Abweichungen	$\sum_j \sum_i /y_{ij}$-Md..| (Median v.y)	Bedingter Median	$\sum_j \sum_i /y_{ij}$-Md.$_j$/	Vorhersage $\Delta_{y/x}$
Ordinal-ordinal (n Paare) (a)	Anzahl der Inversionen	$\frac{1}{2}\binom{n}{2}$ (zufällige Reihenfolge)	Gleiche Reihenfolge wie bei x-Werten (b)	N($(i;j)$ $x_i<x_j \wedge y_i>y_j$)	Absolutwert von Kendalls Tau $\lvert\tau\rvert$
Nominal-Intervall (a)	Summe der quadratischen Abweichungen	$\sum_j \sum_i (y_{ij}-\bar{y}..)^2$ (Mittelwert von y)	Bedingter Mittelwert	$\sum_j \sum_i (y_{ij}-\bar{y}._j)^2$	Korrelation $\eta^2_{y/x}$
Intervall-Intervall	Summe der quadratischen Abweichungen	$\sum_i (y_i-\bar{y})^2$ (Mittelwert von y)	Regressionsgerade	$\sum_i (y_i-\hat{y}_i)^2$	Quadrat des Korrelationskoeffizienten r^2.

a) Die Reihenfolge zwischen diesen beiden Klassen ist nicht eindeutig

b) x ist entweder abhängig oder unabhängige Größe

Ist die Anzahl der Übereinstimmungen in den Rangzahlen der Paare größer oder gleich der Anzahl der Abweichungen, so dient die gleiche Reihenfolge wie bei den unabhängigen Größen als Vorhersage, andernfalls kehrt man die Reihenfolge der unabhängigen Größen vor der Vorhersage um.

von Tabelle 2) ist zwar nicht so weit verbreitet, aber es ist nach demselben Prinzip konstruiert und kann ähnlich interpretiert werden.

Gewöhnlich macht man sich nicht auf den ersten Blick klar, daß der absolute Wert des Kendallschen Tau - in Lehrbüchern (Siegel, 1956, S. 213-223, und Hays, 1973, S. 792-796) definiert als Differenz zwischen dem Anteil der Paare, die gleiche Rangzahlen haben, und dem Anteil der Paare, die verschiedene Rangzahlen d.h. Inversionen haben - eigentlich ein Spezialfall von Θ in Formel (1) ist. Eine einfache Umformung liefert:

$$\tau = \frac{\left(\text{Anzahl der Paare mit gleichen Rangzahlen}\right) - \left(\text{Anzahl der Paare mit verschiedenen Rangzahlen}\right)}{\text{Anzahl aller Paare}}$$

$$= \frac{\frac{1}{2}\binom{n}{2} - \text{Anzahl der Inversionen}}{\frac{1}{2}\binom{n}{2}},$$

wobei n die Anzahl aller Paare ist. Wir definieren den Verlust als "Anzahl der Inversionen".

Ohne Vorhersageverfahren kann man nur eine rein zufällige Reihenfolge der n möglichen Rangzahlen der y vorhersagen; daher ist der Erwartungswert des Verlusts die halbe Anzahl der möglichen Inversionen, d.h. $V(y) = \frac{1}{2}\binom{n}{2}$. Ist die Rangfolge der x-Werte bekannt, so wäre im Fall positiver Abhängigkeit die optimale Vorhersage dieselbe Rangfolge für die y-Werte, nämlich dann wenn die Anzahl der Inversionen kleiner als $\frac{1}{2}\binom{n}{2}$ ist, und genau die umgekehrte Reihenfolge, wenn die Abhängigkeit negativ ist. Diese Art der Vorhersage stellt sicher, daß die Anzahl der Inversionen bezüglich der tatsächlichen Rangfolge, also $V(y/x)$, den Wert $\binom{n}{2}$ nicht übersteigt. Das Maß Θ, das auf dieser Vorgehensweise beruht, ist gleich $|\tau|$ und kann als Anteil des speziellen Verlusts interpretiert werden, der bei der Vorhersage der Rangzahlen von y mit Hilfe der x-Werte eingespart wurde (siehe Tabelle 2).

Schlußfolgerung

Das Problem, das Schüler bewegt: "Was bedeutet r = 0,7" kann nun folgendermaßen beantwortet werden:

r^2 = 0,49 bedeutet, daß 49 % von dem Verlust (- definiert als die Summe der quadratischen Abweichungen -), den wir in Ermangelung eines geeigneten Vorhersageverfahrens in Kauf genommen hätten, durch optimale Vorhersage der y-Werte durch Einsetzen der x-Werte in die Gleichung der Regressionsgerade eingespart werden können.

Wir glauben, daß die vorgeschlagene Interpretation von r und anderer Maße, sinnvoller ist als nur von "einem Maß für den Grad der Abhängigkeit zwischen 0 und 1" zu sprechen. Darüber hinaus führt sie auch zu einer Vereinheitlichung der Interpretation verschiedener Maße. Für Schüler könnte es von Nutzen sein, einige der Abhängkeitsmaße, die sie in verschiedenem Zusammenhang kennengelernt haben, als Spezialfälle desselben Grundprinzips zu begreifen. Dieses Konzept der Maße für Vorhersagbarkeit kann im Unterricht auch dazu genützt werden, die verschiedenen Mittelwerte als die Werte aufzufassen, die gewisse Verlustfunktionen minimieren.

Fußnote

Es sei noch vermerkt, daß "Abhängigkeit" ein symmetrischer Begriff ist, während "Vorhersagbarkeit" nur eine Richtung beinhaltet: Schluß von x auf y oder umgekehrt von y auf x (die beiden "Vorhersagbarkeiten" sind nicht notwendig gleich).

Bemerkung

Wir danken Ozer Schield und Lea Shatil-Alon für die Anregung zu diesen Überlegungen. Die Vorarbeiten zu diesem Aufsatz wurden zum Teil durch das Sturman Center for Human Development von der Hebräischen Universität Jerusalem unterstützt.

Universität Tel-Aviv
Hebräische Universität von Jerusalem

Literatur

FALK, R. (1980). Minimize your losses. Teaching Statistics, 2, 80-83

HAYS, W.L. (1973). Statistics for the Social Sciences (2nd Edition). New York, Holt, Rinehart & Winston

SIEGEL, S. (1956). Non-Parametric Statistics for Behavioral Sciences. New York: McGraw-Hill