# Proportionalität ist „relativ“

ROLF MANTYK, RATINGEN

***Zusammenfassung:*** *Anhand der Lagebeziehungen von zwei Vektoren im n-dimensionalen Raum werden die Begriffe „Proportionalitätsmaß“ und „Modifikationsfaktor“ entwickelt. Fasst man die einzelnen Komponenten der beiden Vektoren paarweise zusammen, so lassen sich diese Objekte als zweidimensionale Punktwolke interpretieren. Eine geeignete Koordinatentransformation führt in diesem neuen Modell schließlich zur Definition des Korrelations- und Regressionskoeffizienten.*

## Vorwort

*Beschreibende Statistik* – insbesondere *Regression* und *Korrelation* – gehört laut den neuen Richtlinien für die Sekundarstufe II seit Jahren zum obligatorischen Themenkatalog der Jahrgangsstufe 11. Das mathematische Hinterfragen dieser Standardinstrumente der empirischen Wissenschaften erscheint zu diesem Zeitpunkt – beim Einstieg in die gymnasiale Oberstufe – mehr als sinnvoll, wird doch von nun an den Schülern in zunehmendem Maße eigenständiges Analysieren und Beurteilen von statistischen Daten in vielen Fächern abverlangt.

Allerdings hat die tägliche Unterrichtspraxis in den Schulen gezeigt, dass diesem Anspruch nicht immer genügend Raum geboten werden kann. Bei den Bemühungen um die *Grundlegung der Differentialrechnung* und eine vertiefende *Betrachtung der Koordinatengeometrie* muss gerade die *Statistik* aus Zeitgründen oft außen vor bleiben.

Die Reduktionen und Umverteilungen von Lerninhalten im Zusammenhang mit einer Verkürzung der Schulzeit bis zum Abitur um ein Jahr (G8-Modell) wird diese Situation zunehmend verschärfen.

Als mögliche Folge dieser „Trends” muss davon ausgegangen werden, dass die *Beschreibende Statistik* unter Umständen auf ein mehr oder weniger formelhaftes Anwenden zurückgedrängt wird.

Deshalb wird der „engagierte Statistiker” den im Folgenden vorgestellten Entwurf einer Einführung in die *Regression* und *Korrelation* im Rahmen der *Linearen Algebra* wahrscheinlich begrüßen, erhält er doch damit eine realisierbare[1] Möglichkeit, diese Instrumente – wenn auch im Nachhinein – für die Oberstufenschüler mathematisch einwandfrei zu begründen. Dabei entstand die Idee zu diesem „neuen Ansatz” interessanterweise bei dem Versuch, *Regression* und *Korrelation* für den Unterricht in der Jahrgangsstufe 11 (innerhalb des G9-Modells) zugänglicher zu machen.

In unserer Eigenschaft als Moderatoren für die Lehrerfortbildung wurden mein Kollege Michael Spielmann und ich von der Bezirksregierung Düsseldorf im Rahmen der Unterregionalisierten Lehrerfortbildung damit beauftragt, hier entsprechende Hilfestellungen für die Unterrichtspraxis zu entwickeln (vgl. Fußnote 1).

Nachdem wir unseren Entwurf für die Lehrerfortbildung formuliert hatten, stellte Michael Spielmann im Verlauf der folgenden Planungstagung im Februar 2004 ein Papier vor, in welchem er auf die Interpretationsmöglichkeiten der Begriffe *Regression* bzw. *Korrelation* als *Länge einer Projektion* bzw. als *Skalarprodukt* entsprechender *n*-dimensionaler Vektoren hinwies.

Inspiriert durch seine abschließenden Gedankengänge und motiviert durch die vielfältigen Anregungen, die er mir aufgrund seiner jahrelangen, intensiven Vertrautheit mit dem Thema bei unserer gemeinsamen Arbeit freundlicherweise zukommen ließ, hatte ich den Wunsch, eine Einführung der Begriffe *Regression* und *Korrelation* zu entwerfen, welche über die formalen Bezüge hinausgehend vollständig auf dem vektorgeometrischen Ansatz aufbaute.

Das Ergebnis meiner Arbeit erlaube ich mir hier zu präsentieren. Nachdrücklich darf ich mich an dieser Stelle nochmals für die vorbereitenden Inspirationen und Anregungen und das abschließende Lesen der Korrektur bei meinem lieben Kollegen Michael Spielmann bedanken.

## Die Problemstellung

Seit Januar 2002 haben wir „neues“ Geld in den Taschen. Die gute, alte DM musste endgültig dem Euro Platz machen. In einem logistisch beispiellosen Akt wurden allein in der Bundesrepublik 3 Mrd. Euro-Banknoten im Wert von 76,1 Mrd. € und 12 Mrd. Euro-Münzen im Wert von 4,4 Mrd. € in Umlauf gebracht. Obwohl bis Ende Februar mit der alten Währung bezahlt werden konnte, waren in

[1] In diesem Aufsatz geht es nicht vornehmlich um die Umsetzung dieses Entwurfs für die konkrete Unterrichtspraxis. Dies würde den Rahmen sprengen. Zunächst einmal soll vor allem die Idee skizziert werden.
Zur Zeit offeriert die Bezirksregierung Düsseldorf allerdings eine „konventionelle“ *Einführung in die Lineare Regression* in Form einer 2tägigen Veranstaltung. (Nähere Informationen unter www.mathe-treff.de)

Deutschland bis zum 9. Januar 2002 bereits 90 Prozent der DM aus dem Verkehr gezogen. Dem vorausgegangen war eine zweifache Ausweisung der Preise bei einem wesentlichen Anteil des Warensortiments seit Mitte 2001. Grundlage für diese Doppelauszeichnung war der Umrechnungsfaktor 1,95583. Jeder Euro-Preis musste mit diesem Faktor modifiziert werden, um den entsprechenden DM-Preis zu erhalten. Wer allerdings auch nach dem 28. Februar 2002 an den DM-Preisen interessiert war – und das waren wir alle, denn unser Preisgefühl konnte sich anfangs nur sehr schwerfällig mit der neuen Währung anfreunden – der musste selbst nachrechnen, weil ab diesem Zeitpunkt nur noch die Euro-Auszeichnung erlaubt war.

Während sich in unseren Köpfen schnell die „Verdopplung“ des Euro-Betrages als der „wahre Wert“ aller Waren etablierte, nutzte der Handel ab März 2002 die zunehmende Vergesslichkeit seiner Kunden den alten DM-Preisen gegenüber, um die neuen Euro-Preise verstärkt einer „eurospezifischen Dynamik zu überantworten“. In Tabelle 1 sind die Preise für 20 Artikel der Filiale einer Baumarktkette in Zusammenhang mit der Währungsumstellung notiert. Solange die zweifache Auszeichnung aktuell war, fand sich auf den Etiketten der DM-Preis und daneben der (theoretische) Euro-Preis, der vor dem 28. Februar 2002 dem realen Euro-Preis entsprach.

| Art. Nr. | durchschn. VK-Zahlen | alter Preis in DM | theoret. Preis in € | realer Preis in € | theoretische Brutto-einnahmen | reale Brutto-einnahmen | realer Faktor | bewertet realer Faktor |
|---|---|---|---|---|---|---|---|---|
| 1 | 200 | 4,98 | 2,55 | 2,89 | 510 | 578 | 1,72318 | 344,64 |
| 2 | 40 | 15,78 | 8,07 | 7,95 | 322,8 | 318 | 1,98491 | 79,4 |
| 3 | 17 | 138,50 | 70,81 | 69,89 | 1203,77 | 1188,13 | 1,98169 | 33,69 |
| 4 | 302 | 17,98 | 9,19 | 9,85 | 2775,38 | 2974,7 | 1,82538 | 551,26 |
| 5 | 1007 | 0,99 | 0,51 | 0,49 | 513,57 | 493,43 | 2,02041 | 2034,55 |
| 6 | 58 | 23,45 | 11,99 | 11,99 | 695,42 | 695,42 | 1,95580 | 113,44 |
| 7 | 92 | 45,78 | 23,41 | 23,49 | 2153,72 | 2161,08 | 1,94891 | 179,3 |
| 8 | 218 | 9,78 | 5,00 | 4,99 | 1090 | 1087,82 | 1,95992 | 427,26 |
| 9 | 24 | 19,95 | 10,20 | 9,98 | 244,8 | 239,52 | 1,99900 | 47,98 |
| 10 | 9 | 44,78 | 22,90 | 24,75 | 206,1 | 222,75 | 1,80929 | 16,28 |
| 11 | 35 | 64,25 | 32,85 | 32,85 | 1149,75 | 1149,75 | 1,95586 | 68,46 |
| 12 | 98 | 13,28 | 6,79 | 6,85 | 665,42 | 671,3 | 1,93869 | 189,99 |
| 13 | 186 | 7,75 | 3,96 | 3,99 | 736,56 | 742,14 | 1,94236 | 361,28 |
| 14 | 285 | 16,20 | 8,28 | 8,75 | 2359,8 | 2493,75 | 1,85143 | 527,66 |
| 15 | 18 | 38,27 | 19,57 | 19,98 | 352,26 | 359,64 | 1,91542 | 34,48 |
| 16 | 7 | 248,50 | 127,06 | 126,98 | 889,42 | 888,86 | 1,95700 | 13,7 |
| 17 | 19 | 145,90 | 74,60 | 74,98 | 1417,4 | 1424,62 | 1,94585 | 36,97 |
| 18 | 44 | 82,85 | 42,36 | 44,98 | 1863,84 | 1979,12 | 1,84193 | 81,04 |
| 19 | 37 | 18,47 | 9,44 | 9,98 | 349,28 | 369,26 | 1,85070 | 68,48 |
| 20 | 4 | 324,75 | 166,04 | 167,75 | 664,16 | 671 | 1,93592 | 7,74 |
| | | | | | | | | 5217,6 |
| | 2700 Summe | 1,95583 Faktor offiziell | | | 20163,45 | 20708,29 | 1,91718 Mittelwert | 1,93244 Mittelwert bewertet |
| | | | | | Saldo 544,84 | | | |

Tabelle 1

Mit dem Verschwinden des DM-Bezuges konnte man jedoch allenthalben eine „Modifizierung der Euro-Preise im Rahmen der gängigen Preispolitik“ beobachten. So wurde z.B. für den Artikel Nr. 9 der theoretische Preis gesenkt, um die magische 10€-Marke nicht überschreiten zu müssen. Artikel Nr. 1 wurde aufgrund seiner hohen Verkaufszahlen angehoben. Insgesamt ergibt sich für den der Tabelle 1 zugrunde liegenden Zeitraum – was natürlich kaufmännisch betrachtet sehr zufriedenstellend ist – ein Umsatzplus von mehr als 500 €, wenn man die theoretischen und die realen Bruttoeinnahmen miteinander vergleicht. Entsprechend sind die Umrechnungsfaktoren ohne die Bewertung bzgl. der Umsatzzahlen im Schnitt fast 4 Hundertstel kleiner als der offizielle Umrechnungsfaktor. Bewertet liegen sie durchschnittlich noch 2 Hundertstel unter dem offiziellen Wert.[2]

Und damit sind die grundsätzlichen Probleme, die im Folgenden angesprochen werden sollen, auch bereits im Wesentlichen vollständig umrissen.

[2] Um von DM auf Euro zu kommen, muss man durch den Umrechnungsfaktor dividieren. Ein kleinerer Faktor bedingt also einen Euro-Wert, der über dem offiziellen Wert liegt.

- Lässt sich bzgl. der Umrechnung von den realen Euro-Preisen zu den ursprünglichen DM-Preisen ein „mittlerer Modifikationsfaktor“ festlegen?
- Lassen sich die durchaus individuellen Preisangleichungen verschiedener Filialen der besagten Baumarktkette miteinander vergleichen?

## Lineare Beziehungen zwischen Datenmengen

Wir haben eine Liste von $n$ Zahlen, die wir *Ursprung*, kurz $\vec{u}$ nennen werden. Eine zweite Liste mit derselben Anzahl von Zahlen bezeichnen wir als *Ziel*, kurz $\vec{z}$ genannt. Wir interessieren uns für die Lage von $\vec{u}$ und $\vec{z}$ zueinander, wobei wir sinnvollerweise davon ausgehen, dass weder $\vec{u}$ noch $\vec{z}$ gleich $\vec{0}$ sind. Präziser geht es damit um folgende Fragen (vgl. Abbildung 1):

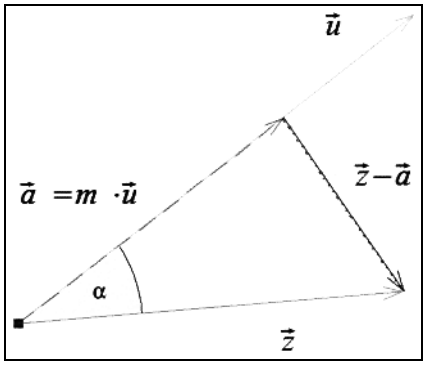

Abbildung 1

- Wir schauen uns die Projektion $\vec{a}$ von $\vec{z}$ auf $\vec{u}$ bzw. $m \cdot \vec{u}$ an. Welches $m$ beschreibt diese Projektion? Oder, mit welchem $m$ müssen wir $\vec{u}$ *modifizieren*[3], damit der Abstand[4] zwischen dieser *Annäherung* $\vec{a} = m \cdot \vec{u}$ von $\vec{u}$ an $\vec{z}$ minimal wird?
- Lässt sich darüber hinaus ein geeignetes Maß für die „Abweichung“ zwischen $m \cdot \vec{u}$ und $\vec{z}$ finden, welches unabhängig von $m$ ist?

Im trivialen Fall, dass sämtliche Zahlen aus *Ziel* aus den Zahlen von *Ursprung* durch eine einfache Multiplikation mit demselben Faktor $m \neq 0$ hervorgegangen sind, haben wir eine sog. „strenge“ Proportionalität vorliegen, sodass $m \cdot u_i = z_i$ für alle $i = 1,...,n$ ist. $\vec{u}$ und $\vec{z}$ liegen also auf derselben Geraden. $\vec{a}$ und $\vec{z}$ sind identisch. Ihr Abstand beträgt Null, was sich vektoriell folgendermaßen notieren lässt:

$$m \cdot \begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix} = \begin{pmatrix} m \cdot u_1 \\ \vdots \\ m \cdot u_n \end{pmatrix} = \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} z_1 \\ \vdots \\ z_n \end{pmatrix}$$

$$\Leftrightarrow m \cdot \vec{u} = \vec{a} = \vec{z}\,.$$

Falls $m$ positiv ist, sprechen wir bzgl. der Daten aus den beiden Listen von einer *100%ig positiven Proportionalität* bzw. sagen, dass das sog. *Proportionalitätsmaß* $\boldsymbol{r}$ gleich 1 sei, $r = 1$.

Ist $m$ negativ, gehen wir von einer *100%ig negativen Proportionalität* aus bzw. sagen, dass das sog. *Proportionalitätsmaß* $\boldsymbol{r}$ den Wert –1 annehme.

Geometrisch betrachtet haben $\vec{u}$ und $\vec{z}$ im $n$-dimensionalen Raum in beiden Fällen dieselbe Richtung. Während ihre Orientierung bei $r = 1$ identisch ist, verläuft sie im Fall $r = -1$ jedoch entgegengesetzt. Der Winkel zwischen $\vec{u}$ und $\vec{z}$ hat das Winkelmaß 0 bzw. π. Stellen wir uns vor, diese „Grenz“-Zustände seien aus einer „Drehung“ von $\vec{u}$ und $\vec{z}$ zueinander hervorgegangen. Dann existieren aber auch „Zwischen“-Stadien. Wenn z. B. das Winkelmaß zwischen $\vec{u}$ und $\vec{z}$ den Wert π/2 annimmt, liegt es nahe, das *Proportionalitätsmaß r* mit 0 zu bewerten. In diesem Fall stehen die beiden Vektoren $\vec{u}$ und $\vec{z}$ senkrecht zueinander.

Als ein Abweichungsmaß zwischen *Ursprung* bzw. *Annäherung* und *Ziel,* welches unabhängig von $m$ ist, wäre demnach die *Cosinus*-Funktion geeignet, welche den Winkel α zwischen $\vec{a}$ bzw. $\vec{u}$ und $\vec{z}$ beschreibt und den bisher genannten Zuständen sehr wohl genügen kann:

$$\cos:[0;\pi] \rightarrow [-1;1].$$

Ausgehend von $\vec{u}$ und $\vec{z}$ bestimmen wir $r$ also mit $\cos\alpha$ bzw. mit Hilfe des Skalarproduktes[5] (vgl. Abbildung 2),

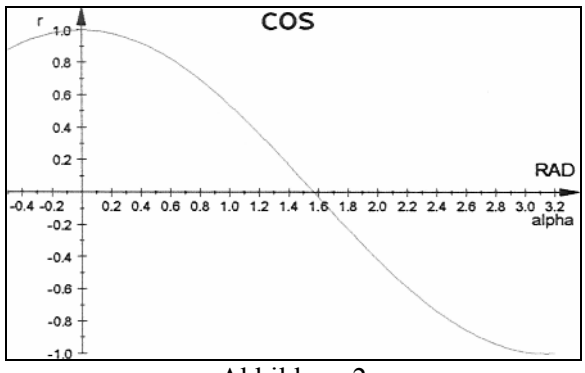

Abbildung 2

[3] Wir gehen davon aus, dass hierbei $\vec{u}$ und $\vec{z}$ nicht senkrecht zueinander stehen. Andernfalls setzen wir $m = 0$.

[4] Die Definition für Länge und Abstand im $n$-dimensionalen Raum folgt weiter unten.

[5] Falls auf die $n$-dimensionale Situation besonders verwiesen werden soll, bezeichnen wir die Länge eines Vektors auch mit $\|\vec{u}\| = \sqrt{\vec{u} \cdot \vec{u}}$ .

Dabei gilt: $\vec{u} \cdot \vec{z} = \sum_{i=1}^{n} u_i \cdot z_i$ .

d. h.

$$r: = cos\ \alpha = \frac{\vec{u} \cdot \vec{z}}{|\vec{u}| \cdot |\vec{z}|} .$$

Falls $\vec{a} \neq \vec{z}$, d. h. der Abstand zwischen $\vec{a}$ und $\vec{z}$ größer als Null ist, die Werte von α somit zwischen 0 und π, $0 < \alpha < \pi$, liegen, bedarf es eines deutlich höheren Aufwandes, um das optimale *m* zu ermitteln[6].

Allgemein ist der Abstand zwischen $\vec{a}$ und $\vec{z}$ bzw. die Länge $|\vec{z} - \vec{a}|$ definiert durch:

$$|\vec{z} - \vec{a}| = \sqrt{(\vec{z} - \vec{a})^2} = \sqrt{\sum_{i=1}^{n} (z_i - a_i)^2}$$

bzw.

$$|\vec{z} - m \cdot \vec{u}| = \sqrt{(\vec{z} - m \cdot \vec{u})^2} = \sqrt{\sum_{i=1}^{n} (z_i - m \cdot u_i)^2} .$$

Wir suchen also ein *m*, womit die Strecke $|\vec{z} - \vec{a}|$ minimiert werden kann. Hierzu betrachten wir den Ausdruck als eine Funktion von *m,* wobei es genügt, die Summe der quadratischen Abweichungen,

$$f(m) = \sum_{i=1}^{n} (z_i - m \cdot u_i)^2 ,$$

auf ihr Minimum hin zu untersuchen. Es gilt:

$$f(m) = \sum_{i=1}^{n} (z_i^2 - 2 z_i m u_i + m^2 u_i^2) .$$

Wir bilden die erste Ableitung

$$f'(m) = \sum_{i=1}^{n} (-2 z_i u_i + 2 m u_i^2) .$$

Hiervon suchen wir die Nullstelle.

$f'(m) = 0$ liefert $m = \dfrac{\sum_{i=1}^{n} 2 z_i u_i}{\sum_{i=1}^{n} 2 u_i^2}$.

Also: $m_{opt} = \dfrac{\sum_{i=1}^{n} z_i u_i}{\sum_{i=1}^{n} u_i^2}$ bzw. $m_{opt} := \dfrac{\vec{u} \cdot \vec{z}}{\vec{u} \cdot \vec{u}}$.

Speziell ergibt sich für $m \cdot \vec{u} = \vec{z}$, wenn also $\vec{u}$ und $\vec{z}$ auf derselben Gerade liegen:

$$m_{opt} = \frac{\vec{z} \cdot \vec{u}}{\vec{u} \cdot \vec{u}} = \frac{\vec{u} \cdot (m \cdot \vec{u})}{\vec{u} \cdot \vec{u}} = m \cdot \frac{\vec{u} \cdot \vec{u}}{\vec{u} \cdot \vec{u}} = m.$$

Falls $\vec{u} \perp \vec{z}$ ist, greift die Formel für das optimale *m* ebenfalls:

$$m_{opt} = \frac{\vec{z} \cdot \vec{u}}{\vec{u} \cdot \vec{u}} = \frac{0}{\vec{u} \cdot \vec{u}} = 0.$$

[6] Bis auf den Fall $\vec{a} = \vec{0}$, wo $\vec{u}$ und $\vec{z}$ senkrecht zueinander stehen, also $m = 0$ ist.

Fassen wir die bisherigen Ergebnisse zusammen:
Wir können die Proportionalität zwischen den Daten aus den beiden Listen $\vec{u} \neq \vec{0}$ und $\vec{z} \neq \vec{0}$ durch das sog. *Proportionalitätsmaß* beschreiben

$$r = \frac{\vec{u} \cdot \vec{z}}{|\vec{u}| \cdot |\vec{z}|} .$$

Wir sind ferner in der Lage, mit Hilfe eines *Modifikationsfaktors m* eine *Annäherung* $\vec{a}$ für den *Ursprung* $\vec{u}$ zu bestimmen, sodass der Abstand zwischen $m \cdot \vec{u} = \vec{a}$ und $\vec{z}$ minimal wird.

Während das *Proportionalitätsmaß r* gegebene (unveränderte) Daten miteinander vergleicht, versucht der *Modifikationsfaktor m* durch eine Modifizierung der *Ursprung*sdaten die Beziehung zu den *Ziel*daten zu „optimieren".

In den „Grenz"-Fällen, in denen das *Proportionalitätsmaß* den Wert $|r| = 1$ annimmt, sprechen wir von strenger Proportionalität zwischen den Daten des *Ursprungs* und den *Ziel*daten. Der *Modifikationsfaktor m* ergibt sich hierbei direkt aus der Beziehung

$$m = \frac{z_i}{u_i} \text{ bzw. } z_i = m \cdot u_i \text{ für } i = 1,..., n.$$

Ein *Proportionalitätsmaß* von $0 < |r| < 1$ signalisiert hingegen, dass sich die Proportionalität „relativiert" hat. Die Berechnung des *Modifikationsfaktors m* wird damit aufwändiger.

In jedem Fall ergibt sich das optimale *m* über

$$m = \frac{\vec{u} \cdot \vec{z}}{\vec{u} \cdot \vec{u}} .$$

## Veranschaulichungen im zweidimensionalen Raum

Wir betrachten die einzelnen Komponenten von $\vec{u} \neq \vec{0}$ und $\vec{z} \neq \vec{0}$,

$$\vec{u} = \begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix} \text{ und } \vec{z} = \begin{pmatrix} z_1 \\ \vdots \\ z_n \end{pmatrix}$$

und fassen sie „horizontal" zu Paaren

$$\begin{pmatrix} u_1 \\ z_1 \end{pmatrix}, \begin{pmatrix} u_2 \\ z_2 \end{pmatrix}, \ldots, \begin{pmatrix} u_n \\ z_n \end{pmatrix}$$

zusammen, die sich in einem zweidimensionalen Koordinatensystem als Graphenpunkte verdeutlichen lassen, wobei eine sog. *Punktwolke* entsteht.

Obwohl wir damit für unsere *n*-dimensionalen Daten das anschauliche Modell ausgetauscht haben, bleiben wir trotzdem weiterhin unserem zuvor auf *n*-dimensionalem Terrain definierten Abstandsbegriff treu.

Damit positionieren sich in diesem zweidimensionalen Rahmen die *Annäherungs*punkte $(u_i|a_i)$ dadurch, dass die Summe ihrer quadrierten vertikalen Abstände (Strecken) zu den jeweiligen $(u_i|z_i)$ minimal wird. Somit bleibt der „Abstand" in der Punktwolke weiterhin ein globales Phänomen, welches sich nicht zwingend über Abstandsbetrachtungen im zweidimensionalen Kontext lokal erschließen lässt. Aus der zweidimensionalen Perspektive heraus müsste man vielmehr in Anlehnung an alle bis dahin gewonnene Erfahrung vom Lot der $(u_i|z_i)$ auf die *Annäherungs*gerade[7] ausgehen. Analog zu den Überlegungen im *n*-dimensionalen Raum lassen sich 3 Fälle unterscheiden[8]:

Für das *Proportionalitätsmaß* gilt:

$\boldsymbol{r = 1}$ (vgl. Abbildung 3), bzw.
$\boldsymbol{r = -1}$ (vgl. Abbildung 4), bzw.
$\boldsymbol{0 < |r| < 1}$ (vgl. Abbildung 5).

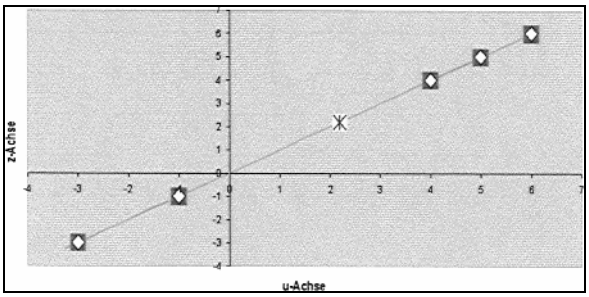

Abbildung 3

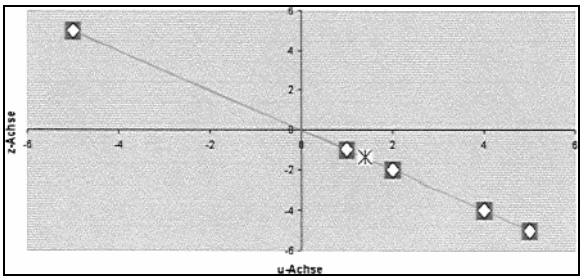

Abbildung 4

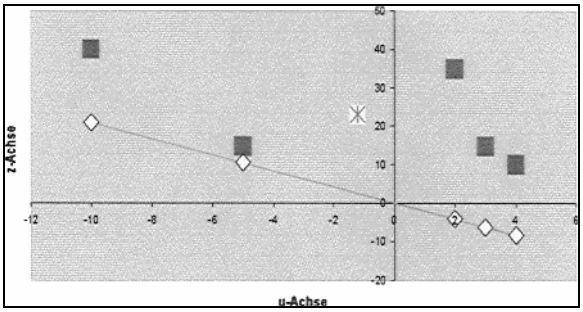

Abbildung 5

[7] Definition der *Annäherungs*geraden siehe weiter unten.
[8] Die Punkte $(u_i|z_i)$ sind als Quadrate und die *Annäherungs*punkte $(u_i|a_i)$ sind als Rauten dargestellt. Der „Schwerpunkt" der Punktwolke wird durch ∗ gekennzeichnet. Die Definition des Schwerpunktes erfolgt weiter unten.

Sämtliche *Annäherungs*punkte $(u_i|a_i)$ liegen aufgrund ihrer Definition auf einer Geraden, im Folgenden *Annäherungs*gerade genannt, die durch den Punkt (0|0) des *zu-Koordinatensystems* verläuft. Ihre Steigung wird durch den *Modifikationsfaktor m* beschrieben.

Der „Schwerpunkt" der Punktwolke $S(\overline{u}\,|\,\overline{z}\,)$ wird durch die Werte $\overline{u}$ bzw. $\overline{z}$ definiert, die jeweils das arithmetische Mittel der *Ursprungs-* bzw. *Ziel*daten darstellen.

$$\overline{u} := \frac{1}{n}\sum_{i=1}^{n} u_i \text{ bzw. } \overline{z} := \frac{1}{n}\sum_{i=1}^{n} z_i .$$

Offensichtlich liegt dieser Schwerpunkt $S(\overline{u}\,|\,\overline{z}\,)$ im Fall $|r| = 1$ stets auf der *Annäherungs*geraden (vgl. Abbildung 3 und Abbildung 4).

Im Fall $0 < |r| < 1$ findet sich seine Position jedoch i. A. nicht auf der besagten Geraden, was natürlich eher unbefriedigend ist, da doch insbesondere der Schwerpunkt bzw. das arithmetische Mittel gegebener numerischer Daten als deren „erste, repräsentative Annäherung" betrachtet wird (vgl. Abbildung 5).

Untersuchen wir stattdessen nur *Ursprungs-* und *Zieldaten*, welche die Eigenschaft

$$\sum_{i=1}^{n} u_i = 0 \text{ bzw. } \sum_{i=1}^{n} z_i = 0$$

haben, so liegt der Schwerpunkt der Punktwolke $(u_i|z_i)$ mit $i = 1,..,n$ bei (0|0) und somit in jedem Fall – wie gewünscht – auf der *Annäherungs*geraden.

Die Frage ist nun, wie wir mit Punktwolken verfahren sollen, deren Schwerpunkt sich nicht im Ursprung des gegebenen Koordinatensystems befindet?

## Korrelation und Regression

Unser Augenmerk richten wir nun auf solche speziellen Listen *Ursprung* $\vec{u} \neq \vec{0}$ und *Ziel* $\vec{z} \neq \vec{0}$ mit

$$\sum_{i=1}^{n} u_i = 0 \text{ bzw. } \sum_{i=1}^{n} z_i = 0,$$

wobei jeweils
$u_i := x_i - \overline{x}$ bzw. $z_i := y_i - \overline{y}$ für $i = 1,..., n$ gilt, was wiederum bedeuten soll, dass wir von zwei beliebigen[9] Datenlisten $\vec{x}$ und $\vec{y}$ ausgehend die oben notierte Bedingung für den Schwerpunkt $S(\overline{u}\,|\,\overline{z}\,)$ im zugehörigen *zu-Koordinatensystem* erfüllen.

[9] Die Beliebigkeit wird dadurch eingeschränkt, dass nicht sämtliche $x_i$ bzw. $y_i$ jeweils identisch sein dürfen.

In dem entsprechenden *xy-Koordinatensystem* stellen die Punkte $(x_i|y_i)$ eine Punktwolke dar, deren Schwerpunkt bei $S(\overline{x}\,|\,\overline{y})$ liegt. Geometrisch ergeben sich die $u_i$ aus den $x_i$ bzw. die $z_i$ aus den $y_i$ durch eine Verschiebung des *xy-Koordinatensystems* in den Schwerpunkt $S(\overline{x}\,|\,\overline{y})$.

$u_i$ bzw. $z_i$ sind die „neuen" Koordinaten der $x_i$ bzw. $y_i$ in dem neuen *zu-Koordinatensystem.*

Unter diesen Voraussetzungen nennen wir den *Modifikationsfaktor m*, der die Steigung der Geraden liefert, welche durch den „Abstand der *Annäherungs*punkte $(u_i|a_i)$ von den transformierten Punkten $(u_i|z_i)$", kurz durch den Abstand[10] von $\vec{a} = m \cdot \vec{u}$ und $\vec{z}$, definiert wird,

*Regressionskoeffizient* der *xy-Daten*[11],

und die zugehörige *Annäherungs*gerade, die durch die Punkte $(u_i|a_i)$ festgelegt wird, heißt

*Regressionsgerade* der *xy-Daten*.

Schließlich bezeichnen wir das *Proportionalitätsmaß r,* d. h. den *cos α* des eingeschlossenen Winkels zwischen den beiden Vektoren $\vec{u}$ und $\vec{z}$ in diesem Kontext als die

*Korrelation r* der *xy-Daten*[12].

Allgemein lässt sich der *Regressionskoeffizient* über den Ansatz

$$\frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{(\vec{x} - \overline{x}) \cdot (\vec{y} - \overline{y})}{(\vec{x} - \overline{x}) \cdot (\vec{x} - \overline{x})} = \frac{\vec{u} \cdot \vec{z}}{\vec{u} \cdot \vec{u}} =: m$$

bestimmen.

Die zugehörige *Regressionsgerade* verläuft stets durch den Nullpunkt des *uz-Koordinatensystems*, der durch den Schwerpunkt $S(\overline{x}\,|\,\overline{y})$ der *xy-Daten* definiert wird.

Die *Regressionsgerade* wird im *uz-Koordinatensystem* durch $z = m \cdot u$ und im *xy-Koordinatensystem* durch $y = m(x - \overline{x}) + \overline{y}$ beschrieben (vgl. Abbildungen 6, 7 und 8 folgende Seite).

Wir hatten oben im *n*-dimensionalen Kontext den *Regressionskoeffizienten m* so festgelegt, dass der Abstand zwischen den *Annäherungs*daten $\vec{a}$ und den *Ziel*daten $\vec{z}$ minimal wurde.

In der zweidimensionalen Perspektive realisiert sich diese Situation als Minimierungsproblem über die Summe der Quadrate aus den vertikalen Abständen, welche die gegebenen Punkte $(x_i|y_i)$ bzw. $(u_i|z_i)$ zu den *Annäherungspunkten* $(x_i|a_i)$ bzw. $(u_i|a_i)$ haben.

Dabei wird $a_i$ im *xy-System* durch

$$a_i = m(x_i - \overline{x}) + \overline{y}$$

und im *uz-System* durch

$$a_i = mu_i$$

repräsentiert. Denn es gilt:

$$|z_i - a_i| = |z_i - m \cdot u_i|$$

(Darstellung im *zu-System*)

$$= |(y_i - \overline{y}) - m \cdot (x_i - \overline{x})|$$

(Koordinatentransformation)

$$= |y_i - [m \cdot (x_i - \overline{x}) + \overline{y}]|$$

(Umstellen)

$$= |y_i - a_i|$$

(Darstellung im *xy-System*).

Genauer betrachtet, muss die Summe der quadrierten vertikalen Abstände

$$\sum_{i=1}^{n}(z_i - a_i)^2$$

minimal werden (vgl. vertikal verlaufende Strecken in Abbildung 8).

Bleibt noch die Frage nach der Berechnung des *Korrelationskoeffizienten* beliebiger Datenlisten.

Wir definieren den *Korrelationskoeffizienten* von zwei *n*-dimensionalen Datenlisten $\vec{x} \neq \vec{0}$ und $\vec{y} \neq \vec{0}$ als:

$$\frac{\sum_{i=1}^{n}(x_i - \overline{x}) \cdot (y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

$$= \frac{(\vec{x} - \overline{x}) \cdot (\vec{y} - \overline{y})}{|(\vec{x} - \overline{x})| \cdot |(\vec{y} - \overline{y})|}$$

$$= \frac{\vec{u} \cdot \vec{z}}{|\vec{u}| \cdot |\vec{z}|}$$

$$:= r.$$

## Wie war das mit dem Euro?

Inzwischen mussten wir uns alle an die neue Währung gewöhnen, was die Leute in bestimmten Situationen natürlich nicht davon abhalten kann, nachdrücklich auf den DM-Wert einer Ware hinzuweisen, falls man ihren hohen Preis ganz besonders hervorheben möchte.

[10] Gemeint ist $\|\vec{z} - \vec{a}\| = \|\vec{z} - m \cdot \vec{u}\|$.

[11] regredere (lat.) zurückführen, zurückgehen

[12] oder als *Korrelationskoeffizienten r* der *xy-Daten*

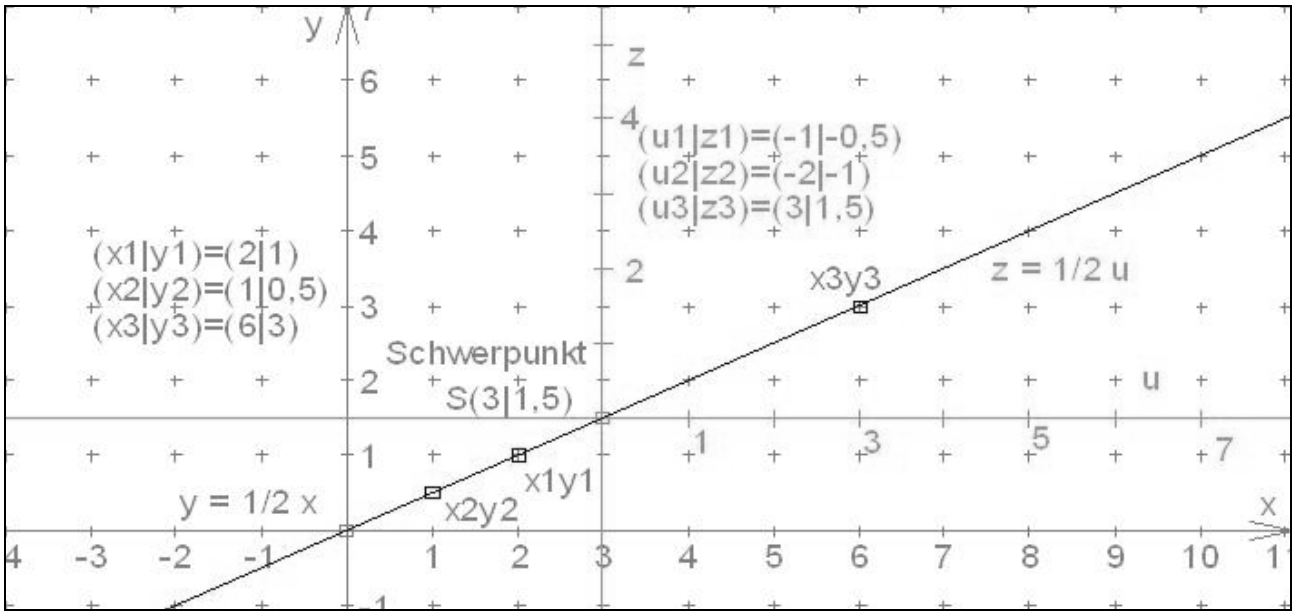

Abbildung 6

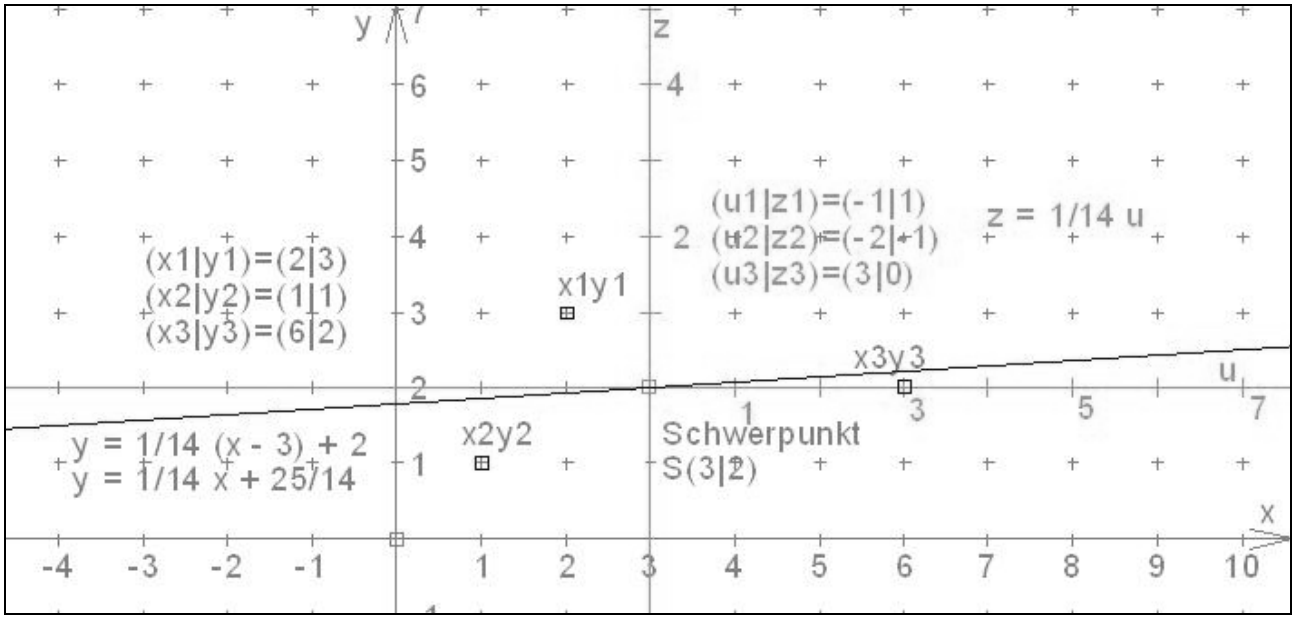

Abbildung 7

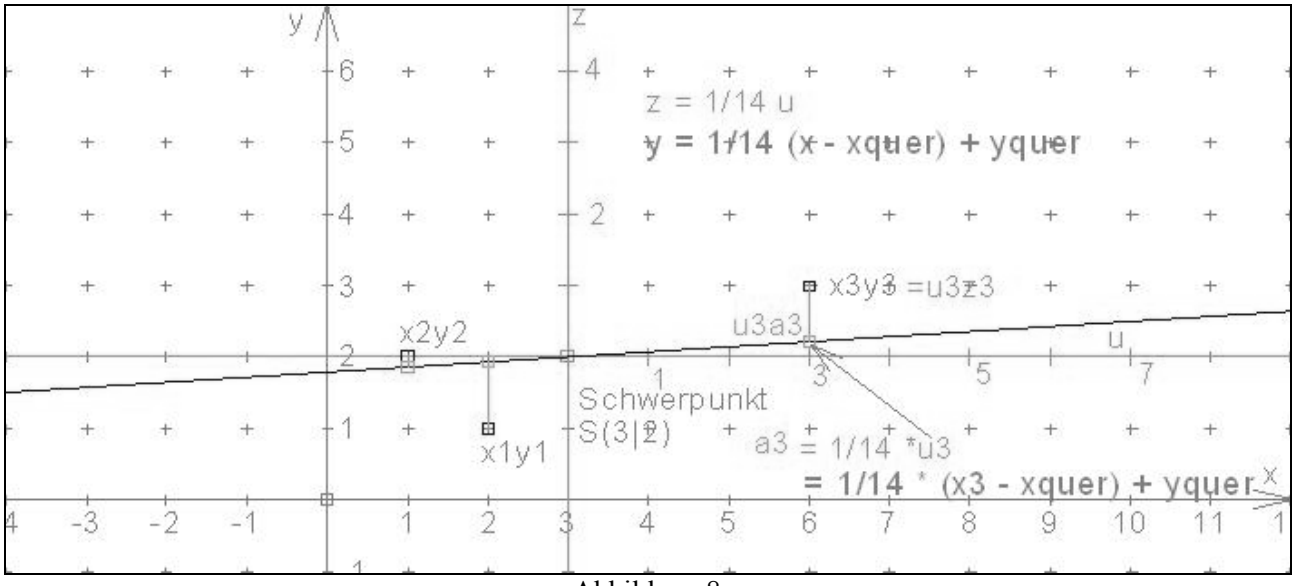

Abbildung 8

Wir nehmen Bezug auf die Tabelle 1 bzw. auf die Tabelle 2, wo in der *x*-Spalte die Warenpreise in DM und in der *y*-Spalte die realen Warenpreise in Euro notiert sind.

| ***x*** | ***y*** | ***u*** | ***z*** | $u^2$ | $z^2$ | ***u·z*** |
|---|---|---|---|---|---|---|
| 4,98 | 2,89 | -59,13 | -30,28 | 3496,36 | 916,88 | 1790,46 |
| 15,78 | 7,95 | -48,33 | -25,22 | 2335,79 | 636,05 | 1218,88 |
| 138,5 | 69,89 | 74,39 | 36,72 | 5533,87 | 1348,36 | 2731,6 |
| 17,98 | 9,85 | -46,13 | -23,32 | 2127,98 | 543,82 | 1075,75 |
| 0,99 | 0,49 | -63,12 | -32,68 | 3984,13 | 1067,98 | 2062,76 |
| 23,45 | 11,99 | -40,66 | -21,18 | 1653,24 | 448,59 | 861,18 |
| 45,78 | 23,49 | -18,33 | -9,68 | 335,99 | 93,7 | 177,43 |
| 9,78 | 4,99 | -54,33 | -28,18 | 2951,75 | 794,11 | 1531,02 |
| 19,95 | 9,98 | -44,16 | -23,19 | 1950,11 | 537,78 | 1024,07 |
| 44,78 | 24,75 | -19,33 | -8,42 | 373,65 | 70,9 | 162,76 |
| 64,25 | 32,85 | 0,14 | -0,32 | 0,02 | 0,1 | -0,04 |
| 13,28 | 6,85 | -50,83 | -26,32 | 2583,69 | 692,74 | 1337,85 |
| 7,75 | 3,99 | -56,36 | -29,18 | 3176,45 | 851,47 | 1644,58 |
| 16,2 | 8,75 | -47,91 | -24,42 | 2295,37 | 596,34 | 1169,96 |
| 38,27 | 19,98 | -25,84 | -13,19 | 667,71 | 173,98 | 340,83 |
| 248,5 | 126,98 | 184,39 | 93,81 | 33999,67 | 8800,32 | 17297,63 |
| 145,9 | 74,98 | 81,79 | 41,81 | 6689,6 | 1748,08 | 3419,64 |
| 82,85 | 44,98 | 18,74 | 11,81 | 351,19 | 139,48 | 221,32 |
| 18,47 | 9,98 | -45,64 | -23,19 | 2083,01 | 537,78 | 1058,39 |
| 324,75 | 167,75 | 260,64 | 134,58 | 67933,21 | 18111,78 | 35076,93 |
| | | | | 144522,79 | 38110,24 | 74203 |
| 64,11 | 33,17 | 0 | 0 | 380,16 | 195,22 | |
| *x*-quer | *y*-quer | | | Länge *u* | Länge *z* | |
| | Korrelation: | 0,99984 | | | | |
| | Regression: | 0,51343 | 1,94769 | Umrechnungsfaktor | | |

Tabelle 2

Wie nicht anders erwartet, korrelieren die DM- und €-Preise in hohem Maße. Die Regression liefert hingegen die Steigung für die Annährungsgerade. Da wir von den DM-Preisen als *Ursprung* ausgehen, ergibt sich als Regressionskoeffizient folgerichtig der Wert 0,51343. Der Kehrwert 1,94769 stellt den „mittleren Umrechnungsfaktor“ von € nach DM dar.

## Abschließende Bemerkungen

Bei der hier vorliegenden Entwicklung eines Korrelations- bzw. Regressionsmodells für Punktwolken in der Ebene sind die *n*-dimensionalen Daten aus den beiden Listen *Ursprung,* $\vec{u} \neq \vec{0}$, und *Ziel*, $\vec{z} \neq \vec{0}$, der Ausgangspunkt unserer Betrachtungen.

Solange $\vec{u}$ und $\vec{z}$ als *n*-dimensionale Vektoren auf derselben Geraden liegen, definieren die Begriffe *Proportionalitätsmaß* und *Modifikationsfaktor* eingängige mathematische Objekte, deren Sinn sich direkt aus dem gegebenen (*n*-dimensionalen) Kontext ableiten lässt.

Sobald jedoch diese lineare Abhängigkeit von $\vec{u}$ und $\vec{z}$ aufgehoben wird, liefert die zweidimensionale Veranschaulichung der paarweise zusammengehörenden Komponenten von $\vec{u}$ und $\vec{z}$ einen quasi disharmonischen Zustand, da der Schwerpunkt der zweidimensionalen Punktwolke nun nicht mehr auf der *Annäherungs*geraden liegt. Um dies generell zu gewährleisten, werden einschneidende Restriktionen notwendig.

Weil sämtliche Annäherungsgeraden als Ursprungsgeraden nur einen gemeinsamen Punkt – den Ursprung des Koordinatensystems – haben, ergeht an alle Punktwolken die Forderung, eben dort ihren Schwerpunkt zu platzieren, damit dieser stets auf der *Annäherungs*geraden liegen muss.

Wie lassen sich aber die Punktwolken beliebiger Datenmengen in dieses Korsett zwängen?

*Transformation* heißt das Schlüsselwort, welches an dieser Stelle klare Verhältnisse schafft und gleichzeitig in Bezug auf die *xy-Daten* neue Begrifflichkeiten erfordert.

Damit die Forderung nach $\vec{u} \neq \vec{0}$ und $\vec{z} \neq \vec{0}$ weiterhin erfüllt werden kann, dürfen außerdem in $\vec{x} \neq \vec{0}$ und $\vec{y} \neq \vec{0}$ nicht jeweils sämtliche Komponenten identisch sein.

Bleibt abschließend zu erwähnen, dass dem primären Anliegen, *Lineare Trends in zweidimensionalen Punktwolken zu quantifizieren*, über diesen Ansatz

im *n-dimensionalen Kontext somit* insgesamt mit einer in sich konsistenten Argumentationskette Genüge getan werden kann.

Bis auf das sehr gut begründbare „Zugeständnis“, dass der Schwerpunkt der Punktwolken stets auf deren Annäherungsgeraden liegen sollte, wird der gesamte Begriffsrahmen für die *Regression* bzw. *Korrelation* von Daten ansonsten komplett auf die Lagebeziehungen der beiden *n*-dimensionalen Vektoren *Ursprung* und *Ziel* reduziert.

Insbesondere ist der explizite Zugriff auf die Standardabweichung nicht notwendig. Auch die mitunter schwer nachvollziehbare Bevorzugung der vertikalen Abstandsquadrate im zweidimensionalen Kontext steht auf der *n*-dimensionalen Schiene an keiner Stelle zur Diskussion.

Vielmehr kann bereits im Vorfeld erkannt werden, dass lokale Abstandsbetrachtungen bzgl. einzelner Punkte zu einer Geraden (Lot auf die Gerade fällen!) sehr wohl vom globalen Abstandsproblem aller Punkte zur gesuchten Geraden zu unterscheiden sind.

Dem gegenüber steht der Aufwand, die erforderlichen Instrumente aus der Linearen Algebra zur Verfügung zu stellen, hier insbesondere das Skalarprodukt und damit verbunden der Abstandsbegriff und natürlich die Messung von Winkeln im *n*-dimensionalen Raum.

### Literatur


Baum, M., Riemer, W., Schermuly, H., Stark, J., Weidig, I., Zimmermann, P.: Lambacher Schweizer 11, Mathematisches Unterrichtswerk für das Gymnasium, Ausgabe NRW, Ernst Klett Verlag, Stuttgart 2000

Biehler, R.: Explorative Datenanalyse – Eine Untersuchung aus der Perspektive einer deskiptiv-empirischen Wissenschaftstheorie. IDM Materialien und Studien Band 24, Bielefeld: Universität Bielefeld 1982

Eipelt, B., Hartung, J.: Grundkurs Statistik, Lehr- und Übungsbuch der angewandten Statistik, Oldenbourg Verlag, München Wien 1987

Fischer, G.: Stochastik einmal anders. Friedr. Vieweg & Sohn Verlag, Wiesbaden 2005

Vogel, F.; Beschreibende und schließende Statistik. Oldenbourg Verlag, München Wien 1997

Wirths, H.: Beziehungshaltige Mathematik in Regression und Korrelation. Stochastik in der Schule (1, 1991)


### Autor


Rolf Mantyk
Carl Friedrich von Weizsäcker-Gymnasium
Karl-Mücher-Weg 2
40878 Ratingen
Email: rolman@cultpro.de


---

# Presse-ΣΠΛITTEP (2)

24. Oktober 2004

Frankfurter Allgemeine Sonntagszeitung.

**Geschafft:** Weltmeister im „Schere, Stein, Papier“ darf sich nun schon seit acht Jahren Lee Rammage aus dem kanadischen Burlington nennen. Am Samstag vor einer Woche schlug Rammage bei den mehr oder weniger offiziellen Weltmeisterschaften in der bislang eher unbekannten Wettkampfdisziplin (F.S.A. vom 17. Oktober) Heather Birrell in deren Heimatstadt Toronto; Chris Berggeren aus dem amerikanischen Bundesstaat Michigan erhielt die Bronzemedaille. Der Wettbewerb hatte Spieler aus ganz Kanada, aus Australien, aus der Tschechischen Republik, aus Norwegen und 21 amerikanischen Staaten angelockt. Zum Nachspielen folgt hier noch einmal der dramatische dritte sowie der fünfte und letzte Satz; die Figuren des späteren Siegers zuerst (zur Erinnerung: Schere schlägt Papier, Stein schlägt Schere, Papier schlägt Stein). Dritter Satz: Stein – Stein; Papier – Schere; Stein – Stein; Papier – Papier; Schere – Schere; Papier – Papier; Stein – Stein; Papier – Schere (Satz an Birrell). Fünfter Satz, zu diesem Zeitpunkt hatte jeder Finalist je zwei Sätze gewonnen: Papier – Stein; Stein – Schere (Satz und Sieg an Rammage).