Daten, die auf der Erde liegen - auf Spurensuche im Supermarkt

ANDREAS EICHLER, MÜNSTER UND WOLFGANG RIEMER, KÖLN

Zusammenfassung: *Die Analyse von Kassenbons, die in Supermärkten achtlos weggeworfen werden, ermöglicht eine detektivische Datenanalyse, die mit unterschiedlichen Schwerpunkten in allen Schulstufen möglich ist. In diesem Artikel wird Schritt für Schritt die Rekonstruktion des Mikrokosmos' eines Supermarkts aus zunächst leblosen Zahlenkolonnen nachgezeichnet – vom Konsumentenverhalten bis hin zum Umsatz des Supermarkts. Ein 'historisches' Nachwort zur Verwendung von Kasenzetteln im Stochastikunterricht zur Zeit der Euro-Umstellung beschließt den Artikel.*

1 Einführung

Sie liegen buchstäblich auf der Erde herum, zusammengeknüllt, weggeworfen oder im Einkaufswagen zurückgelassen: nach einem prüfenden Blick haben die Bons, die man an der Kasse eines Supermarkts erhalten hat, offenbar keinerlei Wert mehr für deren kurzfristige Besitzer. Doch für angehende Statistiker aller Schulstufen können sie spannende Einblicke in die Welt eines Supermarkts und Ansätze für die Analyse realer Daten geben, denen in ihrer Vielfalt quasi nur durch die Kreativität von Datendetektiven Grenzen gesetzt sind. Durch unterschiedliche Untersuchungsperspektiven und flexible Verwendung statistischer Methoden entsteht so Stück für Stück aus zunächst leblosen Zahlenkolonnen der Mikrokosmos eines Supermarkts.

Die Idee, Kassenbons auszuwerten, stammt vom zweiten Autor (2003, 2006) und wurde dort als Projekt vorgeschlagen. Einige der vielen Möglichkeiten, diese von der Erde gesammelten Daten auszuwerten, sollen in dieser Arbeit behandelt werden. Dabei geht zunächst der Blick in die unteren Klassen (auch in die Primarstufe), in der Fragen nach dem Konsumentenverhalten anhand eindimensionaler Datensätze thematisiert werden können. Der Bogen reicht dann bis zum Untersuchen von Zusammenhängen durch die Auswertung zweidimensionaler Datensätze, mit denen die Finanzlage eines Supermarktes aufgedeckt und fachlich die Datenanalyse mit der Wahrscheinlichkeitsrechnung verbunden wird. Einige weitere Betrachtungen, wie etwa die Auswirkung der Mehrwertsteuererhöhung oder die Frage nach der 'Verkäuferin des Monats' sind in die Datenanalyse eingeflochten. Die Arbeit ist insgesamt als Pool von Anregungen zu verstehen. Die Beispiele werden exemplarisch ausgeführt, ohne dass diese allerdings als fertigen Unterrichtsvorschlag für spezielle Klassenstufen auszubauen.

2 Datensammlung und erste Einsichten

2.1 Der Supermarkt

Der Supermarkt, um den es hier geht, liegt in Braunschweig und dort in einer gutbürgerlichen Gegend. Es ist ein Markt, der Kunden aus der umliegenden Gegend bedient und von Montag bis Samstag in der Zeit von 8 bis 20 Uhr geöffnet ist. Viele Kunden kommen zu Fuß und mit dem Rad, der Parkplatz ist relativ klein. Der Supermarkt unterscheidet sich damit deutlich von den großen Märkten in der Peripherie der Stadt mit großen Parkplätzen, geschaffen für einen großen Wocheneinkauf und Angeboten, die weit über Nahrungsmittel hinausreichen.

So ist das Angebot des Supermarkts zwar eher auf gehobene Bedürfnisse ausgelegt, bedient aber neben den üblichen Haushaltswaren wie Reinigungsmittel oder Pflegebedarf in erster Linie Nahrungsmittel und Getränke. Neben Angestellten, die ausschließlich für die Bestückung der Obst- und Gemüseabteilung oder für die Fleisch- und Käsetheke zuständig sind, bedienen 14 Verkäuferinnen bzw. Verkäufer die vier Kassen des Supermarkts.

2.2 Die Datensammlung

Als Blatt voller Daten kommt ein Kassenbon daher (siehe Abb. 1).[1] Einige der darauf enthaltenen Daten sind unmittelbar einsichtig, das sind die (meisten) Artikel-Bezeichnungen und zugehörigen Preise. Die Angaben zur Preis-Summe aller Waren, zum Hin- und Rückgeld sowie zur Mehrwertsteuerauszeichnung sind ebenso einsichtig. Dann folgen allerdings im unteren Bereich (siehe Pfeil) des Bons Angaben, die erst nach dem Betrachten mehrerer Kassenbons de-codiert werden können. So bezeichnet

- die erste vierstellige Nummer (Zahl) die Nummer des Kunden an einer bestimmten Kasse,
- die zweite vierstellige Nummer den speziellen Supermarkt,
- die erste dreistellige Nummer die Kasse und

- die zweite dreistellige Nummer die Verkäuferin bzw. den Verkäufer.

FRISCHEMARKT
[redacted]
[redacted] BRAUNSCHWEIG
TEL. 0531-[redacted]

GRANINI	0016 B %0	1,49
PUFFER	0037 A %0	2,49
SCHLAGSAHNE	0036 A %0	0,35
SCHLAGSAHNE	0036 A %0	0,35
OBST UND GEMÜSE	0061 A %0	1,76
MAGERQUARK	0036 A %0	0,35
MAGERQUARK	0036 A %0	0,35
KRAEUTERTOPF	0061 A %0	1,99
SALAT	0061 A %0	0,69
Posten 9	S u m m e	9,82
BAR	EUR	10,00
Zurück	EUR	-0,18
Summe-Netto	MWST A	7,79
MwSt	07.0%	0,54
Summe-Netto	MWST B	1,25
MwSt	19.0%	0,24
6065 0907/003/011	06.03.07 12:40	VA-00
Es bediente Sie:	[redacted]	

[redacted] Discount GMBH
Steuer-Nr.: [redacted]

QUITTUNG
**** Vielen Dank für Ihren Einkauf *****
**** Besuchen Sie uns bald wieder !*****

Wir lieben Lebensmittel
[redacted]

Abb. 1: Ein Kassenbon des Supermarkts

Das Datum und die Uhrzeit des Einkaufs sind schließlich ebenfalls noch auf dem Kassenbon enthalten. Was hier profan erzählt wird, ist schon Teil des Detektivspiels 'dem Supermarkt auf der Spur'. Die Zahlen bzw. Ziffern, die hier mit dem im Mathematikunterricht der Sekundarstufen eher vereinzelt behandelten Kodieraspekt daherkommen (vgl. z.B. Herget 1994), enthalten eine zunächst unbekannte Botschaft, die entschlüsselt werden muss. Dieses für jeden Supermarkt unterschiedliche Rätsel zu lösen ist überhaupt der Ansporn gewesen, mehr als einen Bon zu betrachten und hat schließlich die leidenschaftliche Jagd auf das Wegwefprodukt Kassenzettel ausgelöst.

Seit dem 6. November 2006 sind insgesamt bisher 372 Bons unterschiedlicher Tage und unterschiedlicher Tageszeiten zusammengekommen.

2.3 Die Güte der Daten

Da die Kassenzettel von einer Einzelperson bei diversen eigenen Einkäufen in dem Supermarkt gesammelt wurden, kann natürlich nur bedingt von einer Zufallsauswahl oder repräsentativen Auswahl gesprochen werden. Das liegt zum einen an der Zeit der eigenen Einkäufe und der zu diesen Zeiten verfügbaren Bons. Abbildung 2 zeigt die Verteilung der gesammelten Bons über die Wochentage.

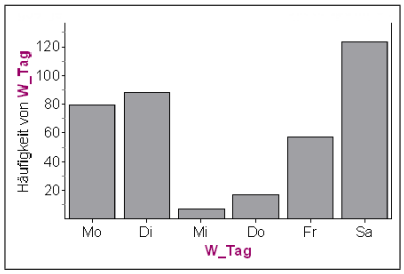

Abb. 2: Verteilung der Kassenbons über die Wochentage

Auch die Verteilung der Kassenbons über die Öffnungszeiten des Marktes ist nicht repräsentativ. Insbesondere die Randzeiten sind deutlich unterrepräsentiert (siehe Abb. 3)

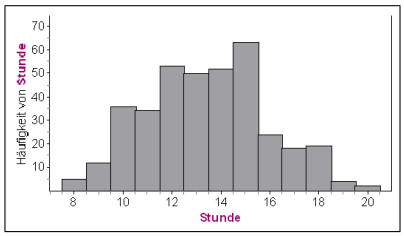

Abb. 3: Verteilung der Kassenbons über die Stunden

Ein anderes Manko der Datengüte ist die Frage, wer überhaupt seinen Kassenbon annimmt (wird er gleich an der Kasse gelassen, landet er nämlich in einem unerreichbaren Korb), wer ihn annimmt und kurze Zeit später achtlos zurücklässt und wer diesen mitnimmt. Rein qualitativ, aus den dreimonatigen Beobachtungen herrührend, ist kein Muster erkennbar gewesen, ob zum Beispiel erst bei besonders großen Einkäufen der Bon zur Kontrolle mitgenommen wird. Eine ebenso qualitative Einschränkung gibt es dabei: so scheinen Kartenzahler den Bon zusammen mit dem Einzugsbeleg für die Karte eher mitzunehmen als fallenzulassen. Ein Indiz dafür könnte das geringe Aufkommen an solchen Bons zu sein, auf denen die Kartenzahlung erkennbar dokumentiert ist. Andererseits gibt es keinen harten Beleg dafür, wie viele Kunden mit Karte bezahlen, sondern eher einen, der in der Literatur unter dem Stichwort 'Verfügbarkeitsheuristik' (Tversky und Kahnemann 1973) bekannt geworden ist. Diese 'Verfügbarkeitsheuristik' bedeutet in diesem Zusammenhang die subjektiv empfundene Häufigkeit (bzw. Wahrscheinlichkeit) von Kunden, die mit Karte bezahlen, denn schließlich ist man überzeugt davon, dass zumindest in der eigenen Schlange vor der Kasse prinzipiell alle mit einer (langsamen) Karte bezahlen, deren halb defekter Magnetstreifen mehrere Anläufe verursacht.

3 Wer kauft was - eine einfache Datenanalyse

Eine einfache, mit kleinen Datensätzen bereits in der Primarstufe anwendbare Analyse, betrifft das Konsumentenverhalten. Was wird eigentlich vorwiegend in solch einem Supermarkt gekauft? Ein Teildatensatz von 64 Kassenbons ergibt dabei folgendes Bild (siehe Tabelle 1):

Warenart	Umsatz (in Euro)	Anteil am Gesamtumsatz (in %)
Fleisch	126,48	13,9
Obst & Gemüse	123,21	13,5
Sonstiges	122,91	13,5
Süßigkeiten	77,77	8,5
Käse	68,0	7,5
Alkohol	63,23	7
Backwaren	37,85	4,1
Milch	35,4	3,9
Fertiggerichte	31,55	3,5
Getränke	30,34	3,3
Zigaretten	28	3,1
Backzutaten	25,73	2,8
Eier	25,68	2,8
Joghurt	24,71	2,7
Konserven (Gemüse)	14,59	1,6
Gewürze	14,31	1,6
Sanitärbedarf	13,29	1,5
Zeitschriften	12,6	1,4
Nudeln & Reis	9,83	1,1
Kaffee & Tee	9,68	1,1
Knabberwaren	8,38	1,0
Körperpflege	5,58	0,6
Konserven (Obst)	3,37	0,4

Tabelle 1: Verteilung der Umsätze auf einzelne Produktklassen

Eine Datenaufbereitung in diesem Sinne berührt das elementare Umgehen im Größenbereich Geld. Aus der Sicht der Datenanalyse ist hier der Aspekt der Klassierung interessant, der eine fundamentale Aufgabe jeglicher (Ein)Ordnung von Daten ist. So hat ein Supermarkt unüberschaubar viele Waren im Angebot. Unter welchen Kategorien oder eben Klassen soll man diese erfassen? Hier müssen Datendetektive vorher überlegen und kommen möglicherweise zu der Einsicht, die die Statistik insgesamt durchzieht: es gibt eigentlich keine 'richtigen' oder 'falschen' Klasseneinteilungen, sondern nur sinnvolle und weniger sinnvolle. Sinnvoll bedeutet hier allein, dass man die Fragen, die man an die Daten stellt, durch eine mehr oder minder willkürliche Klasseneinteilung beachtet. Auch die oben gegebene Klasseneinteilung ist willkürlich, eine Vielzahl anderer Einteilungen sind möglich. Je nach Fragestellungen könnte man andere, auch erheblich einfachere Klasseneinteilungen vornehmen, wie etwa *Getränke, Nahrungsmittel, Anderes* oder *Frischprodukte, konservierte Produkte* etc.

Eine Deutung oder Interpretation der Daten, ein grundsätzlicher Schritt des statistischen Denkens (Pfannkuch und Wild 1999), ist bei diesem Datensatz schwierig. Aus der Auswahl der Kassenbons ergibt sich etwa ein vermutlich zu hoher Anteil an Backzutaten und Süßigkeiten, da einige Kassenbons aus der Vorweihnachtszeit stammen und einen erheblichen Anteil an den jeweiligen Gesamtsummen haben. Ebenso wird man an den Tagen vor dem Wochenende einen erhöhten Umsatz an Alkoholika entdecken.

Sicher sind weitere Interpretationen anhand der Datentabelle möglich. Interessant und propädeutisch für die Statistik wichtig wird es aber dann, wenn man *vor* der Datenaufbereitung (aber nach einer Klasseneinteilung) eine wiederum fundamentale statistische Methode angewendet hat: nämlich die Hypothesenbildung bezüglich der Verteilung der Waren innerhalb des Gesamtbudgets. Dann hat man möglicherweise Argumentationsbedarf: Stimmen die Daten mit den Hypothesen überein? Ab welcher Abweichung geht man davon aus, dass die Hypothese falsch war? Gut wäre es dabei, einen weiteren Datensatz zur Verfügung zu haben, der ein erstes Erkennen der Variabilität statistischer Daten (Pfannkuch und Wild 1999) ermöglicht. An solch einem Beispiel lässt sich die Hypothesenbildung sowie die Beibehaltung bzw. Ablehnung von Hypothesen vorbereiten.

4 Dem Einzel-Umsatz auf der Spur

Spannend wird es, wenn man die Daten im Sinne des Umsatzes des Supermarkts hin- und herwendet. Auch das geht bezogen auf untere Klassenstufen (quasi) eindimensional.

Alle folgenden Bearbeitungen sind sinnvoll nur mit dem Rechner zu bearbeiten. Eine sehr geeignete Software, die unmittelbar das mehrfach bemühte Bild des Datendetektivs unterstützt, ist *fathom*, das seit vergangenem Jahr in deutscher Sprache auf dem Markt ist. Der ungeheure Vorteil ist die unmittelbare, sehr einfach handhabbare Verfügbarkeit von grafischen Datenaufbereitungen in allen schulrelevanten Repräsentationen. Der Nachteil ist, etwa gegenüber Excel, die (noch) geringe Verbreitung. Alle Grafiken in dieser Arbeit sind mit *fathom* erstellt.

4.1 Der Umsatz

4.1.1 Der Umsatz pro Kunde

Nimmt man alle Kassenbons ($n = 372$), dann ergibt sich eine leicht linkssteile Verteilung der Einzelumsätze (vgl. Abb. 21). Der Begriff 'Summe' in der Grafik und ebenso im Folgenden bezeichnet stets den Einzelumsatz pro Kunden.

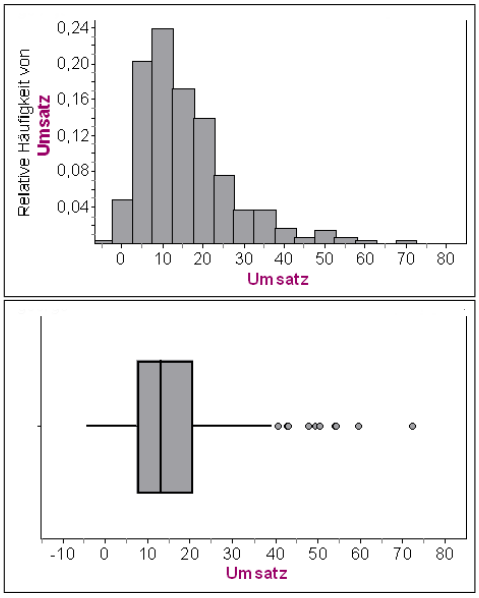

Abb. 4: Verteilung der Einzelumsätze

Man erkennt die oben bereits gegebene Charakterisierung des Supermarkts: 'Großeinkäufe' mit einem Umsatz über 50 Euro sind selten (zumindest in der Stichprobe), das Maximum liegt bei rund 72 Euro. Der Modalwert liegt bei der gewählten Klasseneinteilung bei 10 Euro, die Hälfte der Kunden tätigt Einkäufe zwischen etwa 10 und 20 Euro.

Unter der Annahme, dass es sich bei der Stichprobe um eine zumindest annähernd repräsentative Stichprobe handelt (siehe oben), erhält man über das arithmetische Mittel ($\bar{x} \approx 15,3$) eine Schätzung hinsichtlich des Umsatzes pro Kunden bei einer Standardabweichung von etwa 11 (Euro). Diese Schätzung wird im weiteren Verlauf der Analyse wichtig.

Mit Hilfe der (sinnvollen) Annahme, dass die Mittelwerte der Verteilung des Umsatzes pro Kunden in Stichproben von je 372 Kassenbons annähernd normalverteilt sind, ergibt sich nach dem Standardverfahren zur Berechnung von Konfidenzintervallen ($\bar{x} \pm \frac{s}{\sqrt{n}}$, mit $\bar{x} \approx 15$ und $s \approx 11$ aus der Stichprobe mit $n = 372$ Kassenbons) das Intervall $I_\mu = [14,2; 16,4]$ (Euro) für den tatsächlichen Umsatz (μ) pro Kunden (Konfidenzniveau: $\alpha = 0,95$).

4.1.2 Der Umsatz zu verschiedenen Zeitpunkten

Ein weiterer Aspekt bei der Untersuchung der Kassenbons könnte von der Frage ausgehen, ob sich beispielsweise an verschiedenen Wochentagen oder auch zu verschiedenen Tageszeiten Unterschiede in den Umsätzen ergeben. Diese Aspekte werden im Folgenden zunächst qualitativ, über den grafikgesteuerten Vergleich der Verteilung sowie den numerisch gesteuerten Vergleich von Kennzahlen der Verteilung, betrachtet. Anschließend folgt mit der Beurteilung mit Hilfe des Anpassungstest ein 'hartes' statistisches Verfahren.

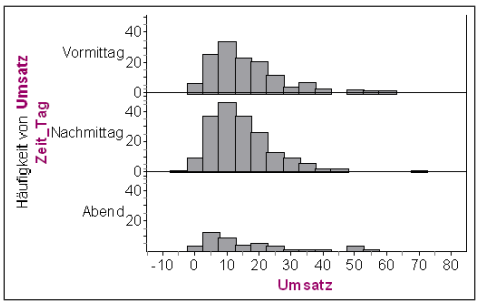

Abb. 5: Umsätze pro Kunden nach Tageszeiten

Bezüglich der Tageszeiten (vgl. Abb. 6) ergibt sich ein ähnliches Bild der Umsatzverteilung pro Kunden, wie es sich bereits in der Verteilung bezüglich aller Kassenbons gezeigt hat. Die Aufteilung in Tageszeiten bedeutet hier eine Drittelung der Öffnungszeiten (Morgen: 8-12 Uhr, Mittag: 12-16 Uhr, Abend: 16-20 Uhr). Methodisch betreibt man hier eine Clusterung der Daten und damit einen ersten, qualitativen Schritt in Richtung der Beurteilung von Abhängigkeiten zwischen zwei Merkmalen. Die Verfahren bleiben allerdings weiterhin die gleichen, die bei der Analyse eindimensionaler Daten verwendet werden.

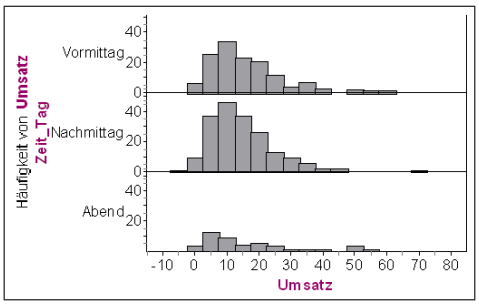

Abb. 6: Umsätze pro Kunden nach Tageszeiten

Ebenso ergibt sich das gleiche Bild, wenn man als Clusterung der Gesamtdaten die Wochentage nimmt (vgl. Abb. 7). Im Histogramm sind hier die Tage Mittwoch und Donnerstag auf Grund ihres geringen Vorkommens im Datensatz ausgenommen worden.

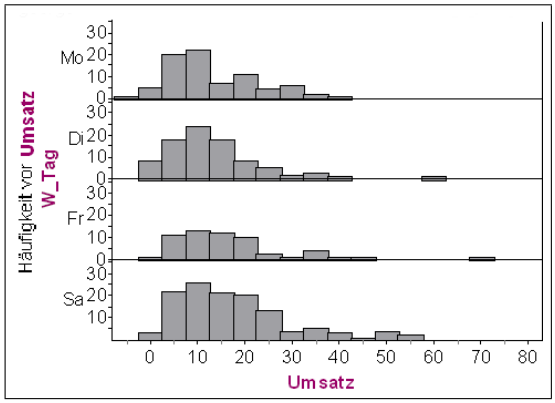

Abb. 7: Umsätze pro Kunden nach Tageszeiten

Im Sinne des Aufbaus des statistischen Denkens sind in diesem Schritt der Datenanalyse mehrere Aspekte wichtig.

Zunächst ist in nahezu allen realen Datensätzen die Frage nach Abhängigkeiten von Interesse und hier speziell die Frage, ob die Umsätze pro Kunden von der Tageszeit oder dem Tag selber abhängig sind. Diese Frage wird hier mit einem zunächst qualitativen Verfahren, der Clusterung der Daten und deren Visualisierung, aufgenommen.

Als zweiter wichtiger Aspekt ergibt sich der Vergleich von Verteilungen, der als fundamentaler Schritt im Aufbau statistischen Wissens eingeschätzt wird (Makar und Confrey 2004). Der Vergleich kann hier zunächst über die Betrachtung der Form der Verteilung, dann aber über die Betrachtung der charakteristischen Kennzahlen, d.h. der verfügbaren Lage- *und* Streuparameter erfolgen (vgl. Tabelle 2).

Zeitpunkt	$\bar{x}_{0,5}$	$\bar{x}$	SW	$Q_{0,5}$	σ
Morgen	13,0	15,7	59,6	13,7	11,2
Mittag	12,4	14,5	76,9	11,9	10,1
Abend	10,5	16,0	55,1	17,2	17,2
Montag	13,1	9,9	42,7	14,1	9,6
Dienstag	13,2	11,6	59,6	10,1	9,8
Freitag	16,5	14,5	72,0	12,3	12,4
Samstag	17,4	15,0	55,1	13,7	12,0

Tabelle 2: Kennzahlen zur Umsatzverteilung: Median, arithmetisches Mittel, Spannweite, Quartilsabstand und Standardabweichung

Eine wichtige Erweiterung zum Vergleich der Verteilungen ist die Simulation. Dabei wird nebenbei ein wichtiger Aspekt des statistischen Denkens berührt, nämlich die Erkenntnis der Variabilität statistischer Daten (Pfannkuch und Wild 1999). Nimmt man an, dass die Umsätze pro Person unabhängig von den Zeitpunkten der Messung sind, so kann man zufällige Stichproben aus dem Gesamtdatensatz ziehen (mit dem gleichen Umfang, der für verschiedene Messzeitpunkte vorhanden ist). Diese simulierten Stichproben sollten dann wiederum eine der realen Stichprobe 'ähnliche' Verteilung aufweisen. Durch die Simulation wird nun gerade die Erkenntnis der Variabilität gefördert. Jede Simulation ergibt eine neue, den anderen zwar ähnliche, im Detail dennoch unterschiedliche Verteilung (vgl. Abb. 8). Man kann aber trotz der *Variabilität* der Einzelverteilungen das Muster, die immer ähnliche Form der Verteilungen erkennen.

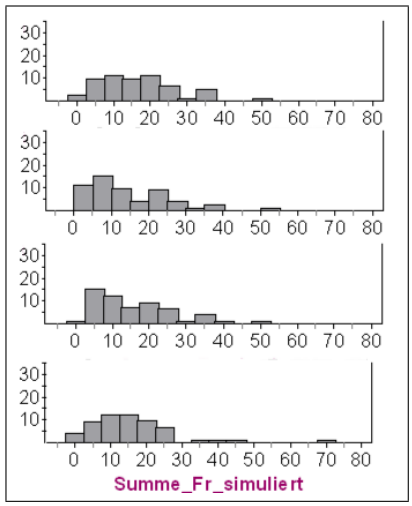

Abb. 8: Simulation von vier Freitagen (bzw. 57 Umsätzen)

Mit der statistischen Methode des Anpassungstests, der – wenn überhaupt – ein Thema am Ende der Schulstochastik ist, kann man die qualitative Untersuchung schließlich statistisch absichern. Dazu kann man beispielsweise die Cluster verschiedener Wochentage heranziehen und untersuchen, ob Unterschiede in den Verteilung hinsichtlich der verschiedenen Mess-Zeitpunkte innerhalb des Zufälligen liegen, d.h. mit dem χ^2-Anpassungstest sich jeweils die Hypothese der identischen Verteilung beibehalten lässt (die folgenden Schritte sind nur in Kürze beschrieben. Sie dienen nicht dazu, das Vorgehen des Anpassungstests zu erlernen, sondern allein das Nachvollziehen in Kenntnis dieses Tests).

Die Frage, die bei dem χ^2-Anpassungstest (hier in der Ausführungsvariante von Fisz 1973) beurteilt werden soll, ist die folgende: Passt die empirisch ermittelte Verteilung der Umsätze eines bestimmten Wochentags mit der ebenfalls empirisch ermittelten Gesamtverteilung der Umsätze überein, oder sind die Abweichungen der beiden Verteilungen voneinander überzufällig? Dazu werden beispielsweise hinsichtlich des Freitags die vorhandenen 57 Daten

in 5 Klassen mit 10 Daten und ein Randklasse mit 7 Daten aufgeteilt. Die Häufigkeit dieser Klassen ist $H_{fr}(k_i) = 10$ für $i = 1,...,5$ und $H_fr(k_i) = 6$ für $i = 6$; vgl. Tabelle 2). Als Vergleichsmaßstab gilt die Verteilung aller Umsätze. Mit der gleichen Klassierung ergeben sich die absoluten Häufigkeiten $H_{ges}(k_i)$. Daraus ergibt sich – ebenfalls mit der gleichen Klassierung – die theoretische Häufigkeit $H_{theo}(k_i)$ der Umsätze (wenn man die relative Häufigkeit hinsichtlich der Gesamtstichprobe als Wahrscheinlichkeit für die Klasse setzt) der 6 Klassen durch:

$$H_{theo}(k_i) = \frac{H_{ges}(k_i)}{n_{ges}} \cdot n_{fr} = \frac{H_{ges}(k_i)}{372} \cdot 57$$

Mit diesen Vorgaben lassen sich die einzelnen Summanden der χ^2-Testverteilung mit

$$t_i = \frac{(H_{theo}(k_i) - H_{fr}(k_i))^2}{H_{theo}(k_i)}$$

bestimmen, die Summe berechnen und mit Hilfe der χ^2-Verteilung die Anpassungshypothese beurteilen. In diesem Fall ergibt sich als Testgröße

$$t = \sum_{i=1}^{6} \approx 5,37$$

Der Vergleich mit der χ^2-Verteilung (mit 5 Freiheitsgraden) ergibt hinsichtlich des Wertes ihrer Verteilungsfunktion: $F_{\chi^2_{r=5}}(5,37) \approx 0,627$, d.h. die Hypothese der identischen Verteilung kann nicht abgelehnt werden.

Klasse (i)	$H_fr(k_i)$	$H_{ges}(k_i)$	$H_{theo}(k_i)$	t_i
1	10	83	12,7	0,58
2	10	68	10,4	0,02
3	10	70	10,7	0,05
4	10	37	5,67	3,31
5	10	82	12,6	0,52
6	7	32	4,9	0,90
Summe	57	372	57	5,37

Tabelle 3: Intervalle zum Anpassungstest

Man erkennt an der Tabelle, dass eine große Abweichung der theoretischen Häufigkeit von der in der Stichprobe vorhandenen (z.B. in der Klasse 4) einen großen Summanden hinsichtlich der Testgröße t ergibt. Insgesamt erkennt man ebenso, dass die für die reale Verteilung am Freitag durch die Wahl der Klassengrenzen künstlich erzeugten annähernden Gleichverteilung ebenso annähernd in der theoretischen Verteilung auftritt.

Ebenso wie bei der Verteilung für den Freitag kann auch für die übrigen Wochentage keine Abweichung von der Verteilung aller Umsätze pro Kunden statistisch nachgewiesen werden. Das ist ein erstaunliches Resultat, das möglicherweise wiederum auf die Charakteristik des Supermarkts, kein Ziel größerer Wochenendeinkäufe zu sein, zurückzuführen ist.

4.1.3 Kassen und ihre Umsätze pro Kunde

Die vier Kassen sind wie oben erwähnt nur selten alle geöffnet. Nahezu durchgehend ist die Kasse 1 in Betrieb, die anderen werden je nach Bedarf dazugeschaltet. Auch für die Kassen soll abschließend einzeln der Umsatz pro Kunde betrachtet werden (vgl. Abb. 9).

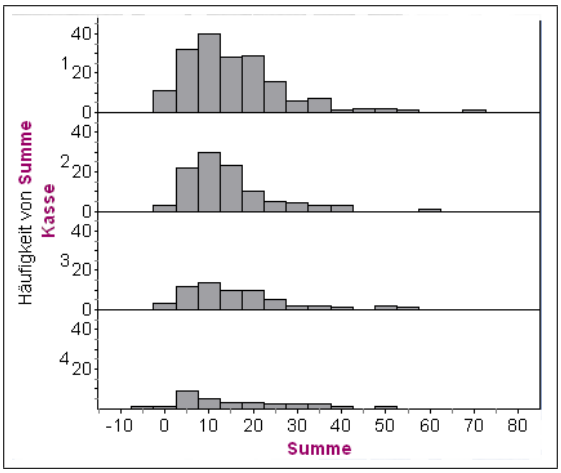

Abb. 9: Unterschiede zwischen den Kassen 1-4

Hinsichtlich der Kassen 1 bis 3 gibt es weder qualitativ noch quantitativ einen Zweifel, dass die Verteilung der Umsätze pro Kunden an den einzelnen Kassen identisch ist zur Gesamtverteilung dieser Umsätze. Hinsichtlich der Kasse 4 könnte man dagegen einen Zweifel hegen, denn hier deutet die visuelle Darstellung im Histogramm auf eine Verschiebung des Zentrums in Richtung der niedrigen Umsätze hin. Diese Verschiebung des Zentrums bei den leicht linkssteilen Verteilungen bezieht sich auf den Median, da das arithmetische Mittel hier durch die wenigen hohen Umsätze stark beeinflusst ist (vgl. Eichler 2007):

Kasse	$\bar{x}_{0,5}$	$\bar{x}$	x_{Mod}
1	13,23	15,53	10
2	11,88	14,19	10
3	13,62	16,29	10
4	11,36	14,55	5

Tabelle 4: Lageparameter zur nach Kassen geclusterten Verteilung der Umsätze

Allerdings kann man der Tabelle (4.1.3) entnehmen, dass diese Verschiebung hinsichtlich der dargestellten Kennzahlen auch für die Kasse 2 vorliegt. Auch

mit der Hinzunahme von Streuparametern lässt sich der visuell entstandene Zweifel nicht belegen. Einzig der Modalwert, der aber abhängig ist von der gewählten Klassierung, stützt den Verdacht.

Ein vertiefte (und quantitative) Untersuchung kann sich wiederum durch einen χ^2-Anpassungstest ergeben. Legt man wiederum die Gesamt-Verteilung der Umsätze pro Kunden zu Grunde, so gibt es hinsichtlich der Teil-Verteilung zu den Kassen 1-3 keinerlei Zweifel, dass diese den gleichen Mechanismen wie die Gesamtverteilung unterliegen. Auch für die Kasse 4 ergibt sich statistisch kein Zweifel ($\alpha \approx 0,314$).

Die in der Datenanalyse stets notwendige Verbindung der statistischen Untersuchungen mit dem Sachkontext (vgl. Pfannkuch und Wild 1999) kann dennoch eine Erklärung geben, warum möglicherweise die niedrigen Umsätze pro Kunde an der Kasse 4 leicht überrepräsentiert sind. Die Kasse 4 ist diejenige Kasse, die kaum über einen längeren Zeitraum geöffnet ist (vgl. dazu auch Abb. 10). Sie wird meist im Bedarfsfall kurzfristig zugeschaltet. Wird eine neue Kasse in dem Supermarkt geöffnet, so zeigt sich (nach Erfahrung), dass insbesondere diejenigen mit wenig Waren in der Hand eher den schnellen Kassenwechsel vollziehen, als diejenigen mit einem gefüllten Einkaufswagen. Ist die Kasse länger geöffnet, so wirkt sich das Verhalten der wenigen ersten 'Kassenwechsler' nicht aus. Wird dagegen die Kasse sehr schnell wieder geschlossen, so kann dies durchaus eine Auswirkung auf die Verteilung der Umsätze pro Kunden haben, wie man sie – ohne statistische Untermauerung – hinsichtlich der Kasse 4 vermuten kann.

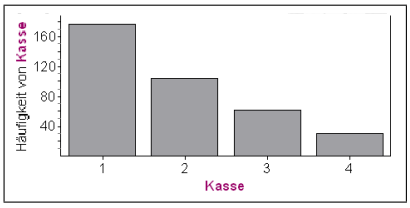

Abb. 10: Unterschiede zwischen den Kassen 1-4

4.1.4 Verkäuferin des Monats?

Für die Filialleitung eines Supermarkts, die nicht über eine Stichprobe, sondern über die Gesamtheit der Kassenbons verfügen, wäre es sicherlich von Interesse, wenn man Unterschiede im Umsatz pro Kunde zwischen den VerkäuferInnen feststellen könnte. Die 14 VerkäuferInnen in dem untersuchten Markt weisen die in Abb. 11 dargestellten Umsätze pro Kunden auf (hier ist die Darstellung auf diejenigen Verkäuferinnen beschränkt, von denen mehr als 40 Kassenbons vorhanden sind).

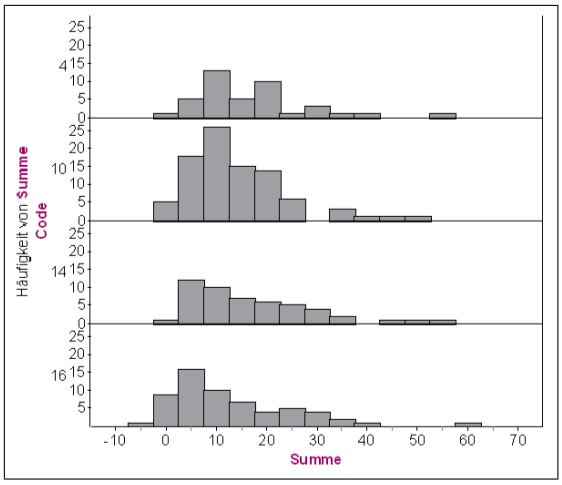

Abb. 11: Unterschiede zwischen vier Verkäuferinnen

Wieder könnte eine qualitative und statistische Beurteilung der Verteilungsunterschiede erfolgen. Tatsächlich ergibt sich für die Verkäuferin mit der Codenummer 16 ein signifikant oder überzufälliger Unterschied zur Gesamtverteilung der Umsätze pro Kunden. Mit einer Notation analog zur Betrachtung der Umsätze am Freitag verwendet wurden, ergibt sich:

Klasse (i)	$H_{Verk.nr.16}(k_i)$	$H_{ges}(k_i)$	$H_{theo}(k_i)$	t_i
1	10	18	2,9	17,35
2	10	49	7,9	0,56
3	10	36	5,8	3,03
4	10	104	16,77	2,74
5	10	65	10,48	0,022
Summe	50	372	50	26,02

Tabelle 5: Vergleich der Umsatzverteilung der Verkäuferin Nr. 16 und aller Umsätze

Man erkennt zunächst ohne Betrachtung der Testgröße, dass die künstlich herbeigeführte Gleichverteilung hinsichtlich der realen Daten zu dieser Verkäuferin in den theoretisch ermittelten Häufigkeiten kaum zu erkennen ist. Es scheinen in der realen Stichprobe die geringen Umsätze pro Kunde überrepräsentiert, die hohen Umsätze unterrepräsentiert zu sein. Es ergibt sich folglich ein Wert für die Testgröße $t = 26,02$, so dass die Hypothese der identischen Verteilung (fast sicher: $\alpha < 0,0001$) abgelehnt werden kann.

Was bedeutet das für diese Verkäuferin? Plakativ könnte man feststellen, dass diese Verkäuferin mit Sicherheit nicht als 'Verkäuferin des Monats' hinsichtlich des Umsatzes pro Kunden gekürt werden sollte. Bösartig könnte man gar eine Unterschlagung von Umsätzen vermuten (die aber praktisch schwer

zu realisieren wäre). Eine (von vielleicht mehreren) Hypothesen für die Ehrenrettung dieser Verkäuferin ergibt sich durch Zusatzinformationen zusammen mit der in Abb. 12 dargestellten Verteilung der Verkäuferinnen auf die vier Kassen.

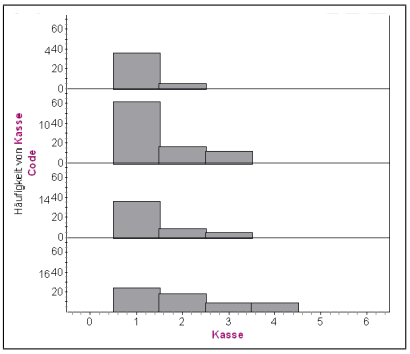

Abb. 12: Unterschiede zwischen vier Verkäuferinnen

Die Verkäuferin mit der Codenummer 16 ist relativ neu in dem Supermarkt. Offenbar wird sie nicht, wie ihre erfahrenen Kolleginnen, deutlich überwiegend an der Kasse 1 eingesetzt, an der der Großteil der Kunden bedient wird. Eine mögliche Hypothese zur Ehrenrettung ist dann folgende: Die noch junge Verkäuferin wird häufig in Überlastsituationen an den Kassen 2-4 eingesetzt (insbesondere an der Kasse 4, an der die anderen Verläuferinnen gar nicht eingesetzt werden), nach Abarbeiten der Überlast wiederum anderswo im Supermarkt beschäftigt. Möglicherweise wirkt sich diese Einsatz-Strategie auch auf die Umsätze pro Kunde dieser Verkäuferin aus (vgl. auch den vorangegangenen Abschnitt).

Evidenz durch die Daten, also eine statistische Untermauerung dieser Hypothese, gibt es allerdings durch die bisherigen Analysen nicht. Auch der Vergleich der Umsätze pro Kunde der Verkäuferin mit den Verteilungen der einzelnen Kassen ergibt jeweils die Ablehnung der Hypothese, dass beide jeweils betrachteten Verteilungen auf identischen Zufallsmechanismen beruhen (vgl. Tab. 6).

Kasse	*t*	α
1	24	0,0002
2	49,6	0
3	24,3	0,0002
4	14,6	0,0123

Tabelle 6: Werte der Testgröße *t* beim Vergleich Verkäuferin Nr. 16 und der Umsätze der vier Kassen

5 Dem Gesamt-Umsatz auf der Spur

Einen wesentlichen Anteil daran, die Spannung bei der Sammlung der Kassenbons aufrecht zu erhalten, hat die Untersuchung des Gesamtumsatzes ausgemacht, der sich aus den Umsätzen der vier Kassen ergibt. Eine weitere Klasse statistischer Methoden kommt bei dieser Analyse der Daten ins Spiel, bestehend aus den Verfahren der Regression und Korrelation. Um einen zumindest kleinen Eindruck der genannten Spannung zu vermitteln, werden im ersten Abschnitt die ersten Erkenntnisse in chronologischer Reihenfolge dargestellt und erst anschließend eine umfassendere Analyse vorgenommen.

5.1 Die ersten Schritte

Nach der Analyse der ersten Kassenbon-Daten (vgl. Abb. 1) war nicht eindeutig geklärt, was die Kodierungen im unteren Teil des Bons tatsächlich bedeuten. Insbesondere war zunächst die Zuordnung der ersten vier-ziffrigen Codes zur Nummer des Kunden (periodisch von 0 bis 9999) sowie die Zuordnung der ersten drei-ziffrigen Codes zur Kasse unklar. Die sich allmählich herausschälende Hypothese in diesem Sinne führte beschreibend zur Darstellung der Kundennummern an einer Kasse in Form einer Punktwolke bzw. einer Zeitreihe (vgl. Abb. 13). Die Skalierung der Zeitachse beschreibt die Öffnungszeit in Stunden seit dem 6. November 2006 (täglich 12 Stunden ohne Sonn- und Feiertage).

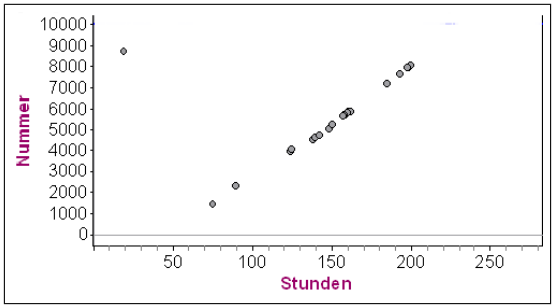

Abb. 13: Zeitreihe der Kundennummern an Kasse 1

Aus einer solchen Darstellung wurden folgende Hypothesen gewonnen:

1. Der erste vier-ziffrige Code beschreibt die Kundennummer, fortlaufend und sich periodisch wiederholend von 1 bis 9999 (ohne diese Hypothese wäre der linke Punkt in der Wolke kaum zu erklären gewesen)
2. Die Kundennummer kann gut durch eine lineare Funktion (in Abhängigkeit von der Öffnungszeit) beschrieben werden.

Mit der zweiten Hypothese geht die Möglichkeit einer Prognose einher, zu welchem Zeitpunkt der 10000. Kunde durch die Kasse 1 geht, wann also der

Umsprung von 9999 auf 1 erfolgt. Diese Prognose ist einerseits ebenfalls eine Hypothese, andererseits aber auch ein Kriterium für die bisher aufgestellten Hypothesen gewesen. Die Schätzung des 'Umsprungs' sowie die Einpassung der Regressionsgeraden (hier noch schlicht nach Augenmaß) sowie die Begrenzung der Kundennummer ($Nummer < 10000$, $Nummer \in \mathbf{N}$) sind in der folgenden Abbildung (Abb. 14) aufgenommen.

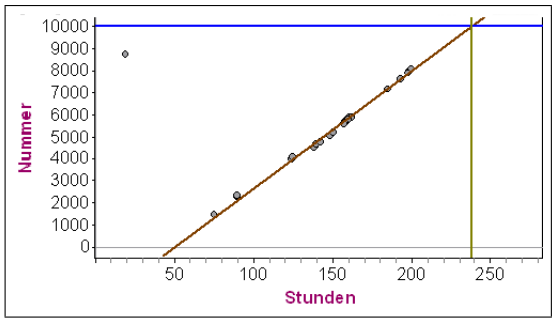

Abb. 14: Zeitreihe der Kundennummern an Kasse 1

Ergebnis dieser ersten Prognose: nach (ungefähr) 234 Stunden (nach dem 6.11.2006, 8 Uhr), d.h. (ungefähr) nach 19 Tagen und 6 Stunden Öffnungszeit, erfolgt der 'Umsprung'. Bezogen auf den Startpunkt bezieht sich die Prognose des 'Umsprungs' auf Dienstag, den 28. November 2006, gegen 14 Uhr. Ab diesem Zeitpunkt sollten dann, so die erweiterte Hypothese, die Punkte der Zeitreihe durch eine Gerade mit identischer Steigung gegenüber der vorangegangenen Periode beschreibbar sein.

Abbildung 15 zeigt Entwicklung der Zeitreihe nach der Sammlung der nächsten Kassenbons bis kurze Zeit nach dem erwarteten 'Umsprung'.

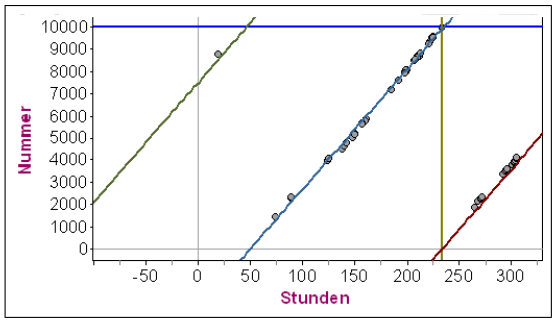

Abb. 15: Zeitreihe der Kundennummern an Kasse 1

Tatsächlich hat – von links kommend – der zuletzt gesammelte Kassenbon der sehr nahe an dem prognostizierten Termin liegt (Dienstag, 28.11.2006, 13.45 Uhr), die Nummer 9946. Ebenso erkennt man zumindest den prognostizierten 'Umsprung' sowie die Periodizität der Regression. Die eingepasste Gerade hat die Gleichung (mit gerundeten Achsenabschnitten)

$$g(Skala) := \begin{cases} 54 \cdot Stunden + 7400 & \text{für} \quad x < 48,15 \\ 54 \cdot Stunden - 2600 & \text{für} \quad x < 233,34 \\ 54 \cdot Stunden - 12600 & \text{sonst} \end{cases}$$

Die weiteren Schritte der Datenanalyse nach diesem Heureka-Effekt sind im Folgenden dargestellt.

5.2 Der Umsatz der einzelnen Kassen

Für die mathematisch gesteuerte Regression sowie die Beurteilung der linearen Abhängigkeit der Kunden-Nummern von der Öffnungszeit kann man zunächst die 'Umsprünge' vermeiden und die Nummern so transformieren, dass sie monoton wachsen (vgl. Abb. 16).

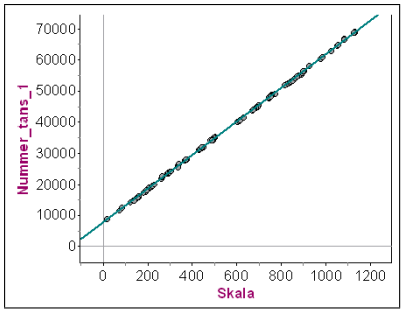

Abb. 16: Zeitreihe der Kundennummern an Kasse 1

Nach der Methode der kleinsten Quadrate für die Regression ist die Gerade mit der Gleichung

$$g(x) = 53,97 + 7370$$

optimal. Der Korrelationskoeffizient für diese Gerade ist 1, d.h. die Güte der linearen Abhängigkeit ist optimal. An den Datenpunkten selbst kann man in diesem Fall erkennen, dass der Wert des Korrelationskoeffizienten von 1 *nicht* bedeutet, dass alle Punkte auf einer Geraden liegen, die Abweichungen sind allerdings gering (vgl. Abb 17).

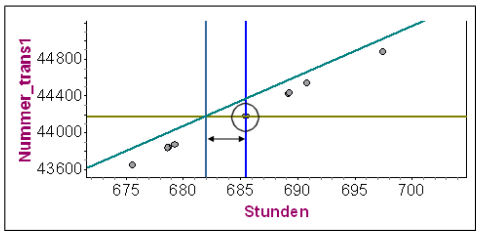

Abb. 17: Ausschnitt aus der Zeitreihe der Kundennummern an Kasse 1

In diesem Fall ist die Residuenbestimmung in Richtung der Zeitachse aussagekräftiger als die übliche

Bestimmung in vertikaler Richtung, also in Richtung der Achse der transformierten Kunden-Nummern. So gibt die Bestimmung der horizontalen Abweichungen an, wie weit zeitlich die Abweichungen der realen von den idealen Kundennummern sind. Für das markierte Datum (vgl. Abb 17) gilt etwa, dass diese Kundennummer gegenüber der idealen Kundennummer (eingepasste Gerade) um rund dreieinhalb Stunden zu spät durch die Kasse gegangen ist. Betrachtet man alle Daten in dieser Weise, so ergibt sich im Mittel eine zeitliche Abweichung von rund 4 Stunden und 40 Minuten (vgl. auch Abb. 18 zu allen Residuen in diesem Sinne).

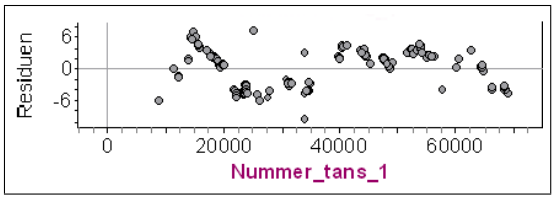

Abb. 18: Zeitliche Abweichungen der Kundennummern an Kasse 1 von der eingepassten Regressionsgeraden

Wie ist nun die Gerade zu interpretieren? Aus der Geradengleichung gewinnt man die Aussage, dass pro Stunde rund 54 Kunden bedient werden. Auch an dieser Stelle wäre es sinnvoll, solch einen Durchlauf-Koeffizienten vorab schätzen zu lassen. So ist der mittlere Durchlauf – trotz Kartenzahler, voller Einkaufswagen, mühsam die Cent suchende Kunden, fehlender Auswägungen von Frischgemüse und den, an Kasse 1 allerdings seltenen, Leerlaufzeiten – recht hoch. Es ist möglicherweise ein Durchlauf, in dem sich wiederum die Struktur des Supermarkts zeigt, in der die großen Wocheneinkäufe eher selten sind (siehe oben). Aus dem Durchlauf pro Stunden ergibt sich weiterhin, dass im Mittel etwa alle 185 Stunden bzw. alle 15 Tage und 4 Stunden 10 000 Kunden die Kasse 1 des Supermarkts passieren.

Kasse	Gerade (h für 'Stunden')	Durchlauf pro h, entspricht 10000 Kunden in
2	$g_1(x) = 46 \cdot h - 3100$ $r = 1$	46,0 18 d 1 h
3	$g_1(x) = 25,9 \cdot h + 4100$ $r = 1$	25,9 32 d 2 h
4	$g_1(x) = 14,4 \cdot h + 4200$ $r = 1$	14,4 57 d 10 h

Tabelle 7: Kunden-Durchläufe

Für die anderen Kassen, die bei Bedarf der Kasse 1 zugeschaltet werden, ergeben sich nach dem gleichen Analysemuster die in Tabelle 7 angeführten Geradenanpassungen bzw. Durchlaufkoeffizienten (vgl. auch Abb. 19):

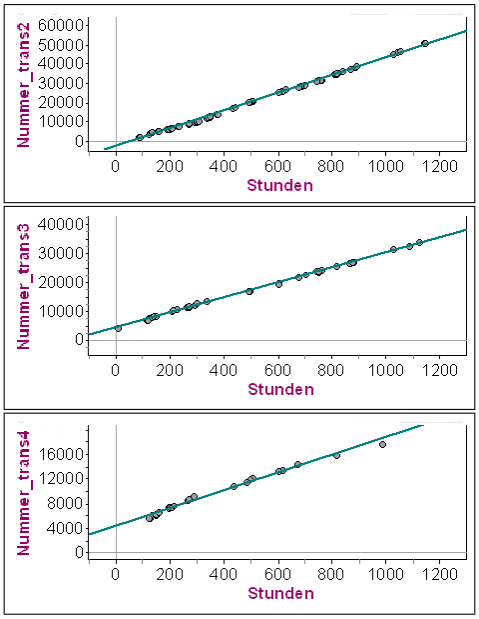

Abb. 19: Zeitreihe der Kundennummern an den Kassen 2, 3 und 4

Man erkennt an den Werten in der Tabelle **??** deutlich die übergreifende Strategie, die Kassen in der Reihenfolge ihrer Nummerierung zur Hauptkasse, der Kasse 1, hinzuzuschalten.

5.3 Der Umsatz des Supermarkts

Sind sowohl der Umsatz pro Kunde als auch der Kundendurchsatz ekannt, so kann man natürlich unmittelbar die Höhe der Kasseneinnahmen des Supermarkts in bestimmten Zeitabschnitten berechnen. Während hier der Kundendurchsatz sehr genau bestimmt werden kann, ist für den Umsatz pro Kunde ein Intervall gegeben, das bei angemessener Güte der Daten mit hoher Wahrscheinlichkeit den realen mittleren Umsatz pro Kunde enthält.

Es ergibt sich z.B. pro Tag im Mittel der Durchsatz von $12 \cdot (53,97 + 46 + 25,9 + 14,4) \approx 1683$ Kunden. Legt man weiterhin das mit dem Standardverfahren bestimmte Konfidenzintervall zum mittleren Umsatz pro Kunde, so erhält man Umsatz pro Tag in Euro das Intervall $[23900; 27600]$ (gerundet). Insgesamt ergibt sich

	Einnahmen in Euro	
Zeitspanne	untere Grenze	obere Grenze
Tag	23900	27600
Woche	167300	193200
Monat (26 Tage)	621500	717700
Jahr (2007)	7242300	8364400

Tabelle 8: Einnahmen des Supermarkts

5.4 Kalkulation eines Supermarktes

Aus dem Kassenumsatz die Kalkulation des Supermarkts zu rekonstruieren, ist ein komplexes Problem und würde den Artikel sprengen. Es sollen nur kurze Hinweise gegeben werden, die projektartig weiter verfolgt werden könnten, aber weit über die hier erfolgte Datenanalyse hinausgehen.

Ein Supermarkt hat das verständliche Ziel, Gewinn zu machen. Von den Einnahmen an der Kasse abzurechnende Kosten entstehen durch

- die monatliche Miete; solche Informationen kann man aus dem Netz beziehen.
- Nebenkosten, die schwer abzuschätzen sind (extensives Heizen im Winter, Kühlung von Waren etc.).
- Kosten entstehen für das Personal. Das monatliche Gehalt einer Verkäuferin bzw. eines Verkäufers ist (im Mittel) auch im Netz verfügbar[3]. Zuzurechnen wären die vom Arbeitgeber zur Hälfte zu bezahlenden Sozialabgaben. Zudem müsste man eine Schätzung machen, wie viele Angestellte ein Supermarkt abgesehen von denjenigen an den Kassen beschäftigt.
- die an den Staat gehende Mehrwertsteuer (vgl. auch den folgenden Abschnitt).
- natürlich durch den Einkauf der Waren. Kann man nach Abzug der an den Staat gehenden Mehrwertsteuer die Miet-, Neben- und Personalkosten bestimmen, so hat man die Möglichkeit, den Aufschlag des Supermarkts zu bestimmen, ab dem ein Gewinn (ohne Beachtung der Steuer) entsteht.

6 Ein letzter Aspekt - die Mehrwertsteuer

Als letzter Aspekt der Datenanalyse soll noch ein Blick auf die Erhöhung der Mehrwertsteuer geworfen werden, die zum Jahresbeginn von 16 auf 19 Prozent angehoben wurde. Was bedeutet das eigentlich für die Preise, falls die Erhöhung unmittelbar an die Kunden weitergegeben wird? Ein einfaches Modell könnte so aussehen: Vor der Erhöhung kostet ein Produkt netto, d.h. ohne Aufschlag der Mehrwertsteuer, 100 Geldeinheiten (als Normierung irgendeines konkreten Euro-Betrags). Damit kostete das Produkt $100 + 0,16 \cdot 100 = 116$ Geldeinheiten vor und $100 + 0,19 \cdot 100 = 119$ Geldeinheiten nach der Erhöhung der Mehrwertsteuer. Das Produkt kostet damit nach der Erhöhung $\frac{119}{116} \approx 1,0259$ mal so viel wie vorher oder: Die Erhöhung der Mehrwertsteuer hat das Produkt um rund 2,6 Prozent verteuert.

Die Rechnung gilt aber natürlich nur für Produkte, die der Mehrwertssteuer von ehemals 16 und nun 19 Prozent unterliegen. Ein Blick auf den Kassenbon (vgl. Abb. 1 zeigt, dass das nicht für alle im Supermarkt verkauften Waren gilt, sondern dass viele der Waren nur mit 7 Prozent (dem ermäßigten Mehrwertsteuersatz). Das Bundesfinanzministerium sagt dazu: „Der ermäßigte Steuersatz wird insbesondere auf die Lieferung, die Einfuhr und den innergemeinschaftlichen Erwerb von fast allen Lebensmitteln – ausgenommen Getränke ... – angewandt.“[2] Das bedeutet also, dass die Lebensmittel (fast) alle nur mit 7 Prozent besteuert werden. Dieser Steuersatz ist entgegen dem ’normalen’ Mehrwertsteuersatz, der im Supermarkt für Getränke (nicht Milchprodukte) gilt, zum Jahreswechsel nicht gestiegen. Zeitschriften unterliegen dem ermäßigten Steuersatz, Waren von Zigaretten, Pflegemitteln, Reinigungsmitteln bis hin zum Klopapier dagegen dem ’normalen’.

Die Erhöhung fast aller Waren, die nicht zu den Nahrungsmitteln gehören, hat sich folgendermaßen ausgewirkt:

Arithmetisches Mittel (Umsatzes pro Person)	
vor und	nach der Erhöhung
15,18€	15,42€

Damit haben sich die Waren (bezogen auf die Stichprobe) in dem Supermarkt um den Faktor $\frac{15,42}{15,18} \approx 1,016$ oder um 1,6 Prozent erhöht. Für eine nähere Analyse kann man noch die Betrachtungen aus Kapitel 2 nutzen. Legt man die dort vorhandene Verteilung der Warengruppen zugrunde, so unterliegen etwa 15,3 Prozent der erhöhten Mehrwertsteuer und die verbliebenen rund 85 Prozent nicht. Legt man diese Verteilung des sicherlich nicht allzu großen Datensatzes zugrunde, dann sollten sich 15,3 Prozent des mittleren Umsatzes um rund 2,6 Prozent erhöht haben. Es wäre also demnach im Jahr 2007 ein folgender mittlerer Umsatz zu erwarten gewesen:

Arithmetisches Mittel (Umsatzes pro Person)	
vor und	nach der Erhöhung
15,18€	$15,18 \cdot (1 + 0,153 \cdot 0,026) \approx 15,24$€

Zu erwarten wäre also durch die erhöhte Mehrwertsteuer nur eine Erhöhung von rund 0,4 Prozent ($0,153 \cdot 0,026 \approx 0,004$) gewesen, statt der offenbaren Verteuerung der Waren im Jahr 2007 von rund 1,6 Prozent. Mit anderen Worten: Der Supermarkt ist

teurer geworden, als es die Erhöhung der Mehrwertsteuer hergibt.

Wie ist das zu erklären? Es gibt zwei Hypothesen, die zwar nicht belegbar, aber dennoch plausibel sind. Die erste ist simpel: Es hat zum Jahreswechsel eine allgemeine Preiserhöhung der Waren gegeben. Auch die zweite Hypothese ist möglich: Die Preise im Supermarkt enden zumeist auf 9 Cent, es hat also etwa ein Getränk vor der Erhöhung 1,19 Euro gekostet. Würde die Mehrwertsteuer unmittelbar und in der richtigen Höhe an die Kunden weitergegeben werden, so müsste das Getränk $1,19 \cdot 1,026 \approx 1,22$ Euro kosten. Das gibt es aber im Supermarkt nicht. So wird man eventuell in einigen Fällen bei 1,19 Euro geblieben sein, im der größeren Zahl der (Waren)Fälle den Preis auf 1,29 Euro erhöht haben. Auch dadurch kann sich die zu große Erhöhung des mittleren Umsatzes erklären.

7 Didaktische Anmerkungen

In der Analyse von Daten, die auf dem Boden liegen, steckt einiges didaktisches Potenzial. Zunächst ist die Analyse der Daten in unterschiedlichen Anforderungsstufen in verschiedenen Klassenstufen möglich, prinzipiell vom Ende der Primarstufe bis in die Sekundarstufe II. Die Spannweite liegt inhaltlich von einfachen Häufigkeitsbestimmmungen bis hin zu Verfahren der beurteilenden Statistik wie Anpassungstest. Innerhalb einzelner Untersuchungsschritte kann die Schwierigkeit variiert werden. So sind an manchen Stellen wie etwa dem Vergleich von Umsätzen pro Kunde an bestimmten Tagen mit dem allgemeinen Umsatz pro Kunde verschiedene Methoden einsetzbar, für die Sekundarstufe I eine qualitative Untersuchung von Verteilungen durch Betrachtung von Mittelwerten und Streuung, für die Sekundarstufe evtl. mit dem Einsatz von inferenzstatistischen Verfahren.

Ein Unterrichtsprojekt zu Kassenbons enthält weiterhin alle Stufen, die eine „komplette statistische Untersuchung“ (Biehler, Hartung 2006, 53) aufweist: „*Problemstellung – Planung der Erhebung – Datenerhebung – Auswertung – Interpretation – Schlussfolgerung und Ergebnisbericht*“ (ebd.). Es ist zudem über die Betrachtung der Kalkulation eines Supermarkts noch erheblich erweiterbar.

Versteht man die Datenanalyse als *Kompetenz*, die Schüler in ihrer Schulzeit erwerben sollen, so kann man die Aufteilung des statistischen Denkens, die Wild und Pfannkuch (1999) formuliert haben, als Auflistung von Teilkompetenzen verstehen:

- Erkennen der Notwendigkeit von Daten
- Flexible Aufbereitung von Daten
- Erkennen der Variabilität von Daten
- Erkennen und Beschreiben von Mustern in den Daten
- Verbindung mit dem Sachkontext.

Im Sinne dieser Teilaspekte des statistischen Denkens soll die Datenanalyse zu den Kassenbons noch einmal rückblickend in Kürze betrachtet werden.

Erkennen der Notwendigkeit von Daten: Für einen Statistiker ist die Formulierung dieser Stufe wahrscheinlich banal, für Schüler zunächst einmal nicht. Hier ist die Erfahrung aufzubauen, dass über das Vergößern von Datensätzen Beurteilungen von Situationen erst einmal über ’Stammtischmeinungen’ bzw. naive Schätzungen hinauskommen.

In dem vorgestellten Projekt wird die Notwendigkeit einer vergößerten Datenbasis in mehreren Phasen deutlich. Zunächst braucht man mehr Daten (Kassenbons), um die Kodierungen des Kassenzettels zu verstehen. Glaubt man zu verstehen (hypothetisch), so braucht man mehr Daten, um seine Hypothesen zu bestätigen (etwas die Kodierung der Kundennummer). Will man weitere Untersuchungen machen (etwa zum Umsatz pro Kunde aufgeteilt nach VerkäuferInnen), so braucht man unter Umständen noch mehr Daten und kann dabei die Grenzen der Ausagefähigkeit eines begrenzten Datensatzes erkennen.

Flexible Aufbereitung: Die flexible Datenaufbereitung durchzieht die explorative Datenanalyse der Kassenzettel. So werden die Daten nach unterschiedlichen Gesichtspunkten grafisch dargestellt und analysiert, etwa als Gesamtdatensatz oder geclustert nach Aspekten wie Einzelkassen, Einzeltagen oder Einzelverkäuferinnen.

Die eingesetzten Methoden können breit gestreut werden. So kann die Analyse eher qualitative Untersuchungen verschiedener Diagramme (Boxplots oder Säulendiagramme), oder eher quantitativ-statistische Verfahren wie etwa den Anpassungstest umfassen. Die flexible Datenaufbereitung hat dabei zum Ziel, in möglichst vielfältiger Weise Muster in den Daten zu erkennen und beispielgebunden Möglichkeiten und Grenzen einzelner statistischer Methoden zu erfahren.

Erkennen der Variabilität von Daten: Dieser Aspekt wird in der Analyse der Kassenbons eher un-

terschwellig behandelt. Dennoch ergibt sich etwa bei der Analyse der nach Wochentagen oder Verkäuferinnen geclusterten Umsätze pro Person stets die Frage, ob die Unterschiede in den Verteilungen auf eben der Variabilität der Daten beruhen oder ob systematische Unterschiede bestehen. Einen etwas expliziteren Zugang zum Erkennen der Variabilität kann dagegen die in Kapitel 4.1.2 (vgl. S.5) simulierte Verteilung von Umsätzen pro Person an vier virtuellen Freitagen leisten.

Erkennen und Beschreiben von Mustern in den Daten: Die Suche nach Mustern in den Daten ist ein Kernstück der Datenanalyse. Hier geht es einerseits um die qualitative Beschreibung von Mustern, etwa der Verteilung von Umsätzen pro Kunde. Die statistische Analyse mit Hilfe des Anpassungstest hilft bei der Beurteilung, ob ein Muster (in den Umsätzen pro Kunde) mit den in den Daten steckenden empirischen Phänomenen wie etwa der Umsatzverteilung pro Kunde an bestimmten Tagen übereinstimmt. Schließlich lässt sich der Durchlauf von Kunden an den Kassen in fast idealer Form mit Hilfe eines Musters, einer Regressionsgeraden, beschreiben und die Güte des Musters mit Hilfe des Korrelationskoeffizienten beurteilen.

Verbindung mit dem Sachkontext: Die Verbindung der Datenanalyse mit dem Sachkontext ist bei realen Daten fast von selbst gegeben. An verschiedenen Stellen wird aber gerade bei der Analyse der Kassenbons der Sachkontext Ausgangspunkt von Erklärungsversuchen, wenn die statistische Analyse keine Erklärungsansätze mehr bieten kann. Eines von mehreren Beispielen sind die Hypothesen zur statistischen Auffälligkeit des Umsatzes pro Kunden einer bestimmten Verkäuferin.

8 Euro-Debatte und die Exponentialverteilung, ein 'historischer' Nachtrag.

Als der zweitgenannte Autor im April 2001 mit seinen Schülern erstmals auf Kassenzettel-Jagd ging, dachte noch niemand an den Euro.

Im April 2002 war er schon 4 Monate als der berüchtigte 'Teuro' unter uns. Und zu allem Ärger hingen da seit Wochen diese Plakate 'im ALDI', die weismachen wollen, dass bei den Lebensmitteldiscountern alles billiger geworden ist, erinnern Sie sich? Wieder so eine geschickte Werbemasche? Hat man doch oft gehört. Und überall gelesen. Kurz vor der Euroumstellung Preise ´rauf, um dann im Frühjahr ein bisschen billiger werden zu können. Und satte Preiserhöhungen unter dem Strich....

Die Stochastik-Kurse am Heinrich-Mann-Gymnasium in Köln gingen erneut auf Zetteljagd... und das Stöbern in dem alten Zettelhaufen begann von vorne. Man musste auf den alten (zum Glück noch nicht entsorgten) und den neuen Kassenzetteln gleiche Waren finden. Die ersten Ergebnisse zeigen: Brechbohnen wurden teurer (+), Toilettenpapier billiger (-). Was sich hinter 'Wacholderbauch' (-) verbirgt, blieb zunächst ungeklärt.

Ware	März 2001 – März 2002 1€= 1,95583DM	Diff.	Vorz.
Backofen Pommes	1,32DM (0,67€) – 0,86€	0,19€	+
Bandnudeln	1,29DM (0,66€) – 0,65€	-0,01€	-
Coca Cola 2l	2,79DM (1,43 €) – 1,27€	-0,16€	-
Fischstäbchen	2,49DM (1,27 €) – 1,39€	0,12€	+
Brechbohnen	1,39DM (0,71€) – 0,79€	0,08€	+
Gouda	4,17DM (2,13€) – 2,27€	0,14€	+
Klarspüler	1,99DM (1,02€) – 1,17€	0,15€	+
Toiletten-papier	4,98DM (2,55€)– 2,49€	-0,06€	-
Wachholder-bauch	1,99DM (1,02€) – 0,99€	-0,03€	-
Waffeleier	1,29DM (0,66€) – 0,69€	0,03€	+

Tabelle 9: kleine Stichprobe vom Umfang 10

Und mit dieser Stichprobe war man beim einfachsten Verfahren der beurteilenden Statistik, beim Vorzeichentest, 'gelandet'. Und zu unserer Verblüffung - und allen Unkenrufen zum Trotz mussten wir feststellen, dass tatsächlich von den 360 verglichenen Preisen, 213 (das sind 59 Prozent) gesenkt worden waren.

Genauere Untersuchungen zerrten dann den tieferen Grund für diese vielen Preissenkungen ans Tageslicht: Nicht eine barmherzige Kundenfreundlichkeit, sondern die sublime Psychologie der 'Schwellenpreise': Wenn früher ein Produkt 1,99DM gekostet hätte, müsste es nun wegen des Umrechnungskurses 1,01€ kosten. Wenn man daraus 0,99€ macht, hat ALDI 2 Cent 'verschenkt', kann aber wieder einen psychologisch günstigen Schwellenpreis anbieten. Diese Vielzahl minimaler Preissenkungen wurden durch massivere Preiserhöhungen bei

einzelnen Produkten ausgeglichen. Unsere 'ALDI-Kassenzettelforschung' ergab insgesamt eine (recht moderate) Preiserhöhung von ca 2 Prozent während eines Jahres. Wer mochte, konnte darauf guten Gewissens reimen: 'Willst du keinen Teuro haben, musst du dich bei ALDI laben'.

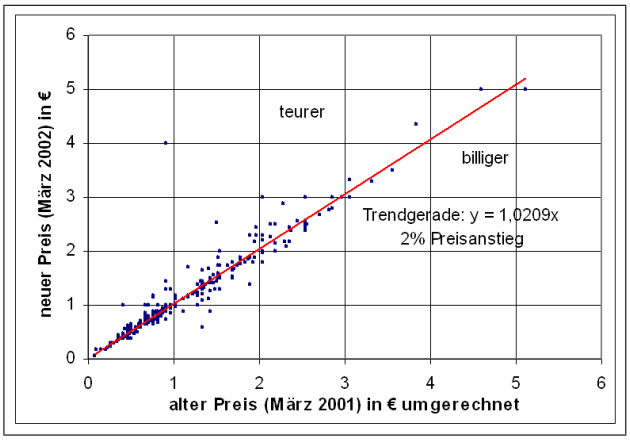

Abb. 20: Alte und neue Preise im Vergleich

Eine weitere überraschende Entdeckung machten Paul, David und Wladi, die inzwischen alle Mathe und Informatik studieren: Die Rechnungsbeträge sind nicht, wie man vielleicht erwarten würde, normalverteilt: Unabhängig von der Supermarktkette erwiesen sie sich als exponentialverteilt. Das gilt für Aldi, für Lidl, für Plus und selbst für Kaiser's. Das 'verrückte' ist, dass man genau diese Verteilung erhält, wenn man im Rahmen einer Modellrechnung 'virtuelle Kunden' beim Einkauf durch den Laden schlendern lässt, die jede Ware, an der sie vorbeikommen, rein zufällig und unabhängig von dem, was schon im Korb ist mit einer festen Wahrscheinlichkeit hinzupacken.

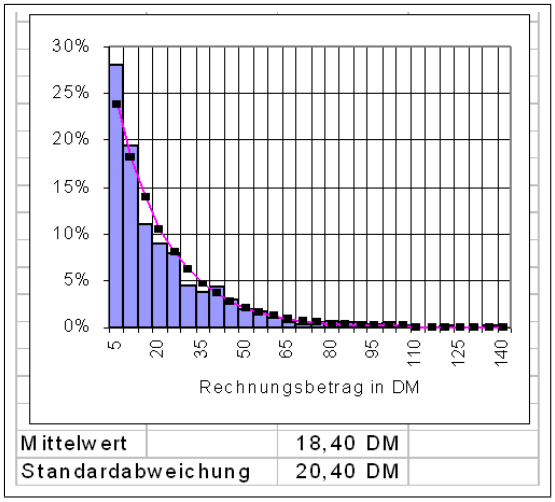

Abb. 21: Exponentialverteilung der Rechnungsbeträge (Säulen) mit virtuellen Kunden (Kurve)

Anmerkungen

[1] Die Supermärkte haben verschiedene Vorgehensweisen bei den Angaben auf ihren Kassenbons. Welche Kodierungen von einzelnen Märkten aufgeführt werden, kann daher nicht allgemein angegeben werden. Die De-Codierung ist damit immer wieder eine neue Aufgabe.

[2] Vgl. http://www.bundesfinanzministerium.de/cln_04/nn_3506/DE/Service/Lexikon_A_Z/U/001.html

[3] Vgl http://www.boeckler.de/cps/rde/xchg/SID-3D0AB75D-FEC5218E/hbs/hs.xsl/550.html.

Literatur

Biehler, R., Hartung, R. (2006). Die Leitidee Daten und Zufall. In Blum, W. Drüke-Noe, C., Hartungs, R., Köller, O. (Hrsg.), Bildungsstandards Mathematik: konkret. Berlin: Cornelsen-Scriptor, S. 51-80.

Eichler, A. (2007) „Geld weg – Arzt weg!"— Was ist dran am Ärzteprotest?. Praxis der Mathematik 49(13), S.20-26.

Fisz, M. (1973) Wahrscheinlichkeitsrechnung und mathematische Statistik. Berlin: VEB.

Herget, W. (1994) Artikelnummern und Zebrastreifen, Balkencode und Pruefziffern - Mathematik und Informatik im Alltag. In Blum, W., Henn, W., Klika, M., Maass, J. (Hrsg.), Materialien fuer einen realitaetsbezogenen Mathematikunterricht. Bd. 1. Hildesheim: Franzbecker. 1994. S.69-84.

Makar, K. und Confrey, J. (2004). Secondary teachers' statistical reasoning in comparing two groups. In Ben-Zvi, D., und Garfield, J. B. (Hrsg.), The challenge of developing statistical literacy, reasoning, and thinking (S. 353-373). Dordrecht: Kluwer.

Pfannkuch, M., Wild, C. (1999). Statistical Thinking in Empirical Enquiry. In: International Statistical Review 67(3) 1999, S.223-248.

Riemer, W. (2003). Stochastik. Stuttgart: Klett-Verlag, S.154.

Riemer, W. (2006). Lambacher-Schweizer 6 NRW. Stuttgart: Klett-Verlag, S.164.

Tversky, A. und Kahnemann, D. (1973). Availability: A heuristic for judging frequency and probability. Cognitive Psachology 5, 207-232.

Anschrift der Verfasser
Andreas Eichler
Institut für Didaktik der Mathematik
Westfälische Wilhelms-Universität Münster
Fliednerstraße 21
48149 Münster
a.eichler@uni-muenster.de
Wolfgang Riemer
Studienseminar für Lehrämter an Schulen
Claudusstraße 1
50678 Köln