Datenanalyse mit einfachen, didaktisch gestalteten Softwarewerkzeugen für Schülerinnen und Schüler

Von Cliff Konold

Wenige Menschen würden der Aussage widersprechen, daß der Computer die Praxis der Statistik verändert hat. Noch weniger würden bestreiten, daß der Computer bisher wenig Einfluß im in Statistikunterricht gehabt hat; viele meinen immer noch, er solle bei einführenden Kursen keine wichtige Rolle spielen. Bei Benutzung geeigneter Software sind häufig gehörte Einwände aber zu entkräften.

Entwurf von Datenanalysewerkzeugen für Schülerinnen und Schüler

In diesem Artikel beschreibe ich Grundsätze, die das Design von Software zur Datenanalyse beeinflußt haben, die wir seit kurzem speziell für Schülerinnen und Schüler mit geringen oder keinen Erfahrungen mit Datenanalyse entwickelt haben. Ich konzentriere mich dabei auf Eigenschaften der Software, die zwingende Gründe gegen die immer noch oft gehörten Einwände liefern, daß Einführungen in die Statistik durch die gleichzeitige Gewöhnung der Schülerinnen und Schüler an den Computer sowie ein Softwarewerkzeug bloß weiter kompliziert werden. Der Kleincomputer eröffnet Statistikanfängern unter den Schülerinnen und Schülern Möglichkeiten der Betrachtung und der Befragung von Daten, die sie potentiell rasch über die technischen Aspekte der Datenanalyse hinaus zu dem vordringen lassen, worum es bei dieser Tätigkeit geht: Wie man interessante Fragen stellt und plausible Antworten produziert. Obwohl sich der Aufsatz hauptsächlich darauf konzentriert, wie Software diesen Prozeß unterstützen kann, versuche ich auch, verschiedene Aspekte des iterativen Charakters der Datenanalyse zu schildern, und die Schwierigkeiten anzudeuten, die Schülerinnen und Schüler dabei haben, Konzentration und Interesse in einer möglichen Flut von Informationen aufrechtzuerhalten.

Die von mir beschriebene Software *DataScope* wurde im Rahmen eines Vier-Jahres-Projektes entwickelt, daß von der National Science Foundation finanziert wurde, um Materialien und Software für den Unterricht in Datenanalyse auf dem Niveau der Highschool und der ersten Semester am College zu entwickeln. Wir haben DataScope entworfen, um eine von uns wahrgenommene Lücke zwischen Einführungskursen zu komplexen professionellen Analysewerkzeugen (zum Beispiel StatView und DataDesk) und vorhandener Unterrichtssoftware (zum Beispiel Statistics Workshop, Data Insights) zu schließen, die nicht leistungsfähig genug waren, um die von uns angestrebte Art der Analyse zu unterstützen. DataScope ist sicher nicht das Unterrichtswerzeug, daß wir uns erträumt hatten (vergleiche Biehler, 1994), doch ermöglicht es den Schülerinnen und Schülern einen leichten Zugang zu einer beachtlichen Datenanalyse-Kapazität.

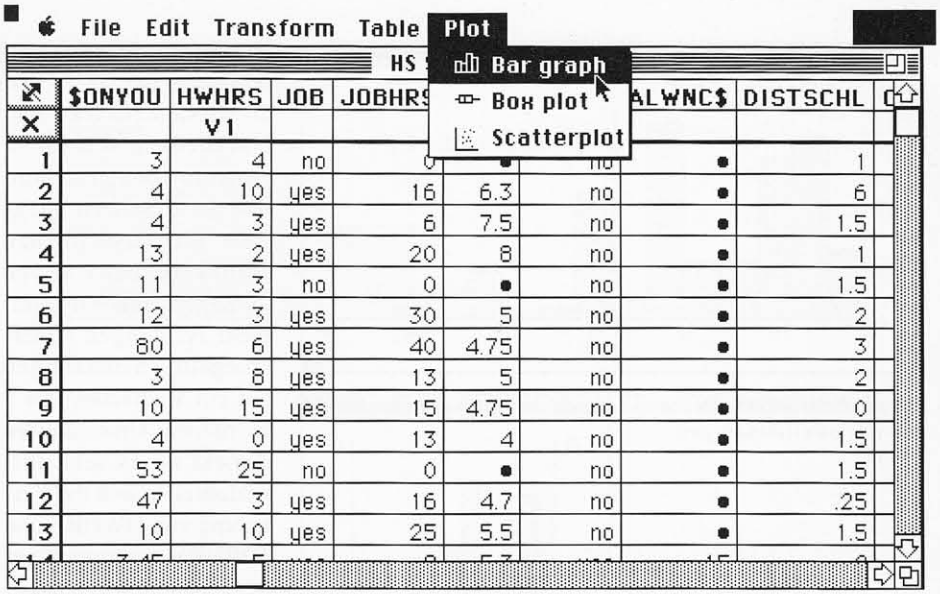
Ziele der Datenanalyse und Softwaredesign

Es gibt allerlei Ansichten darüber, was wir im Namen von Statistik oder Datenanalyse unterrichten sollen (vergleiche Gordon und Gordon, 1992). Steht man vor der Frage, Software für Unterrichtszwecke zu entwerfen oder auszuwählen, sollte man lieber eine klare Vorstellung davon haben, was man erreichen möchte. Das Leitziel für das Entwerfen von DataScope war, Schülerinnen und Schüler so rasch wie möglich an den Punkt zu führen, wo sie die Software benutzen können, um einer Reihe aufeinander bezogener Fragen mit dem Ziel nachzugehen, eine kohärente „Geschichte" zu einem Datensatz zu produzieren. Die Daten sollen dabei i. a. viele Variablen enthalten, die mit zu verarbeiten sind. Datenanalyse wird oft als interaktiver, iterativer Prozeß geschildert, in dem der oder die Einzelne ausgehend von einer Frage relevante Daten sammelt und untersucht und schließlich die Frage umformuliert und zuspitzt, danach weitere Daten betrachtet, und so weiter. Zwar hängt diese Tätigkeit von den Wahrnehmungsfähigkeiten der oder des Einzelnen ab, doch wird sie von Neugier und dem Wunsch getrieben, plausible Darstellungen zu produzieren, mit denen verschiedene Beobachtungen erklärt oder in einen Zusammenhang gebracht werden können. Unterrichtssoftware kann die Erreichung dieses Ziels befördern, indem sie Schülerinnen und Schüler dabei unterstützt, Muster in Daten, die „Textur" von Daten, zu erkennen, indem sie ihre visuellen Fähigkeiten zur Ermittlung von Trends, Mustern und Unterschieden nutzt und weiterentwickelt, und dadurch, daß sie so leicht anzuwenden ist, daß die ansonsten mit dem Erlernen der Software verbrachte Zeit dazu genutzt werden kann, die Daten zu analysieren und zu durchdenken.

Einfachheit

Eine Möglichkeit, DataScope einfach zu gestalten, bestand darin, die Zahl der angebotenen Graphiktypen zu begrenzen. Wir haben zwei Gründe, deren Zahl gering zu halten. Der einleuchtendste war, daß die Software um so leichter zu lernen ist, je weniger Graphiken sie aufweist und je unkomplizierter sie ist. Vorstellbar ist aber auch ein Unterrichtsmittel mit vielen Graphiken und Optionen, das aber so entworfen ist, daß es Anfänger nicht überfordert und dessen Leistungsfähigkeit sich in dem Maße zeigt, wie die Bedürfnisse der Benutzerinnen und Benutzer wachsen (vergleiche Biehler, 1994). Der triftigere Grund, die Zahl der Graphiken in einem Einführungskurs zu beschränken, liegt darin, daß zum Erwerb von Kompetenz zum „Lesen" eines bestimmten Diagramms Zeit erforderlich ist. Für viele Menschen ist die Interpretation eines Histogramms zweite

Natur geworden. Ein Experte kann einen Blick auf ein Histogramm werfen, rasch sowohl typische wie atypische Merkmale vermerken und diese Information dazu nutzen, die weitere Exploration anzuleiten. Doch für viele Anfängerinnen und Anfänger ist ein Histogramm immer noch ein Informationswirrwarr. Sie wissen nicht, worauf sie achten sollten und können das Ungewöhnliche nicht wahrnehmen. Die Begrenzung der Anzahl von Graphiktypen in einem Einführungskurs ermöglicht den Schülerinnen und Schülern, auf jede Graphik soviel Zeit zu verwenden, um die Erfahrung zu erlangen, durch die ein Werkzeug am Ende zur unbewußten Erweiterung unseres üblichen Wahrnehmungssystems wird (vgl. Polanyi, 1969).



	$ONYOU	HWHRS	JOB	JOBHRS		ALWNC$	DISTSCHL
X		V1					
1	3	4	no	0	•	•	1
2	4	10	yes	16	6.3	no	•	6
3	4	3	yes	6	7.5	no	•	1.5
4	13	2	yes	20	8	no	•	1
5	11	3	no	0	•	no	•	1.5
6	12	3	yes	30	5	no	•	2
7	80	6	yes	40	4.75	no	•	3
8	3	8	yes	13	5	no	•	2
9	10	15	yes	15	4.75	no	•	0
10	4	0	yes	13	4	no	•	1.5
11	53	25	no	0	•	no	•	1.5
12	47	3	yes	16	4.7	no	•	.25
13	10	10	yes	25	5.5	no	•	1.5

Abb. 1: Datentabelle mit einer Teilmenge der Daten über 84 Schülerinnen und Schüler.

Aussagefähige Graphiken

Strunk und White (1972, S. x) beschreiben einen guten Schriftsteller als jemand, der „jedes Wort aussagefähig macht". Cleveland (1993, S. 1) fordert ähnliches für die Visualisierung von Daten: „Wir lassen uns gern zu der Vorstellung verleiten, daß wir relevante Informationen aufnehmen, wenn wir viel sehen. Der Erfolg einer Visualisierung jedoch sollte einzig und allein danach bemessen werden, wieviel wir über das jeweils untersuchte Phänomen erfahren." Da wir bei einem Unterrichtswerkzeug die Anzahl der Graphiken begrenzen müssen, ist es besonders wichtig, daß die ausgesuchten Graphiken solche sind, die „etwas aussagen". Derartige Graphiken sind nicht vollgepackt mit irrelevanten Angaben – was Tufte (1983) als „Graphikmüll" bezeichnet hat – sondern in der Lage, kritische Merkmale und Beziehungen der Daten aufzuzeigen. Zur Veranschaulichung nicht-numerischer Daten verwenden wir Häufigkeitstabellen und Säulendiagramme, für numerische Daten Histogramme, Boxplots und Streudiagramme. Für diese entschieden wir uns nicht nur wegen ihrer Leistungsfähigkeit, sondern auch, weil sie so häufig gebraucht werden. Dies brachte uns in gewisse Konflikte. Tortendiagramme nahmen wir nicht auf, obwohl sie in den Massenmedien vielleicht die meistgebrauchten Diagramme sind. Sie sind nicht sehr geeignet, die relevanten Muster in den Daten aufzuzeigen oder Vergleiche zu ziehen. Um Tufte (1983, S. 178) zu zitieren, gibt es nur eins, was schlimmer ist als ein Tortendiagramm: mehrere davon. Und obwohl Boxplots praktisch außerhalb des Felds der Datenanalyse nie gesichtet werden, bezogen wir sie ein, weil sie das Zentrum und die Streuung klar darstellen und für den Vergleich von Datensätzen besonders leistungsfähig sind.

Um vorzuführen, wie die Schülerinnen und Schüler die Software benutzen, werde ich Daten aus einem Fragebogen heranziehen, der 84 Oberschülerinnen und Oberschülern in zwei Städten im westlichen Massachusetts im Jahre 1990 vorgelegt wurde: in der College-Kleinstadt Amherst und in der größeren Industriestadt Holyoke. Dieser anonymisierte Datensatz enthält Angaben über jeden der 84 Schülerinnen und Schüler, darunter über Geschlecht, Alter, Familiengröße, Familienstand der Eltern, religiöse Betätigung, Schulnoten, Bildungsniveau der Eltern, abendliches Ausgehlimit und eine Anzahl anderer Daten, die zum Teil unten beschrieben sind. In DataScope werden Daten in einer Tabelle mit Zeilen verzeichnet, die Angaben über die einzelnen Fälle enthalten (in diesem Fall Schülerinnen und Schüler) und Spalten, die bestimmte Variable repräsentieren. In der Datentabelle in Abb. 1 sind die sichtbaren Variablen: $ONYOU (wieviel Dollar der Schüler/die Schülerin beim Ausfüllen des Fragebogens bei sich hatte), HWHRS (Zeitaufwand für Hausaufgaben in Stunden), JOB (ob ein Nebenjob wahrgenommen wurde oder nicht), JOBHRS (wöchentliche Arbeitsstunden für den Job), JOB$ (Stundenlohn in $), ALWNC (ob der Schüler/die Schülerin Unterhalt bezog oder nicht), ALWNC$ (Unterhalt in Dollar pro Woche), DISTSCHL (Schulweg in Meilen). Dies ist einer der Datensätze, die wir in einem Jahreskurs in der Holyoke Highschool benutzten. Zu den Fragen, die Schülerinnen und Schüler unter Verwendung dieser Daten verfolgen können, gehören:

– Gilt für Mädchen eher eine festere Zeit des Nachhausekommens als für Jungen?
– Hängen die Ansichten einer Person über Abtreibung mit religiösen Überzeugungen zusammen?
– Prognostiziert die Geburtenposition (Altersrangplatz unter den Geschwistern) in der Familie Führungsqualitäten?
– Haben Kinder von Alleinerziehenden schlechtere Noten als Kinder mit zwei Eltern in der Familie?
– Besteht eine Relation zwischen Geschlecht und Stundenlohn einer Schülerin/eines Schülers?
– Wirkt sich ein Job negativ auf die Schulleistung aus?

Über diese letztere Frage machen sich viele Eltern Sorgen, weil sie sich fragen, ob die Schulleistung ihres Kindes bei Übernahme eines Jobs leiden wird. Zwar scheint diese Frage nur zur Betrachtung weniger Variablen zu führen, doch werden weitere Variablen des Datensatzes rasch in die Untersuchung einbezogen. Dies ist einer der Vorteile der Verwendung von Datensätzen mit vielen Variablen: Sie fordern die Schülerinnen und Schüler auf, Mutmaßungen über mögliche Erklärungen für von ihnen in den Daten beobachtete Trends anzustellen und zu überprüfen.

Histogramme enthüllen die Textur der Daten

Beginnen wir eine Untersuchung der letzten Frage mit der Betrachtung des Histo-

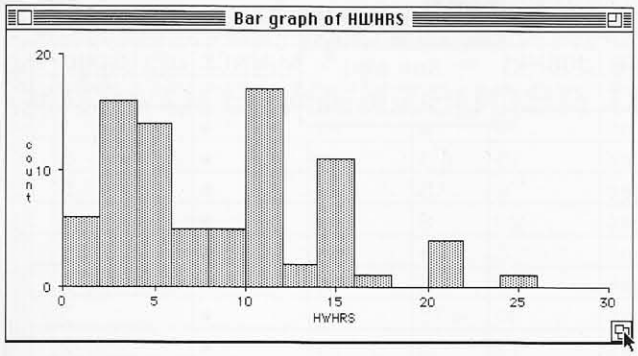


Abb. 2: Histogramm über Hausaufgabenstunden pro Woche.

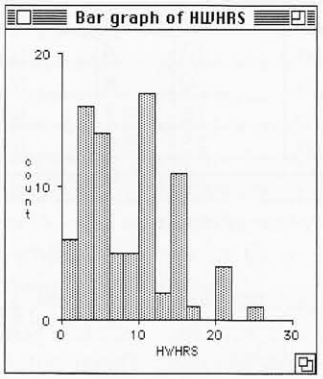


Abb. 3: Histogramm über Hausaufgabenstunden, maßstabsverändert durch Stauchung der horizontalen Achse.

gramms, mit dem dargestellt wird, wieviel Stunden pro Woche die Schülerinnen und Schüler für Hausaufgaben aufwenden. Ein Histogramm wird schlicht dadurch erzielt, daß man eine Spalte der Datentabelle, in diesem Fall HWHRS, durch Anklicken als Variable V1 bezeichnet und aus dem Menü Plot den Bargraph wie in Abb. 1 auswählt. Das Histogramm (Abb. 2) erscheint auf dem Bildschirm, ohne daß der Benutzer/die Benutzerin zunächst verschiedene Parameter angeben muß, weil wir der Ansicht sind, daß die Schülerinnen und Schüler nicht zur Entscheidung über Fragen aufgefordert werden sollten, deren Auswirkungen sie nicht bereits kennen. Das Programm beinhaltet Vorentscheidungen darüber, wie die Daten zunächst am besten dargestellt werden. Nachdem sie dargestellt worden sind, ist der Schüler/die Schülerin sehr gut in der Lage zu entscheiden, ob und wie die Graphik verändert werden sollte, um mehr über die Daten auszusagen. Ein Histogramm kann auf zwei allgemeine Arten modifiziert werden: Durch Größenveränderung mittels der „resize box" unten rechts (vergleiche Pfeil in Abb. 2) oder durch Veränderung der „interval widths" (Intervallbreiten). Die Option resize verändert den Maßstab der Graphik, je nach dem Rechteck, das für das Fenster ausgewählt wird. Die Breitenverkürzung des Histogramm, wie in Abb. 3, verstärkt die Spitzigkeit der Säulen.

Eine Frage, die bei der Betrachtung dieses Diagramms auftaucht, bezieht sich auf die Erklärung der Spitzen, besonders bei 10, 15 und 20 Stunden. Nach kurzem Nachdenken scheint es einigermaßen gesichert, daß diese Spitzen nichts über die tatsächlichen Hausaufgabenstunden aussagen, sondern vielmehr Neigungen widerspiegeln, Schätzungen auf ein Vielfaches von 5 zu runden. Unter diesem Aspekt ist es sehr aufschlußreich, sich die Verteilung von HWHRS und JOBHRS zusammen auf derselben Achse anzusehen (Abb. 4). Man beachte, daß die Spitzen in der Verteilung von JOBHRS an den gleichen Stellen auftreten. Bei JOBHRS ist das vermutlich nicht auf Rundung zurückzuführen, sondern auf die jeweilige Stundenvereinbarung am Arbeitsplatz (zum Beispiel 20-Stunden-Stelle).

Die Veränderung der Intervallbreite ermöglicht der Benutzerin/dem Benutzer eine Betrachtung, wie sich die Gestalt der Verteilung bei feinerer oder gröberer Gruppierung verändert. Die Spitzen in den Daten HWHRS verschwinden in Abb. 5, weil die Intervallbreite von 2 auf 5 verändert worden ist, um den allgemeineren Trend in der Verteilung aufzuzeigen. Von 0 bis unter 5 Stunden für Hausaufgaben zu längerem Hausaufgabenaufwand fortschreitend, finden wir immer weniger Schülerinnen/Schüler dargestellt.

Die Veränderung der Intervallbreite auf 1 zeigt sogar noch mehr Details auf (Abb. 6) und läßt vermuten, daß hier eventuell zwei Verteilungen vorliegen könnten. Vielleicht gibt es eine Verteilung für Schülerinnen und Schüler von Holyoke, die ihre Spitze bei etwa 5 hat und dann abflacht, und eine andere Verteilung für Amherst, die ihre Spitze bei 10 hat und sodann abnimmt. Wenn dies zuträfe, würde es zu lokalen Stereotypen über die Schülerinnen und Schüler der beiden Schulen passen.

Dies ist die Art Mutmaßung, auf die wir bei den Schülerinnen und Schüler hoffen, wenn sie die graphischen Darstellungen der Daten untersuchen. Es ist einer der Gründe dafür, daß Schülerinnen und Schüler Daten analysieren sollten, über die sie bereits etwas wissen. Hintergrundwissen liefert die Grundlage für interessante Fragestellungen und Ergebnisinterpretationen. Und die Software sollte die Untersuchung solcher Hypothesen erleichtern. In DataScope ist es einfach, eine Variable als „gruppierende Variable" zu kennzeichnen. Gemäß dieser Variable wird der Datensatz in Teilgruppen zergliedert. Auf diese Weise können sie Beziehungen zwischen Variablen erforschen, indem sie erkennen, wie die Verteilung einer Variablen quer über die verschiedenen Niveaus einer anderen Variablen erscheinen. Zur Erzeugung der Histogramme in Abb. 7 beläßt man HWHRS als Hauptvariable (V1) und wählt SCHOOL als gruppierende Variable (G1). Damit wurden zwei Histogramme von HWHRS, eine für Schülerinnen und Schüler von Holyoke und eine für Schülerinnen und Schüler von Amherst. (Man beachte, daß die Verteilungen mit den Niveaus der gruppierenden Variablen und nicht mit der Hauptvariable HWHRS bezeichnet sind.)

Die beiden Teilverteilungen für Hausaufgabenstunden werden übereinander dargestellt, um den visuellen Vergleich zu erleichtern. Weil für die beiden Schulen unterschiedliche Schülerzahlen erfaßt wurden, sind auf der y-Achse die relativen Häufigkeiten (proportions) zur Darstellung gebracht worden. Die Anzahl n in jeder Teilgruppe wird rechts von den Histogrammen ausgewiesen. Die Option eines Wechsels von absoluten auf relative Häufigkeiten erleichtert den visuellen Ver-

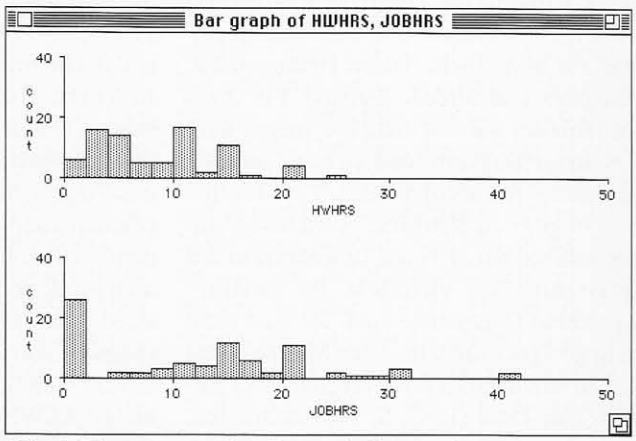


Abb. 4: Histogramme über Hausaufgabenstunden und Jobarbeitsstunden, zum Vergleich über ähnlichen Achsen in demselben Fenster dargestellt.

gleich der beiden Verteilungen, die nur leicht verschieden scheinen. Schlecht erkennen kann man bei den Histogrammen, wo bei den Verteilungen das „Zentrum", der Mittelwert, liegt.

Der Vergleich von Verteilungen mit Boxplots

Abb. 8 stellt Boxplots derselben Daten wie in Abb. 7 dar. Diese zeigen einen Median von 10 für die Schülerinnen und Schüler von Amherst gegenüber von 6 für die Schülerinnen und Schüler von Holyoke. Der Median wird markiert durch den Querstrich innerhalb der Box. Die Enden der Box markieren die beiden Quartile. Die „Antennen" werden bis zum Minimum bzw. Maximum innerhalb der „Zäune" (Quartilwert ± 1,5 · Quartilabstand) gezeichnet. Werte außerhalb werden getrennt aufgeführt, beispielsweise ist der kleine Punkt über der 25 in der Box von Holyoke ein „Ausreißer", ein Wert, der so weit außerhalb des Hauptteils der Daten liegt, so daß er einer besonderen Betrachtung bedarf (siehe den Artikel von Biehler in diesem Heft). Man beachte, wie leicht diese beiden Boxplots verglichen werden können, wenn sie auf einer gemeinsamen Achse übereinander angeordnet liegen.

Was ich bisher getan habe, könnte als erste Untersuchung betrachtet werden, um vertraut zu werden mit der „Textur" einzelner Variablen, bevor man die Frage nach der Beziehung zwischen ihnen angeht. Kommen wir zurück auf unsere Frage und präzisieren sie etwas: „Verwenden Schülerinnen und Schüler mit Job weniger Zeit auf Hausaufgaben als Schülerinnen und Schüler ohne?" Abb. 9 ist ein Boxplot von HWHRS, gruppiert nach der Variablen Job. Überraschenderweise haben die 56 Schülerinnen und Schüler mit Job („yes") einen höheren Median für Hausaufgabenstunden als die 26 Schülerinnen und Schüler ohne Arbeit („no").

Geschichten erzählen statt Berichte abgeben

Man könnte versucht sein, hier innezuhalten und zu glauben, wir hätten die Frage beantwortet. Doch gerade die Aufklärung des überraschenden Ergebnisses kann nun die weitere Analyse antreiben. Es gibt eine große Anzahl möglicher Erklärungen für

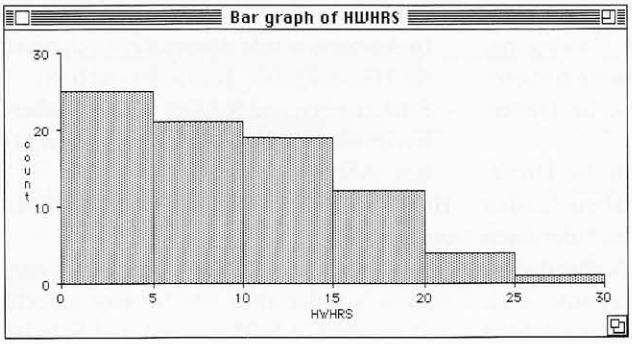


Abb. 5: Histogramm der Hausaufgabenstunden mit einer Intervallbreite 5.

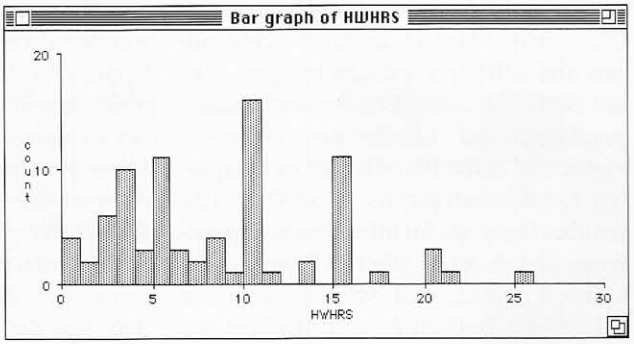


Abb. 6: Histogramm der Hausaufgabenstunden mit einer Intervallbreite von 1.

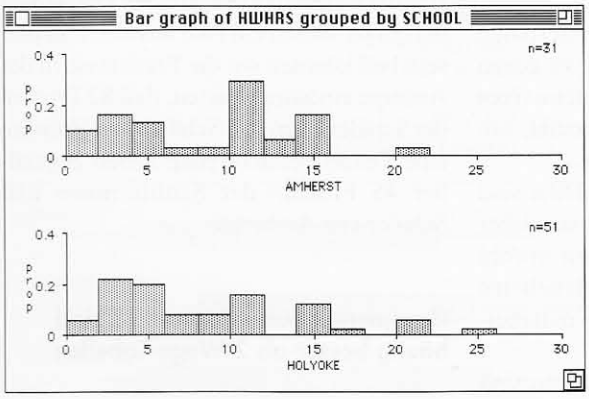


Abb. 7: Getrennte Histogramme für HWHRS (Hausaufgabenzeit) für jedes Niveau der Variable SCHOOL: Amherst und Holyoke.

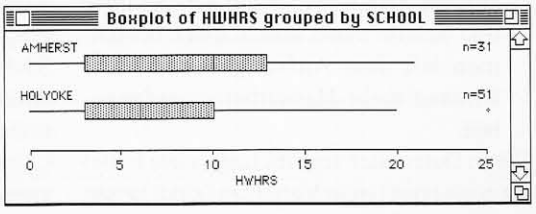


Abb. 8: Getrennte Boxplots von HWHRS für jedes Niveau der Variable SCHOOL: Amherst und Holyoke.

den beobachteten Unterschied und viele mögliche Interpretationen. Zunächst könnten wir einen zufälligen Unterschied vermuten, der sich beim Ziehen der Stichprobe ergeben hat. Immer wieder einmal (tatsächlich etwa 2 mal pro tausend) kann man beim Pokerspiel von fünf einen „Flush" aufeinanderfolgender Karten einer Farbe erhalten, auch wenn das Kartenspiel gut gemischt worden ist. Wenn dieser Flush Pik ist, könnte jemand kommen, der sich mit Pokerkarten nicht auskennt, das Spiel in der Hand ansehen und darauf schließen, daß alle Karten im Stapel Pik sind. Man könnte nachweisen, daß dies falsch ist, indem man der Person den ganzen Stapel zeigt und erklärt, daß sich eine solche Zusammenstellung nicht sehr oft ergibt, zumindest nicht unter ehrlichen Spielerinnen und Spielern. In ähnlicher Weise könnten wir bei Untersuchung der gesamten Schülerschaft jeder Schule vielleicht Schülerinnen und Schüler finden, die gleiche Zeitumfänge in Hausarbeit investieren, ungeachtet, ob sie nun in Holyoke oder in Amherst wohnen oder ob sie arbeiten oder nicht. Die Frage lautet, wie schwer es ist, eine Stichprobe aus dem „Kartenstapel" Schule zu ziehen und ein Ergebnis wie das obige zu erhalten. Diese Frage wird mit Beurteilender Statistik angegangen und mit DataScope durch Randomisierung behandelt (vgl. Konold, 1994; Noreen, 1989; Engel, 1987). Wir verfolgen die „Erklä-

rung durch Zufalls" hier nicht weiter, doch handelt es sich eine wichtige Überlegung, die Schülerinnen und Schüler höherer Klassen bei der Datenanalyse im Hinterkopf behalten sollten.

Nehmen wir einmal an, daß die Unterschiede bei den Hausaufgabenstunden tatsächlich typisch für alle Schülerinnen und Schüler in Holyoke und Amherst sind. Für diese Unterschiede gibt es immer noch eine Vielzahl möglicher Erklärungen. Eine der Herausforderungen beim Unterricht in Datenanalyse besteht darin, die Schülerinnen und Schüler soweit zu bringen, über das Berichten „eines Ergebnisses" hinauszugelangen und „Geschichten" zu produzieren und deren Plausibilität zu überprüfen. Letztendlich kommt es auf Geschichten über Daten an. Im folgenden einige einfache Geschichten, die das Ergebnis erklären könnten, daß Schülerinnen und Schüler mit Teilzeitjobs auch fleißiger studieren:

- In Amherst haben mehr Schülerinnen und Schüler vor, aufs College zu gehen und sind daher sowohl um intensiveres Lernen als auch um Jobs bemüht, um Geld fürs College anzusparen.
- Manche Schülerinnen und Schüler sind motivierter als andere und können daher mit höherer Wahrscheinlichkeit sowohl eine Teilzeitbeschäftigung wahrnehmen als auch fleißig in der Schule mitarbeiten.
- Schülerinnen und Schüler mit Jobs sind älter als Schülerinnen und Schüler ohne und Schülerinnen und Schüler bekommen mit dem Aufsteigen in höhere Klassen mehr Hausarbeiten aufgegeben.

Wenn Datensätze sowohl lang (viele Fälle) als auch breit (viele Variablen) sind, lassen sich von Schülerinnen und Schülern formulierte Erklärungen häufig „überprüfen". Da der Datensatz in diesem Fall das Alter der Schülerinnen und Schüler, die Schule, Collegepläne sowie die Selbsteinschätzung zur Motivation enthält, können Schülerinnen und Schüler die oben geschilderten Möglichkeiten explorieren. Aus Raumgründen können hier nicht alle diese Erklärungen dargestellt werden, und tatsächlich ist keine der oben gegebenen Erklärungen durch die Daten gestützt. Doch möchte ich einer der Möglichkeiten nachgehen (daß die Schülerinnen und Schüler von Amherst mehr lernen als auch mehr arbeiten als die von Holyoke) um hervorzuheben, daß die Datenanalyse kein linearer und geradlinig verlaufender Prozeß ist.

Wir wissen

- In Amherst wurde durchschnittlich mehr für Hausaufgaben gearbeitet (Abb. 8).
- Schülerinnen und Schüler mit Jobs arbeiten im Durchschnitt mehr für Hausaufgaben (Abb. 9).

Haben also die Amherster auch häufiger einen Job?

Beim Explorieren dieser Möglichkeit, entdecken Schülerinnen und Schüler jedoch, daß gerade die Schülerinnen und Schüler von Holyoke häufiger arbeiten. Das geht aus der Zwei-Wege-Tabelle in Abb. 10 hervor. Zwei-Wege-Tabellen sind die am häufigsten verwendete Veranschaulichung der Exploration von Abhängigkeiten zwischen zwei nicht-numerischen Variablen, doch ist das Erlernen ihrer Interpretation für Schülerinnen und Schüler und Anfänger bekanntermaßen schwierig. Es müssen fast zu viele Zahlen im Auge behalten werden, was der Aussage widerspricht, daß sich die Beziehung beschreiben läßt, indem man unter diesen ein Paar auswählt. In diesem Fall könnten wir die Ergebnisse in der Aussage zusammenfassen, daß 82 Prozent der Schülerinnen und Schüler von Holyoke eine Teilzeitbeschäftigung haben gegenüber 45 Prozent der Schülerinnen und Schüler von Amherst.

Gruppierte Säulendiagramme sind häufig besser als 2-Wege-Tabellen

Abb. 11 zeigt dieselben Daten in einem gruppierten Säulendiagramm. Das obere Säulendiagramm gibt den Anteil von Schülerinnen und Schülern aus Amherst, die nicht arbeiten („no"), und die arbeiten („yes"). Aus dem gruppierten Säulendiagramm ist viel leichter zu ersehen als aus der 2-Wege-Tabelle, daß es in der Häufigkeit von Arbeitsverhältnissen zwischen den beiden Schulen einen deutlichen Unterschied gibt. Wenn man präzise Werte haben will, braucht man die 2-Wege-Tabelle. Wenn man aber erkennen will, ob eine Beziehung zwischen zwei nicht-numerischen Variablen vorliegt, muß man sich das gruppierte Säulendiagramm ansehen.

Kombinieren wir das Ergebnis von Abb. 11 mit Abb. 8 und ignorieren Abb. 9 einmal, so könnte uns das sogar zu dem Schluß verleiten, daß Teilzeitarbeit sich negativ auf die Lernzeit auswirkt: Die Schülerinnen und Schüler von Holyoke haben zweimal so häufig eine Teilzeitbeschäftigung wie die Schülerinnen und Schüler von Amherst, und da wir außerdem wissen, daß die Schülerinnen und Schüler von Amherst im Durchschnitt vier Stunden mehr pro Woche lernen als die von Holyoke (vergleiche Abb. 8), könnte man daraus schließen, daß Schülerinnen und Schüler mit mehr Teilzeitarbeit weniger Zeit für Hausaufgaben haben als Schülerinnen und Schüler ohne Arbeit. Um den Widerspruch aufzuklären, darf man die 3 Variablen nicht immer nur paarweise betrachten. In Abb. 12 werden die Hausaufgabenstunden getrennt für Schülerinnen und Schüler jeder Schule und innerhalb jeder Schule für solche mit und ohne Teilzeitarbeit ausgewiesen (Der obere Plot beispielsweise zeigt die 17 Schülerinnen und Schüler von Amherst ohne Arbeit.). In unseren Stichproben von jeder Schule bringen die arbeitenden Schülerinnen und Schüler mehr Zeit mit Hausarbeiten zu als die nicht arbeitenden.

Zumindest eine neue Frage hat sich aus dieser Nachprüfung ergeben: Warum haben in Holyoke mehr Schülerinnen und Schüler einen Job? Auch hier sind wiederum mehrere Erklärungen möglich: Eine geht dahin, daß die Arbeit in Holyoke höher geschätzt wird. Das gruppierte Säulendiagramm in Abb. 13 legt jedoch eine andere Erklärung nahe: In der Stichprobe von Holyoke sind die Schülerinnen und Schüler älter als in der von Amherst.

Der nächste logische Schritt in unserer Analyse wäre die Untersuchung, ob eine Relation zwischen Alter und Beschäftigung besteht (sie ist tatsächlich vorhanden). Dies soll hier nicht weiterverfolgt werden. Es sollte mit der Skizzierung einer möglichen Analyse demonstriert werden, daß umfangreiche Datensätze Schülerinnen und Schülern zwar Gelegenheit bieten, einer ganzen Menge von Fragen nachzugehen, doch daß die Ergebnisse i. d. R. komplex und häufig nicht eindeutig sind. Es kann leicht zu Enttäuschungen bei Schülerinnen und Schülern führen, deren Interesse und Wissen bezüglich der verfolgten Fragen nicht stabil ist. Ich habe erlebt, wie Schülerinnen und Schüler nach Betrachtungen von wenigen unklaren oder widersprüchlichen Ergebnissen die Hände hoben und behaupteten, sie seien an der Frage von vornherein nicht so sehr interessiert gewesen. Dies ist eine der größten Herausforderungen bei der Verwendung umfangreicher Datensätze. Der Computer kann das Problem insofern verschärfen, als er es Schülerinnen und Schülern leicht macht, eine Graphik nach der anderen zu betrachten,

ohne über eine davon ernsthaft nachzudenken, und ohne sich die Zeit zu nehmen, das Gefundene zu ordnen und zu reflektieren. Das ist einer der Bereiche, in der die Anleitung durch den Lehrenden und die kluge Auswahl von Aufgaben von entscheidender Bedeutung ist.

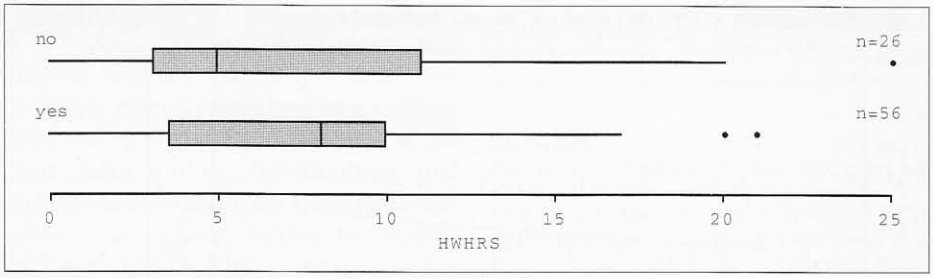


Abb. 9: Boxplot von HWHRS gruppiert nach JOB (yes/no).

Bisweilen läßt sich eine Geschichte aus einer einzigen Graphik entwickeln

Es kommt allerdings auch vor, daß sich manche Ergebnisse für eindeutige Erklärungen hergeben. Abb. 14 ist ein Streudiagramm aus einem Datensatz über 104 Länder, welcher hauptsächlich einem Almanach entnommen wurde. Die meisten Angaben stammen von 1990, damit ist dieser Datensatz bereits Geschichte. 1990 war die UdSSR noch ein statistischer Monolit, waren Ost- und Westdeutschland noch durch eine Mauer getrennt. (Eines der Probleme beim Aufbau großer Datensätze mit aktuellen Angaben besteht darin, daß man mit diesem Aufbau nie fertig wird.) Das Streudiagramm zeigt die Geburtenziffer (Geburten pro 1.000 Einwohnerinnen und Einwohner) auf der x-Achse und die Sterbeziffer auf der y-Achse. Hinzugefügt wurde die Regressionslinie mit Angabe ihrer Formel in der oberen linken Ecke und der Wert von r, dem Pearson'schen Korrelationskoeffizienten in der oberen rechten Ecke.

SCHOOL	JOB no	JOB yes	total
AMHERST	17 (0.55)	14 (0.45)	31
HOLYOKE	9 (0.18)	42 (0.82)	51
total	26 (0.32)	56 (0.68)	82

Abb. 10: Zwei-Wege-Tabelle von JOB (yes/no) und SCHOOL (Amherst/Holyoke).

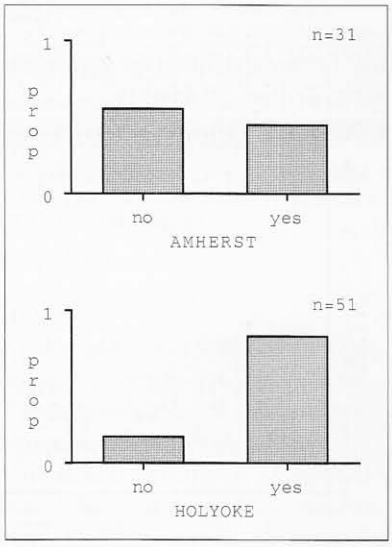


Abb. 11: Säulendiagramm von JOB (yes/no) gruppiert nach SCHOOL (Amherst/Holyoke).

In Statistikkursen, die ich als Erstsemester belegte, wurden Graphiken diesen Typs gezeigt, kurz nachdem man uns dann mit dem Gedanken bekannt gemacht hatte, durch bivariate Daten eine passende Gerade zu ziehen. Der Lehrende pflegte eine Gerade durch ein U-förmiges Streudiagramm zu ziehen und zu fragen „Was ist los?" Einer der besseren Studierenden wies dann darauf hin, daß die Daten nicht linear seien. Dann klärte uns der Lehrende auf, daß r nur die Paßgenauigkeit um die Gerade mißt und daher ein schlechter Näherungswert für die Stärke der Beziehung ist, wenn die Daten nicht linear sind. Nach gelernter Lektion ging der Lehrende dann zu einem weiteren Streudiagramm über. Wegen des Charakters des Streudiagramms wurde die Graphik üblicherweise mit ungekennzeichneten Achsen gezeigt. Die Lektion ist letztendlich eine allgemeine, die für jedes nicht-lineare Streudiagramm gilt. Die Angaben über Geburten- und Sterbeziffern in den Ländern der Welt können die Aufmerksamkeit auf andere Eigen-

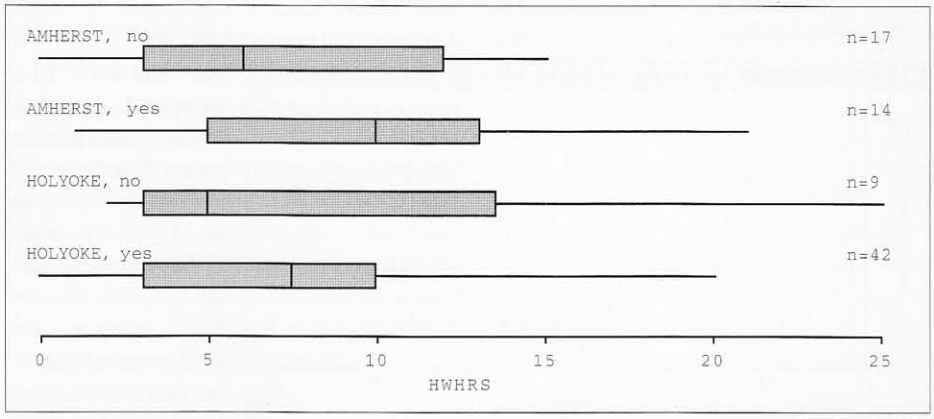


Abb. 12: Boxplots für HWHRS gruppiert nach SCHOOL (Amherst/Holyoke) und JOB (yes/no).

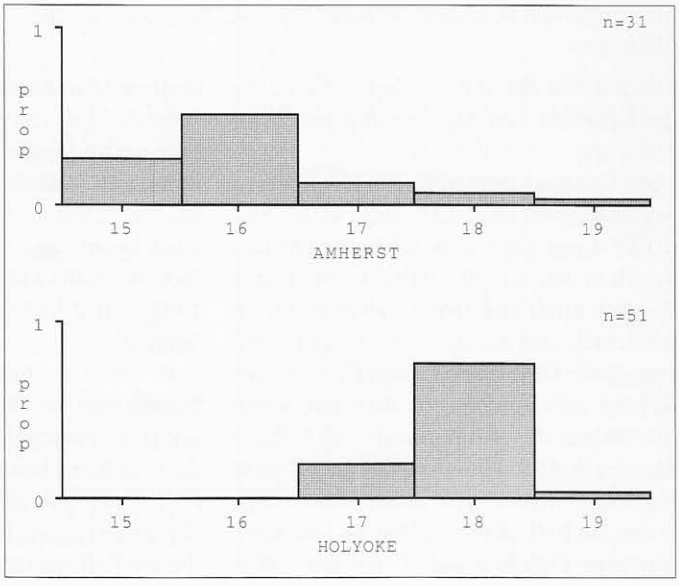


Abb. 13: Getrennte Histogramme (relative Häufigkeiten) für die Altersverteilung der Schülerinnen und Schüler.

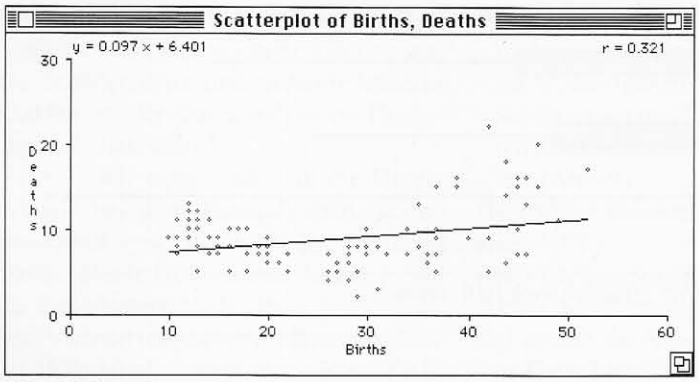


Abb. 14: Streudiagramm mit Regressionsgerade für Geburten- und Sterbeziffer von 104 Ländern.

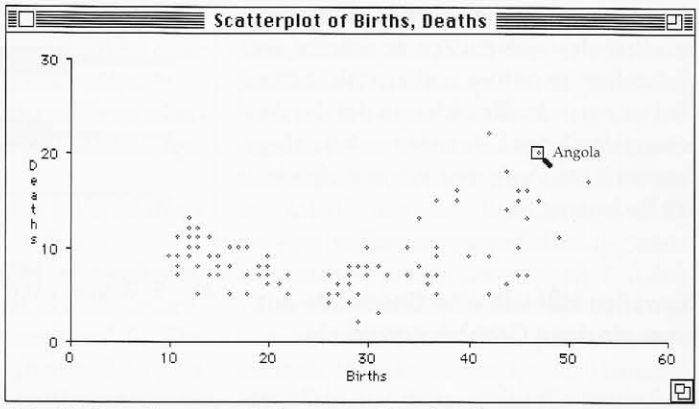


Abb. 15: Streudiagramm Geburten- und Sterbeziffer mit Identifizierung von Angola.

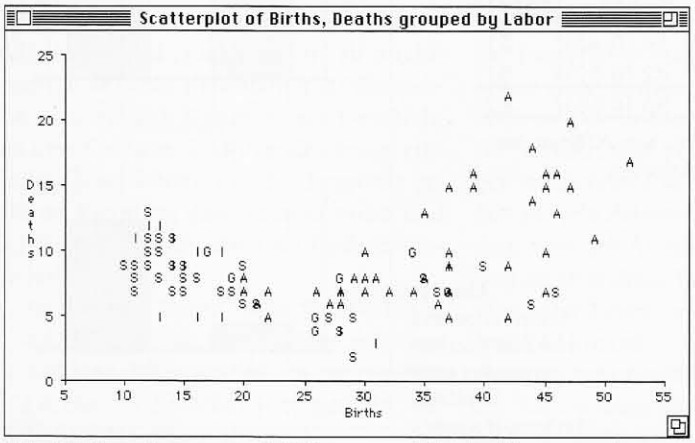


Abb. 16: Streudiagramm mit Geburten- und Sterbeziffern gruppiert nach Typ des Landes.

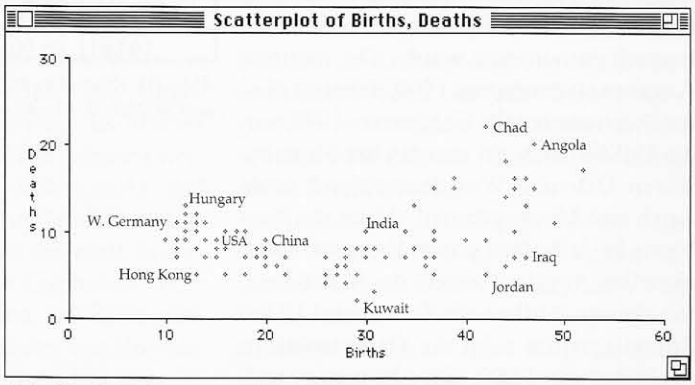


Abb. 17: Streudiagramm der Geburten- und Sterbeziffern mit Identifizierung mehrerer Länder.

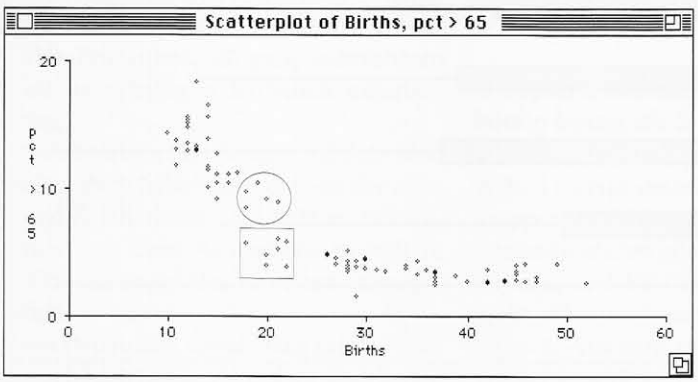


Abb. 18: Streudiagramm mit Geburtenziffer und Prozentsatz der über 65jährigen.

schaften als die statistischen lenken, die der Lehrende den Studierenden nahebringen will.

Natürlich sind Begrenzungen der linearen Regression wichtige Lektionen in der Statistik. Wenn man aber Schülerinnen und Schülern zeigen will, wie sie etwas über die Welt erfahren können, indem sie Daten betrachten und darüber nachdenken, wird eine große Gelegenheit verpaßt, wenn die Achsen unbezeichnet bleiben und wenn das Wissen der Schülerinnen und Schüler über die Welt und ihre Neugier darauf nicht angezapft werden. Die interessante Frage in diesem Fall ist nicht „Was ist los, wenn man eine Gerade anpaßt?" sondern „Was geht hier vor?" Eine Möglichkeit, Schülerinnen und Schüler unter Verwendung von DataScope in die Frage einzubeziehen, besteht darin, erst den Cursor über einem der Punkte anzuklicken, um den Namen eines Landes nachzusehen (vergleiche Abb. 15). Nach einigem Nachdenken können sie eine mündliche Beschreibung des Streudiagramms geben, die eine verblüffende Frage beleuchtet. Wie kommt es, daß die Sterbeziffer eine Zeitlang zusammen mit der Geburtenziffer zurückgeht, also weitere Abnahmen der Geburtenziffern nach Erreichen von 25 pro 1.000 mit Zunahme der Sterbeziffer einhergeht?

Auf manche möglichen Erklärungen kommt man, wenn man nach der Variablen „Arbeit" gruppiert, die die Länder nach dem vorherrschenden Typ von Arbeit einteilt (Industrie (I), Landwirtschaft (A), Dienstleistungen (S), Verwaltung(G)). In diesem Fall ersetzt das Gruppierungsmerkmal jeden Punkt im Scatterplot durch den zum entsprechenden Arbeitstyp gehörenden Buchstaben (Abb. 16). Zivilisationsmüde könnten aufgrund dieser Daten folgern, daß mit der Entwicklung von landwirtschaftlich geprägten zu Dienstleistungsvolkswirtschaften die Sterberate in zunehmendem Stress steigt. Ein Blick auf die Namen anderer Länder in der Graphik ermöglicht den Schülerinnen und Schülern einige mögliche Erklärungen zu formulieren (Abb. 17). Eine davon: Länder, denen es gelungen ist, über etliche Jahre eine niedrige Geburtenziffer zu halten, weisen eine ältere Bevölkerung auf. Daher sind die Sterbeziffern in diesen Ländern höher als in Ländern wie China, die erst in jüngster Zeit ihre Geburtenziffer gesenkt haben, aber immer noch über eine relativ junge Bevölkerung verfügen. Eine derartige Erklärung könnten wir überprüfen, wenn wir Daten über das Durchschnittsalter der Bevölkerung in jedem Land hätten, oder besser noch Daten über eine Periode von 100 Jahren, anhand deren wir für verschiedene Länder, Geburten, Todesfälle und durchschnittliches Bevölkerungsalter im Zeitverlauf verfolgen könnten (siehe hierzu auch die Beiträge von Portscheller und Kohorst in diesem Heft).

Der Scatterplot in Abb. 18 zeigt die Relation zwischen Geburtenziffer und Anteil der Bevölkerung, die über 65 Jahre alt sind. Die Graphik ist eine starke Grundlage für einen Teil unserer Erklärung – daß der Anteil der älteren Bevölkerung mit sinkender Geburtenrate zunimmt. Die Beschleunigung dieser Zunahme mit abnehmender Geburtenrate könnte dazu beitragen, die U-förmige Relation zwischen Geburten und Sterbeziffer zu erklären.

Der zweite Teil unserer Erklärung lautete, daß die Zunahme der Sterbeziffer erst sichtbar wird, wenn ein Land eine Zeitlang eine relativ niedrige Geburtenziffer gehalten hat. Leicht gestützt wird diese Argumentation durch Vergleichen der fünf Länder im Kreis (darunter Argentinien, Israel und die frühere UdSSR) mit den sieben Ländern im Rechteck (darunter China, Sri Lanka und Thailand). Die Geburtenziffern in diesen beiden Gruppen liegen im Vergleich zu dem Bevölkerungsanteil über 65 relativ eng bei einander. Im 20-Jahres-Abschnitt von 1970 bis 1990 hat jedoch die Gruppe im Kreis ihre Geburtenziffern im Durchschnitt um 5,3 pro Tausend gesenkt, während die Gruppe im Rechteck ihre Geburtenziffern im Durchschnitt um 12 gesenkt hat. Sogar dann noch sind die Geburtenziffern in diesen beiden Gruppen fast gleich, weil die Länder im Kreis ihre relativ niedrigere Geburtenziffer schon länger beibehalten haben, weil ihre Bevölkerungen im Durchschnitt älter sind und daher erwarten können, infolgedessen eine etwas erhöhte Sterbeziffer zu haben.

Zusammenfassung

Meine Absicht in diesem Artikel war, nachzuweisen, wie speziell für Unterrichtszwecke entwickelte Software zur Datenanalyse Schülerinnen und Schüler darin unterstützen kann, über rudimentäre Fertigkeiten zu den fesselnderen und schwierigeren Aufgaben vorzudringen und zu lernen, wie man einen Datensatz kritisch untersucht und aus Informationsbruchstücken Geschichten entwickelt, die diese Bruchstücke zu einem verständlichen Ganzen verknüpfen. Ich glaube gewiß nicht, daß der Computer ein Allheilmittel ist, oder daß es unwichtig wäre, Schülerinnen und Schülern verschiedene Aspekte der graphischen Darstellung und Statistik außerhalb des Computers nahe zu bringen. Es gibt viele Lerntätigkeiten, die meiner Ansicht nach besser ohne den Computer stattfinden. Meine Schülerinnen und Schüler zeichnen zum Beispiel ihre ersten Boxplots unter Verwendung bescheidener Datenmengen von Hand. Wenn aber die Zeit dafür reif ist, Schülerinnen und Schüler bestimmte, für die Explorative Datenanalyse typische, Zyklen durchlaufen zu lassen, möchte ich sie vor dem Computer sitzen haben. In diesem Artikel habe ich die Probleme, die Schülerinnen und Schüler auch dann noch mit der Explorativen Datenanalyse haben, wenn sie einfache Software verwenden, nur andeuten können, und diese Probleme erscheinen mir keineswegs einfach. Unseren Weg zur Überwindung dieser Schwierigkeiten zu suchen, wird ein aufschlußreicher Teil der Geschichte von Unterrichtserfolgen werden, die wir hoffentlich in etwa zehn Jahren darüber erzählen können sollten, wie der Computer den Unterricht in Datenanalyse verändert hat.

Danksagung

Ich danke Amie Robinson für ihre kritischen Bemerkungen zu einem ersten Entwurf sowie Rolf Biehler und Günther Seib vom IDM dafür, daß sie die Übersetzung ins Deutsche vorgenommen haben. Die in diesem Artikel beschriebene Software wurde mit Unterstützung der National Science Foundation (#MDR 8954626) entwickelt und von Intellimation veröffentlicht. Ich möchte noch bemerken, daß die hier dargelegten Standpunkte meine eigene Meinung darstellen und sich nicht unbedingt mit denen der Stiftung decken.

Literatur

Biehler, R.: Cognitive technologies for statistics education: Relating the perspective of tools for learning and of tools for doing statistics. In: L. Brunelli & G. Cicchitelli (Eds.), Proceedings of the First Scientific Meeting of the International Association for Statistics Education (S. 173–190). Università di Perugia 1994.
Cleveland, W. S.: Visualizing data. Summit, NJ: Hobart Press 1993.
Engel, A.: Stochastik. Stuttgart: Klett 1987.
Gordon, F. S., and Gordon, S. P.: Statistics for the twenty-first century. MAA Notes #26 Washington, DC: Mathematical Association of America 1992.
Konold, C.: Understanding probability and statistical inference through resampling. In: L. Brunelli & G. Cicchitelli (Eds.), Proceedings of the First Scientific Meeting of the International Association for Statistics Education (S. 199–211). Università di Perugia, Italy 1994.
Noreen, W.: Computer intensive methods for testing hypotheses. New York: John Wiley & Sons 1989.
Polanyi, M.: Knowing and being. Chicago: University of Chicago Press 1969.
Strunk, W. Jr., and White, E. B.: The elements of style. New York: Macmillan 1972.
Tufte, E. R.: The visual display of quantitative information. Cheshire, Conn.: Graphics Press 1983.

Software

Data Desk 4.1. Velleman, P., Data Description Inc., P.O.Box 4555, Ithaca, NY 14852 (Macintosh)
Data Insights. Edwards, Luis A. & Keogh, K. M., Sunburst/Wings for Learning, 1600 Green Hills Road, P.O. Box 660002, Scotts Valley CA 95067-0002 (PC MS-DOS, Apple II)
DataScope. Konold, C. & Miller, C.D., Scientific Reasoning Research Institute, Hasbrouck Lab, Univ. of Massachussetts, Amherst, MA 01003, publiziert durch Intellimation Santa Barbara, CA (Macintosh)
Statistics Workshop. Rubin, A. and Bruce, B. (BBN Inc., Cambridge, MA, USA). Wings for Learning/ Sunburst, 1600 Green Hills Road P.O. Box 660002, Scotts Valley, CA 95067-0002 (Macintosh)
StatView 3.0/4.0. Feldman, D. S., jr & Gagnon, J., Abacus Concepts, Inc., 1984 Bonita Ave., Berkeley, CA 94704, USA. (Macintosh).

Zum Autor
Clifford Konold, geb. 1949, Studium der Psychologie, Promotion in Psychologie über Natur stochastischen Denkens bei Erwachsenen, derzeit Professor am Scientific Research Institute, Hasbrouck Laboratory University of Massachusetts, Amherst, MA 01003 und Co-Direktor des Mathematikzentrums am TERC, 2067 Massachusetts Avenue, Cambridge, MA 02140.

Liebe Leserin, lieber Leser,

haben Sie schon Erfahrungen mit unterrichtsgeeigneter, deutschsprachiger Software zur Datenanalyse sammeln können?
Wir würden uns freuen, wenn Sie darüber kurz in unserer Zeitschrift berichten möchten.

Ihre Beiträge senden Sie bitte an die Redaktion:

Dr. Bernhard Uher
Am Hinteranger 5
91301 Forchheim