Daten analysieren mit dem Computer: Unterstützung von Begriffsbildung und Anwendungsorientierung in der Stochastik

von Rolf Biehler

1. Einleitung

Welche neuen didaktischen Möglichkeiten ergeben sich für den Stochastikunterricht, wenn leicht zu bedienende Softwarewerkzeuge zur Verfügung stehen, mit denen man Daten flexibel und interaktiv, graphisch und numerisch analysieren kann? Diese Frage soll im Zentrum dieses Beitrags stehen.

Neue didaktische Möglichkeiten eröffnen sich zum einen für die Betonung einer praktischen, angewandten Statistik. Zum anderen eröffnet sich die Perspektive des *Lernens der Stochastik durch praktische Datenanalyse:* Reale Daten können viel unproblematischer als Beispiel- und Übungsmaterial zur Illustration und Exploration von Techniken, Strategien und Begriffen der Stochastik herangezogen werden.

Möglichkeiten und Anregungen für die Nutzung von Softwaretools sollen im folgenden anhand von zwei Beispielen dargestellt werden. Auf Formen und Bedingungen unterrichtlicher Realisierung soll dabei im Detail nicht eingegangen werden. Im Rahmen des GRAPHDAS-Projektes[1] sind hierzu aber einige Materialien entwickelt und unterrichtspraktische Erfahrungen gesammelt worden (vgl. Biehler 1990, Biehler/Steinbring 1990).

Professionelle Statistiksoftware hat erhebliche Fortschritte gemacht, was die Steigerung der Benutzerfreundlichkeit bei gleichzeitiger Aufrechterhaltung oder sogar Steigerung einer komplexen Funktionalität anbetrifft. Bestimmte neuere Werkzeuge scheinen deshalb unter bestimmten Bedingungen und ggf. nach gewissen Adaptationen durchaus für den Einsatz in der Schule geeignet zu sein.[2] Für schulische Zwecke sind bereits Datenanalysesysteme mit sehr begrenztem Methodenrepertoire und mit Einschränkung auf spezielle Daten entwickelt worden.[3] Gegenüber diesen eher geschlossen zu nennenden Arbeitsumgebungen zeichnen sich außerschulische Datenanalysesysteme durch ihre *Offenheit und Allgemeinheit* aus. Die größere Offenheit beruht vor allem auf den folgenden allgemeinen Funktionen:

- *Dateneingabemöglichkeiten* für viele inhaltlichen Bereiche, Typen und Strukturen von Daten;
- *Datenbankfunktionen*, d. h. auf vielfältigen Möglichkeiten zur Verknüpfung und Umorganisierung von Daten;
- *Graphik*, d. h. auf Möglichkeiten zur Gestaltung und interaktiven Veränderung von Graphiken für die Rohdaten und für Ergebnisse der Datenanalyse;
- *algebraisch-numerische Verarbeitung*, d. h. auf Möglichkeiten, Daten numerisch zu transformieren, numerisch zu aggregieren, selbst entwickelte Algorithmen auf sie anzuwenden oder in das System einzubinden;
- *Simulation und Modellbildung*, d. h. auf Möglichkeiten zur Simulation und Modellierung von Zufallsprozessen, deren Ergebnisse man im Zusammenhang und Kontrast zu realen Daten heranziehen kann;
- *Bibliothek statistischer Verfahren* mit großem Variantenreichtum einzelner Verfahren.

Sonderdruck aus: Der Mathematikunterricht 36 (1990)
Heft 6, 50 - 71

Die Offenheit wird verstärkt, wenn diese verschiedenen Funktionen so zu einem System organisiert sind, daß man sie jederzeit möglichst frei miteinander kombinieren kann. Sind einzelne elementare Methoden und Darstellungsweisen in eine solche Systemumgebung eingebettet, so verstärken sich ihre datenanalytischen Möglichkeiten außerordentlich.

Die größere Offenheit ist Problem und Chance zugleich, insofern sie den NutzerInnen mehr Freiheit und Verantwortung bringt, sie in die Rolle derjenigen versetzt, die statistische Graphik mitgestalten, die Daten umorganisieren, die selber unkonventionelle Methoden entwickeln und anwenden statt, mehr oder weniger passiv, aus vorgegebenen Verfahren auszuwählen. In diesem Zusammenhang hat sich ein ganz neuer Stil der *interaktiven und explorativen Datenanalyse* entwickelt (vgl. Tukey 1977, Biehler 1982, Borovcnik/Ossimitz 1987, Biehler/Rach 1990). Dies stellt neue inhaltliche Anforderungen an die Kompetenz derjenigen, die Daten analysieren. Hinzu kommt noch die Notwendigkeit zu lernen, *wie* man entwickelte Zielvorstellungen in einer konkreten Software umsetzt: Die Umständlichkeit der Bedienung kann dabei sehr hemmend sein. Oder umgekehrt: Wohlgestaltete Benutzerschnittstellen können statistische Denkprozesse sogar konkretisieren und stützen.

Die Frage angemessener Schnittstellengestaltung und Systemarchitekturen für schulische Anwendungen soll im folgenden ausgeklammert werden (vgl. hierzu aber Biehler/Rach 1990). Wir gehen im folgenden davon aus, daß ein System mit den o. g. Grundfunktionen zur Verfügung steht. Bei den ausführlich dargestellten Beispielen werden diese Funktionen weiter konkretisiert und in ihrer Bedeutung deutlich werden. Die Beispiele sollen dabei nicht nur für sich selbst stehen, sondern allgemeine Möglichkeiten exemplifizieren.

2. Interaktive graphische Datenanalyse: Das Wetter in Norderney und Bamberg 1967

2.1 Einführung

Es soll ein Datensatz aus DIFF (1982) re-analysiert werden, der dort unter dem Stichwort »Kennzahlen von Häufigkeitsverteilungen« behandelt wird. Es handelt sich um Daten zum Klima in Norderney und Bamberg, und zwar sind in DIFF (1982) für jeden Tag des Jahres 1967 angegeben:

- die mittlere Tagestemperatur in °C (Mittelwert der Temperatur um 7.00 h, 14.00 h und 21.00 h (doppelt)). Für jeden Tag t_i sollen die Temperaturen mit bam_i und nor_i bezeichnet werden;
- die tägliche Niederschlagsmenge in mm.

Hiermit hat man eine Tabelle bzw. Matrix mit 365 Zeilen und 6 Spalten:

Nr. des Tages	Monat	Bamberg Temperatur	Norderney Temperatur	Bamberg Niederschlag	Norderney Niederschlag

Dies ist die dominante »Datenstruktur« für die Statistik: für bestimmte Objekte (Meßeinheiten), hier Tage des Jahres 1967, sind die Werte bestimmter Variablen gemessen worden, hier Temperatur und Niederschlag in Bamberg und Norderney.

Ein integriertes interaktives Softwaresystem, mit dem man die Daten in dieser und weitergehender Weise analysieren möchte, müßte etwa die folgenden konkreten Funktionen einschließen[4]:

Datenbankfunktionen
- Zerlegen nach Monaten, Jahreszeiten oder anderen vollständigen Einteilungen eines Jahres (Zerlegung von bam_i in z. B. 12 Teildatensätze für die 12 Monate);
- Auswählen von Teilmengen (einzelnen Monaten oder Perioden) des Jahres;
- Weiterverarbeitung von Zwischenergebnissen als neue Daten, z. B. die Monatsmittelwerte.

Statistische Graphik
- Typen: Histogramm, Streudiagramm, Liniendiagramm, Boxplot, Mittelwert-plus-Streuung-Diagramm[5];
- Operationen: nebeneinanderstellen, überlagern, dekomponieren, verbinden, reskalieren.

Statistische Funktionen (Statistische Zusammenfassungen)
- arithmetisches Mittel, Median und andere Prozentwerte, Standardabweichung, Quartilabstand;
- multiple Zusammenfassungen, wandernde Zusammenfassungen (für den Vergleich mehrerer Datensätze).

Algebraisch-numerische Verarbeitung
- Differenzen und Quotienten z. B. $bam_i - nor_i$, um die beiden Städte besser vergleichen zu können.

Mit diesem System von Funktionen ist gewissermaßen ein Rahmen abgesteckt, in dem man die Daten analysieren kann. Je nach konkreter inhaltlicher Fragestellung und in Abhängigkeit von den Zwischenergebnissen der Datenanalyse wird man sich unterschiedlich in solch einem Rahmen bewegen oder ihn auch überschreiten wollen.

Den *Prozeß* von der offenen Frage über verschiedene Entdeckungen in den Daten bis hin zur kritischen Zusammenfassung von Ergebnissen stärker im Unterricht zur Geltung zu bringen, wird mit einem integrierten Softwaresystem erleichtert. Wichtige Voraussetzung dafür ist die nicht immer ganz einfache Entwicklung geeigneter Aufgabenstellungen oder »Projektthemen« für Datenanalysen, die die Untersuchung anleiten können und interessante, spannende Fragen aufwerfen. Eine wichtige kognitive Herausforderung besteht dann darin, die Fragestellungen und Formulierungen in statistische Fragestellungen zu transformieren, die mit einem Softwarewerkzeug bearbeitbar sind, und die Problemlösung im Hinblick auf die Ausgangsprobleme kritisch zu bewerten.

Für die Temperaturdaten lassen sich z. B. die folgenden Fragestellungen formulieren:

1. Die mittlere Jahrestemperatur (Median) beträgt in Norderney 9 °C, in Bamberg 9,3 °C. Dies scheint ein geringfügiger Unterschied zu sein. Worin unterscheidet sich das Wetter nun überhaupt in den beiden Städten? Wo gab es mehr »kalte« oder mehr »warme« Tage? Man arbeite dazu Unterschiede und Gemeinsamkeiten der beiden Temperaturverteilungen heraus.
2. Man beschreibe den Einfluß der Jahreszeit auf die Temperaturen. Welche Gemeinsamkeiten und Unterschiede bestehen zwischen Norderney und Bamberg? Liefern die kalendarischen Jahreszeiten eine gute Einteilung des Jahres in Wetterperioden oder legen die Daten eine andere Einteilung nahe?
3. Bamberg besitzt eher ein Kontinentalklima, Norderney ein »ausgeglicheneres« Seeklima. In welcher Weise läßt sich diese Behauptung anhand der Daten bestätigen und präzisieren?
4. Wegen ihrer aus weltweiter Perspektive gesehenen geographischen Nähe werden Norderney und Bamberg auch ähnlichen Klimafaktoren ausgesetzt sein. Wieweit läßt sich dies an den Daten bestätigen und präzisieren? Läßt sich eine »gemeinsame Variation« der Daten feststellen?
5. Wenn man eine Münze wirft, kann man aus zurückliegenden Wurfergebnissen wegen der Unabhängigkeit nichts für die Zukunft ableiten: »Die Münze hat kein Gedächtnis«. – Das Wetter hat es schon! Wie gut könnte man also allein aus der Kenntnis der Temperatur von heute Vorhersagen für die Temperatur morgen, in einer Woche, in einem Monat ableiten? Man untersuche daraufhin die Daten für 1967.

Im folgenden wird jedoch nicht strikt von einer der genannten Fragen ausgegangen, sondern es geht – in Erweiterung der Darstellung in DIFF (1982), wo nur mit einer Kollektion unverbundener Programme gearbeitet werden kann – im folgenden darum,
- die interaktive Arbeitsweise mit Graphik und Daten zu demonstrieren und dabei von Datenbankfunktionen und algebraisch-numerischen Verarbeitungsmöglichkeiten Gebrauch zu machen[6];
- einige weitere Darstellungsformen und Methoden wie verbundene Histogramme (2.2), Boxplots, Mittelwert-plus-Streuung-Diagramm (2.4) und gleitende Mittelwerte (2.5) einzubeziehen;
- Zusammenfassung der Daten *als Prozeß* zu interpretieren und den Wert *graphischer* Zusammenfassung zu thematisieren;
- weitere Lernmöglichkeiten zu thematisieren, wie beim Thema Abhängigkeit (2.6) und Kennzahlenvergleich durch Anwendung auf mehrere Datensätze (2.3);
- Schwierigkeiten anzudeuten, die schon beim Umgang mit einfachen Diagrammen zu erwarten sind, insbesondere wenn es um inhaltliche Interpretation und In-Beziehung-Setzung zu anderen Diagrammen geht.

Die folgende Darstellung gliedert sich im wesentlichen nach den verwendeten statistischen Methoden und Diagrammen, um deutlich zu machen, welche Erkenntnismöglichkeiten jeweils neu hinzukommen. In diese Darstellungsstruktur werden die anderen o. g. Aspekte integriert.

2.2 Histogramme zum Verteilungsvergleich

An Frage 1 läßt sich mit dem Vergleich der Verteilungen anhand von zwei Histogrammen herangehen. Wir finden in Abbildung 1 (s. S. 54) die zwei Histogramme, wobei wir im Moment von der Schwärzung eines Teilbereiches abstrahieren wollen.

Man entdeckt u. a. als gemeinsame Hauptstruktur in beiden Histogrammen eine Bimodalität. Handelt es sich bei der Bimodalität um eine Art »Wettergesetz«, das auch für andere Jahre und Orte gilt? Zur Überprüfung wären weitere Daten heranzuziehen. Der Median liegt in beiden Fällen zwischen den Gipfeln, in Bamberg sieht man eine breitere Streuung, die höhere Konzentration in Norderney fällt auf: Ausdruck der »Ausgeglichenheit«! Dies würden zwei Graphiken mit gleicher Skalierung noch besser als Abbildung 1 zum Ausdruck bringen. In einer Computer-Umgebung stehen derartige Variationsmöglichkeiten zur Verfügung.

Sind diese Entdeckungen nun von der speziellen Graphik abhängig oder zeigen sie sich auch bei anderen Klasseneinteilungen? Dies könnte man überprüfen durch Variation des Histogramms.

Kann man Bimodalität irgendwie »erklären«? Zerfällt das Jahr vielleicht in zwei Teile mit jeweils symmetrischen Temperaturvergleichungen um einen Mittelwert? Wäre diese Jahreseinteilung für Bamberg und Norderney gleich?

Beide Histogramme sind logisch miteinander verknüpft: ein bestimmter Jahrestag, z. B. der 4. Juli 1967 ist gleichsam als eine in einer Säule versteckte Scheibe in beiden Diagrammen repräsentiert. Welche Bereiche des Bamberg-Histogramms korrespondieren z. B. mit dem »Gipfelbereich« von 12 °C – 18 °C im Norderney-Histogramm? Dies kann man herausfinden, indem man mit Hilfe der »Datenbankfunktionen« nur alle diejenigen Bamberg-Temperaturen (Tage) auswählt und im Histogramm darstellt, für die gilt $12\,°C \leq$ Norderney-Temperatur $< 18\,°C$. Eine direktere Methode besteht in dem Konzept der *multiplen verbundenen Repräsentationen*, das z. B. in der Software DATADESK implementiert ist: Wenn

man im Norderney-Histogramm einen bestimmten Bereich markiert, so wird der zugehörige Bereich im Bamberg-Histogramm praktisch gleichzeitig auch auf dem Bildschirm markiert (s. Abb. 1).

Man erkennt in Abbildung 1 die Korrespondenz der beiden Gipfel. Welcher Jahreszeit entsprechen die Gipfel? Wir vermuten, daß hier vor allem Tage aus dem Sommer vorliegen. In DATADESK kann man jetzt zur Überprüfung eine weitere Graphik öffnen mit der Darstellung der Daten bam_i und nor_i gegen die Zeit t_i, wo die zugehörigen Punkte wieder extra markiert sind.

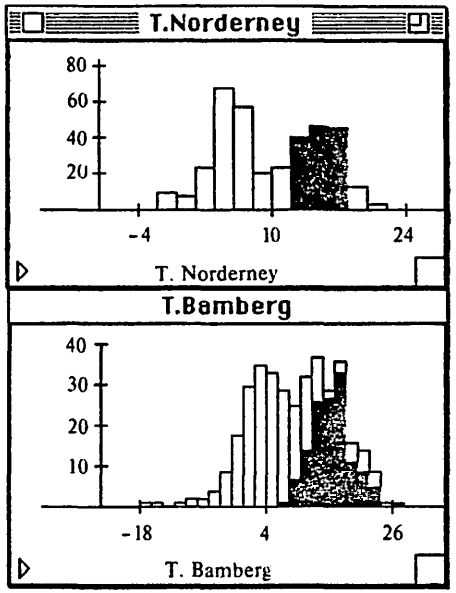

Abb. 1: Verbundene Histogramme für Tagestemperaturen in Bamberg und Norderney 1967

2.3 Mittelwertkurven für die jahreszeitliche Entwicklung

Auf diese Weise kann man auch zu Frage 2 geführt werden: Wie läßt sich nun der jahreszeitliche Verlauf beschreiben? Man kann dazu wie bei DIFF (1982) die Monatsmittel aus Abbildung 2 verwenden.

Man muß lernen, sich auf verschiedene Teilaspekte dieses Diagramms konzentrieren zu können, z. B. auf die beiden Kurven jeweils einzeln oder in ihrem Verlauf zusammen oder auf den Abstand zwischen beiden Kurven. Der Computer erlaubt aber auch auf einfache Weise ein »Herausgreifen« einzelner Aspekte aus der Darstellung, z. B. Darstellung nur einer Stadt. Die Aufmerksamkeitspunkte können unterschiedlichen inhaltlichen Fragestellungen korrespondieren, z. B.: Welche Gemeinsamkeiten im jahreszeitlichen Verlauf weisen die beiden Städte auf (etwa Anstieg auf Maximum im Juli), welche Unterschiede (etwa Jahreszeiten schlagen in Norderney weniger »extrem« zu Buche)?

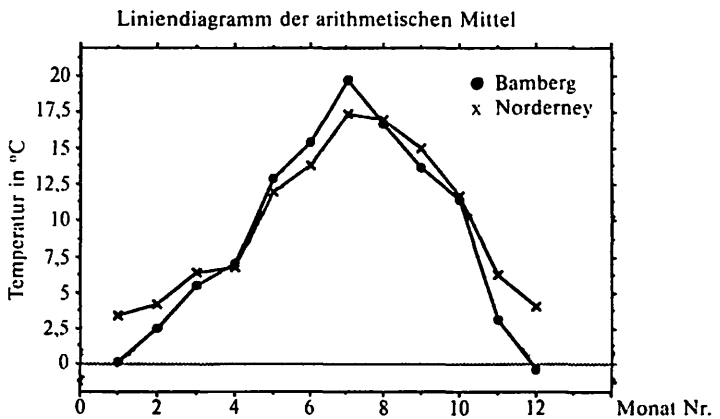

Abb. 2: Monatsmittelwerte für Bamberg und Norderney 1967 (arithmetische Mittel)

Welche Beziehung besteht nun zwischen den Mittelwerten in Abbildung 2 und den Bimodalitäten in den Histogrammen? Kann man die Bimodalität auch anhand von Abbildung 2 erwarten? Dazu muß man von der Zeitdimension abstrahieren und die Daten in Abbildung 2 (gedanklich) auf die Temperaturachse projizieren. Zwei getrennte Streudiagramme (unverbundene Punkte!) wären hierzu besser geeignet, weil man die Punktemenge visuell besser umstrukturieren kann, z. B. sie besser von der Temperaturachse her lesen kann, während in Abbildung 2 die zeitliche Verlaufsweise optisch hervorgehoben wird.

In einer solchen Abbildung würde man eine »Lücke« zwischen April und Mai einerseits und Oktober und November andererseits noch besser entdecken, als dies bei genauem Hinsehen in Abbildung 2 schon der Fall ist. Das Jahr zerfällt im Grunde in zwei Wetterperioden, und zwar Mai-Oktober und November-April. Stimmt dies Ergebnis mit unseren alltäglichen Wettererfahrungen überein?

Man kann nun durch Aufteilen der Daten in zwei Gruppen (»Sommerhalbjahr«, »Winterhalbjahr«) für die beiden Perioden getrennt Histogrammvergleiche und Kennzahlenvergleiche durchführen.[7] Diese Einteilung scheint angemessener als die Einteilung nach kalendarischen Jahreszeiten.

Wie lassen sich Unterschiede im jahreszeitlichen Verlauf zwischen den beiden Städten noch besser analysieren als auf der Grundlage von Abbildung 2? Einfache Operationen mit Diagrammen können den Analyseprozeß unterstützen, z. B. eine getrennte Darstellung für beide Städte in zwei Diagrammen oder die Darstellung der Kurve der Differenzen der Mittelwerte (Abb. 3). Man wendet hier eine allgemeine Strategie graphischer Datenanalyse an, die Chambers et al. (1983, 326 f.) »removing gross structure« nennen: die Hauptstruktur (hier die ähnlich gelagerten Monatseffekte) wird »herausgenommen«, damit man sich auf die verbleibende Feinstruktur besser konzentrieren kann. Diese Strategie wird bekanntermaßen auch bei Wahlanalysen im Fernsehen angewandt, wenn man zu den Gewinn-Verlust-Diagrammen für die einzelnen Parteien übergeht.

In Abbildung 3 sieht man den Verlauf des mittleren Temperatur*unterschieds* viel klarer: Man registriert einen etwa linearen An- und Abstieg mit Maximum im Juli; aber überraschend: der Oktober fällt aus diesem Trend heraus. Könnte dies ein Artefakt der angewendeten Methode sein, schließlich ist die feste und diskrete Aufteilung nach Monaten willkürlich? Wir kommen weiter unten darauf zurück. Oder liegt es an der Wahl des arithmetischen Mittels, das empfindlich gegenüber Ausreißern ist. Sollte man nicht lieber die 12 Monate anhand ihres Medians miteinander vergleichen?

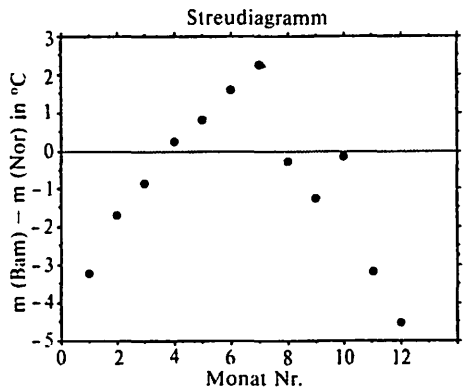

Abb. 3: Differenz der Monatsmittelwerte: Bamberg – Norderney (arithmetische Mittel)

Im Diagramm liegt noch eine wichtige Eigenschaft des arithmetischen Mittels verborgen: Dargestellt ist in Abbildung 3 die Differenz der Monatsmittelwerte. Kann man daraus z. B. schließen, daß es etwa im Januar in Norderney im Mittel ca. 3,2 °C kälter als in Bamberg war? Nein, nicht direkt, denn dargestellt ist die Differenz der Mittelwerte, nicht der Mittelwert der Differenzen. Daß diese beiden Werte für das *arithmetische* Mittel aber immer gleich sein müssen (Linearität), für den Median als Mittelwert aber nicht, kann in diesem Kontext eine unerwartete und bedeutungsvolle Entdeckung werden.

2.4 Boxplots und Mittelwert-plus-Streuung-Diagramme für den Vergleich mehrerer Monate und Städte

Eine wesentliche Anwendung von Kennzahlen besteht in dem Vergleich verschiedener Datensätze. Welche Kennzahlen zu wählen sind, wie stark man die Daten zusammenfassen sollte, um überhaupt Strukturen zu sehen, kann sich schon aus der inhaltlichen Fragestellung ergeben. Häufig ist es aber nützlich, mit mehreren Zusammenfassungen zu experimentieren, bevor man sich entscheidet. Das Experimentieren mit verschiedenen Kennzahlen kann dabei auch unter Lerngesichtspunkten nützlich sein, wenn es benutzt wird, um die Anwendungsmöglichkeiten verschiedener Kennzahlen zu erfahren. Wir setzen die Analyse fort, indem wir über das bisher ausschließlich verwendete arithmetische Mittel hinausgehen.

In Abbildung 4 findet sich für die 12 Monate eine Darstellung, die wir oben Mittelwert-plus-Streuung-Diagramm genannt haben: Für alle 12 Monate ist der arithmetische Mittelwert m durch einen Punkt dargestellt, das Intervall »m ± Standardabweichung« ist durch jeweils eine Strecke markiert.

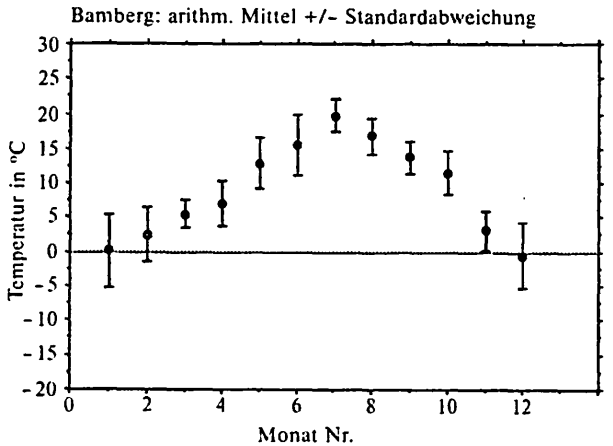

Abb. 4: Tagestemperaturen in Bamberg 1967
(Mittelwert-plus-Streuung-Diagramm für 12 Monate)

Man kann an der Länge der Strecken erkennen, das einerseits die Wintermonate Januar, Februar und Dezember eine größere Streuung aufweisen, dazu noch Mai und Juni. Der April, der angeblich macht, was er will, fällt nicht so auf. Wenn die Daten approximativ normalverteilt wären, könnte man aus dem Mittelwert-plus-Streuung-Diagramm viel mehr Informationen entnehmen, z. B. das ca. 68 % der Daten in dem Bereich m ± s liegen, ca. 95 % in m ± 2s usw. Die Symmetrie des Diagramms stände für eine approximative Symmetrie der Daten. Bei beliebigen Verteilungen darf man dies aber nicht aus dem Diagramm schließen. Will man wissen, in welchem Bereich z. B. die mittleren 68 % der Daten streuen, so kann man diesen Bereich ja »direkt« mit dem Computer ausrechnen, ohne über Umwege zu gehen. Zusammenfassungen auf der Basis der Prozentwerte sind also eine Alternative, die allerdings mit dem einfachen Taschenrechner nicht praktikabel wäre.

Die fünf Zahlen »Minimum, 1. Quartil, Median, 3. Quartil, Maximum« werden zwar auch bei DIFF (1982, 17) zum Vergleich auf tabellarischer Grundlage herangezogen, es lohnt sich aber sehr, für Vergleichszwecke diese Zahlen graphisch darzustellen: in Form des sogenannten Boxplots (Abb. 5, siehe gegenüberliegende Seite).

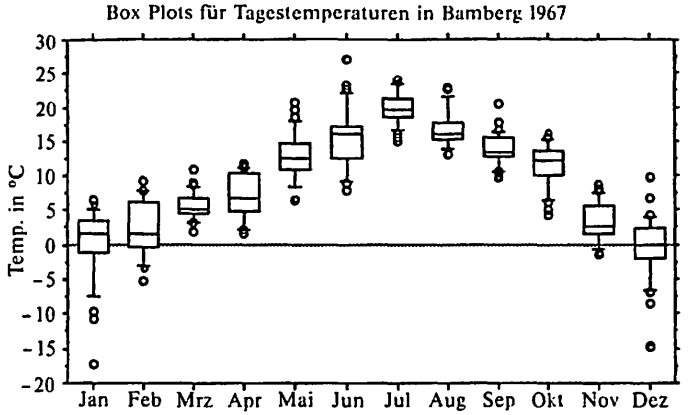

Abb. 5: Tagestemperaturen in Bamberg 1967
Zusammenfassung nach 12 Monaten: Boxplots

In den Boxplots entspricht z. B.

Begriff	Darstellung
Median	mittl. Linie der Box
oberes (3.) Quartil	obere Begrenzung der Box
unteres (1.) Quartil	untere Begrenzung der Box
90 %-Wert	Ende der oberen »Antenne«
10 %-Wert	Ende der unteren »Antenne«
Minimalwert	unterster extra gezeichneter Punkt
Maximalwert	oberster extra gezeichneter Punkt
Streuungsmaß: Quartilabstand	Länge der Box
Streuungsmaß: Spannweite	Abstand der Extremwerte
Streuung der mittleren 80 %	Abstand der Antennenenden
Symmetrie der Verteilung	Symmetrie der Box
einzelner Daten oberhalb des 90 %-Wertes und unterhalb des 10 %-Wertes	einzelne Punkte außerhalb der Box und der Antennen
(keine Bedeutung)	Breite der Box

Median und Quartilabstand sind dabei wesentlich robuster gegenüber dem Einfluß von »Ausreißern« als arithmetisches Mittel und Standardabweichung. Die Boxplots porträtieren die empirische Verteilung sehr viel genauer als die Darstellung in Abbildung 4.

Man kann ein Diagramm wie Abbildung 5 im Vergleich zu Abbildung 4 nutzen, um die Notwendigkeit und den Sinn verschiedener Streuungsmaße zu verdeutlichen. Januar und Februar sehen im Mittelwert-plus-Streuung-Diagramm fast gleich aus, mit Hilfe der Boxplots differenzieren wir zwischen dem Anteil der Streuung durch die Ausreißer und dem Anteil durch die Streuung in der Hauptgruppe der Daten: die mittleren 50 % sind im Februar viel größer, weit abliegende Ausreißer existieren nicht.

Wenn einem die Komplexität von Abbildung 5 zu groß ist, kann man sich durch verschiedene Operationen auf Details konzentrieren, z. B. durch Vergrößern einer Teilmenge der Boxplots oder durch Darstellen der zeitlichen Entwicklung einer oder mehrerer Kennzahlen als Linien- oder Streudiagramm.

Wir wollen die Frage 3 nach dem Kontinentalklima von Bamberg weiterverfolgen. Wir vermuten, daß sich die größere Unausgeglichenheit durch eine größere Variation der Temperaturen in Bamberg zeigt. Dies wissen wir bereits aus dem Histogrammvergleich und untersuchen dies weiter mit Hilfe von Boxplots, die auch einen guten Ansatzpunkt für die Bearbeitung der Frage 1 darstellen (s. Abb. 6).

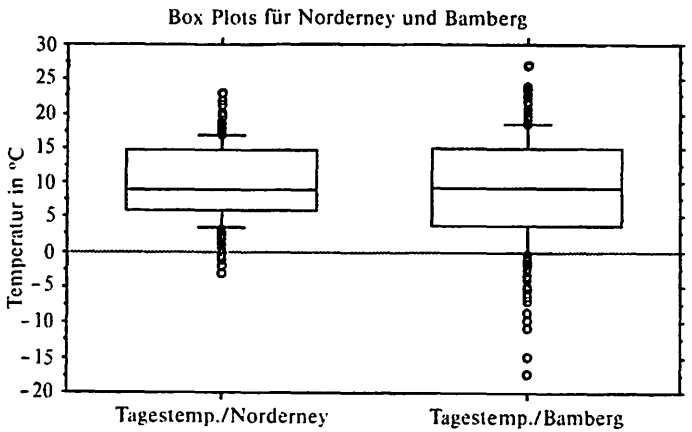

Abb. 6: Boxplots für Temperaturen im ganzen Jahr 1967

Man erkennt u. a., daß die Streuung, gemessen durch Spannweite und Quartilabstand, in Bamberg größer ist als in Norderney, während sich die Mediane kaum unterscheiden. Die Temperaturen schwanken in Norderney weniger, das Klima ist ausgeglichener. Bamberg weist einige sehr kalte Tage auf. Die Bimodalität der Verteilung kommt hier in den Boxplots nicht zum Ausdruck.

Wieweit gelten diese Strukturen auch für einzelne Monate? Welches Erscheinungsbild der Boxplots erwarten wir aufgrund unserer allgemeinen Kenntnisse über Wetter in den verschiedenen Jahreszeiten und Monaten oder unserer bisherigen Analyseergebnisse z. B. für Juli oder für Januar? Wir vermuten vielleicht, daß es im Winter in Bamberg und im Sommer in Norderney kälter ist – wobei wir offenlassen, was wir unter »kälter« verstehen. In Abbildung 7 finden sich die entsprechenden Boxplots für Januar.

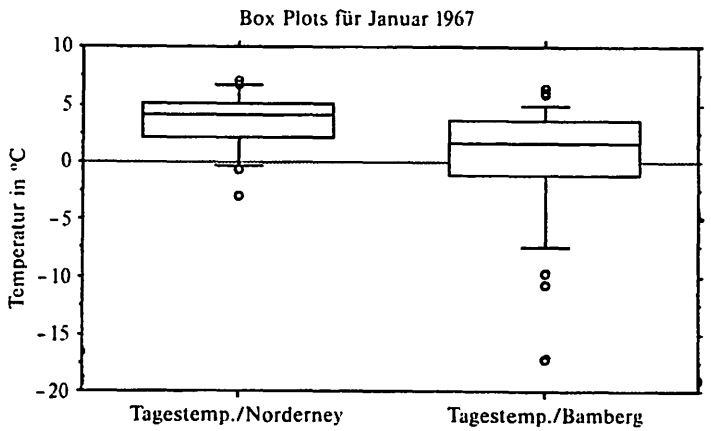

Abb. 7: Januar: Temperaturen in Bamberg und Norderney 1967

Anhand von Abbildung 7 kann man differenziert die beiden Städte vergleichen, u. a.:

- Die Mediantemperatur von Bamberg im Januar ist niedriger als die von Norderney – im Unterschied zu den Medianen im Jahresboxplot aus Abbildung 6.
- Die Streuung gemessen durch die Spannweite, Quartilabstand und Streuung der mittleren 80% ist in Bamberg größer als in Norderney: Die Verteilung scheint gegenüber Norderney nach unten verschoben und gleichzeitig gestreckt worden zu sein. Bamberg hatte einige besonders kalte Tage.

Es ist eine gute Übung, Vermutungen über die Gestalt von Boxplots für andere Monate zu entwickeln, und diese dann zu überprüfen.

Im Hinblick auf die Frage des ausgeglicheneren Klimas kann man jetzt schon differenzieren in zwei vorläufige Antworten:

1. Die 12 Monatsmittelwerte schwanken im Jahr stärker in Bamberg als in Norderney. (Abb. 2)
2. Eine stärkere Streuung haben aber vermutlich auch die Daten *innerhalb* einzelner Monate in Bamberg verglichen mit Norderney (bisher überprüft für Januar).

Das zweite Phänomen ist möglicherweise unerwartet und weist auf einen weiter aufzuklärenden zweiten Aspekt klimatischer Unausgeglichenheit hin.

2.5 Streudiagramme und gleitende Mittelwerte für die Rohdaten

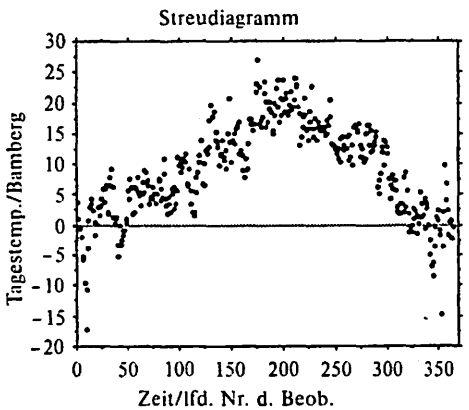

Abb. 8: Temperatur in Bamberg 1967 (Jahresverlauf)

SchülerInnen werden i. d. R. mit Daten schon in einer reduzierten und vorverarbeiteten Form konfrontiert, und zwar schon deshalb, weil man umfangreiche Originaldaten nicht ohne Computer verarbeiten kann. Geht man von umfangreicheren Rohdaten aus, so ist der Wert und die Notwendigkeit der Zusammenfassung der Daten – wie z. B. in Abbildung 2 – als Prozeß besser erfahrbar. In Abbildung 8 findet sich ein Streudiagramm für den Jahresverlauf der mittleren Tagestemperaturen für Bamberg, in dem man durchaus als Grobstruktur das An- und Absteigen eines relativ gleichmäßigen Bandes erkennt.

Würde man in Abbildung 8 noch die Temperaturen für Norderney, z. B. durch ein Kreuzchen markieren, stellte man fest, daß ein Vergleich auf dieser Grundlage sehr schwer fällt, auch wenn man einige Einzelheiten und Entwicklungstrends erkennen könnte. Dies zu erleben ist eine wichtige Lernerfahrung.

Es besteht nun eine wichtige Strategie für die weitere Analyse darin, die Daten zu vergröbern, sie in unterschiedlicher Weise zu aggregieren, z. B. das Jahr in 52 Wochen oder in 12 Monate aufzuteilen. Auch andere Aufteilungen, die sich nicht unbedingt an den Standardunterteilungen orientieren, wären denkbar, z. B.: Aufteilung des Jahres in 10, 20, 30 Teilintervalle. Die Aggregation und Zusammenfassung hat – optisch gesehen – einen »Glättungseffekt«. Eine Alternative wäre, die transformierten Daten $bam_i - nor_i$, $1 \leq i < 365$ näher zu untersuchen.

Die feste Diskretisierung wie in Abbildung 2 ist in gewissen Sinne künstlich. Mit Computerunterstützung kann man leicht sogenannte *gleitende Mittelwerte* als Beispiel für die o. g. »wandernden Zusammenfassungen« anwenden. Für solche lokale Mittelwertbildungen, die über die Daten hinweggleiten, ist in den vergangenen Jahren eine Fülle von neuen

Verfahren entwickelt worden, die dazu dienen, Strukturen in einem Streudiagramm herausarbeiten zu helfen, die man nicht von vornherein »hineinstecken« möchte, z. B. durch die Annahme bestimmter mathematischer Funktionsklassen. Ein rechnerisch elementares wie seit langem angewendetes Verfahren, das aber nicht robust gegen Ausreißer ist, besteht darin, für jeden Punkt das arithmetische Mittel einer Umgebung dieses Punktes zu berechnen.[8]

In Abbildung 9 findet sich ein Liniendiagramm für die gleitenden Mittelwerte mit einer Spanne von 31, d. h. jeweils 15 Werte rechts und links des betreffenden Punktes wurden zur Mittelwertbildung noch mit hinzugezogen, d. h.

$$\text{mgleit 31 (bam)}_i := \frac{1}{31} \sum_{j=i-15}^{i+15} \text{bam}_j$$

Gegenüber der festen Einteilung nach Monaten ist also hier der Rechenaufwand ca. 30 mal so groß.

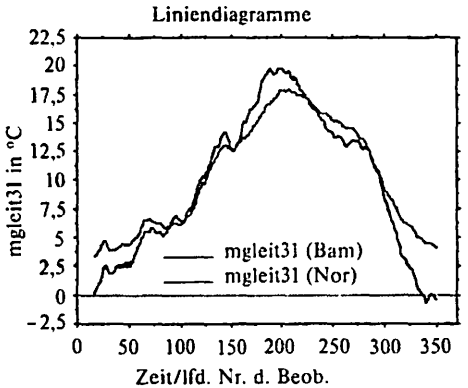

Abb. 9: Temperatur in Bamberg und Norderney (Gleitende arithmetische Mittel: Spanne 31, mgleit 31 (bam)$_i$; $16 \leq i < 350$)

Abbildung 9 muß im Kontext von Abbildung 2 und einer Abbildung wie Abbildung 8 zuzüglich Norderney-Daten gesehen werden. Letztgenannte Abbildung wäre für einen Vergleich praktisch unbrauchbar.

Gegenüber Abbildung 2 kann man in Abbildung 9 nun den mittleren Temperaturverlauf feiner verfolgen. Man erkennt u. a. daß bestimmte lokale Schwankungen z. B. um Tag Nr. 25 und Tag Nr. 75 herum in beiden Städten gleichartig verlaufen. Bemerkenswert ist, wie aus dem schwer zu durchschauenden »Chaos« durch Mittelwertbildung eine Struktur sichtbar wird.[9] An solch einem Beispiel ließe sich auch die Wirkungsweise von Glättern mit verschiedenen Spannen und verschiedenen Mittelwertbildungen graphisch untersuchen und demonstrieren.

2.6 Korrelationen und Streudiagramme für Wetterabhängigkeiten »benachbarter« Tage

Die Fragen 4 und 5 beziehen sich auf das Thema Abhängigkeit, Assoziation und Korrelation von Variablen. Dies läßt sich auf verschiedene Weise mit Computerunterstützung explorieren. Anhand der Streudiagramme für die Zeitreihen haben wir gesehen, wie eng die Temperaturen in Bamberg und Norderney gemeinsam variieren, also assoziiert sind, ohne daß sie gegenseitig füreinander Ursache sind, vielmehr unterliegen beide gemeinsamen generellen Einflußfaktoren (Frage 4). Wie gut kann man nun aber aus der Kenntnis der Bamberg-Temperatur pragmatisch die Temperatur in Norderney vorhersagen? Dies könnte man z. B. näher durch ein Streudiagramm für (bam$_i$, nor$_i$) untersuchen.

Wir wollen nun einem Teilaspekt der Frage 5 nachgehen und untersuchen, welcher Zusammenhang zwischen der Temperatur am Tag i mit der Temperatur am Tag i + 1, i + 5, ..., i + 30 besteht. Hier hat man eine ganze Schar von Variablenpaaren, von denen man erwartet, daß dieser Zusammenhang i. a. »schwächer« wird, wenn der Abstand zwischen den Tagen größer wird. Wie stellt sich dies empirisch für das Jahr 1967 dar? Dazu tragen wir auf die Temperatur für Tag i gegen die von Tag i + 1 bzw. i + 30 in je einem Streudiagramm ein (Abb. 10 und Abb. 11).

Wir erkennen die Abnahme der Stärke der Assoziation deutlich; es ist lohnenswert, die Ausreißer im linken unteren Ende mit den anderen bisherigen Graphiken in Beziehung zu setzen. Ein Zusammenhang geht sowohl numerisch als auch optisch klar hervor. Man kann noch die Frage behandeln, wie gut man hiermit ganz grobe Wetter-(Temperatur-)vorhersagen machen könnte: Die Spannweite der Vorhersage wächst von ca. 10 °C bei einem Tag Abstand auf ca. 20 °C bei 30 Tagen! Interessant wäre es, auch für simulierte unabhängige Zufallszahlen Vergleichsdiagramme zur Verfügung zu haben.

Man kann die Assoziation, die sich hier ganz gut als ein linearer Zusammenhang bis auf einige Ausreißer am linken Rand darstellt, numerisch mit einem Korrelationskoeffizienten zusammenfassen; vor allem dann, wenn man viele Variablen paarweise untersuchen möchte, kann ein explorativer Wert in einer solchen radikalen Zusammenfassung liegen, bei dem allerdings viele Besonderheiten der Daten nicht mehr sichtbar sind.

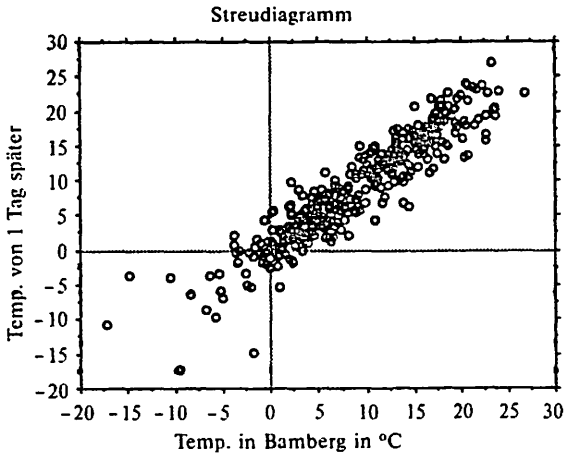

Abb. 10: Temperatur benachbarter Tage in Bamberg (bam_{i+1} vs. bam$_i$; $1 \leq i \leq 364$)

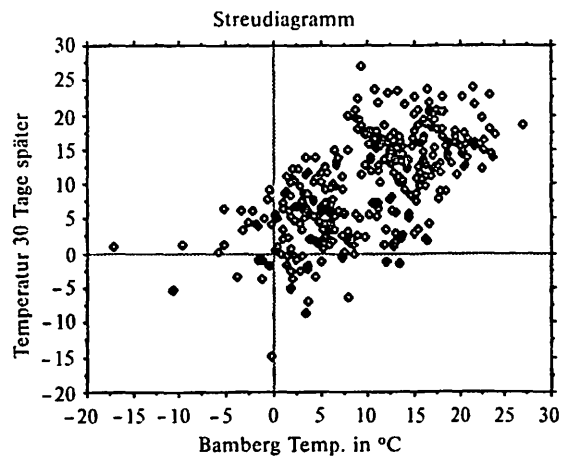

Abb. 11: Temperatur nach einem Monat in Bamberg (bam_{i+30} vs. bam$_i$; $1 \leq i \leq 335$)

Als Korrelationskoeffizienten[10] r(x, y) ergeben sich, wenn man für x immer die Variable bam_i nimmt und für y die in der Tabelle angegebenen Variablen:

x \ y	bam_{i+1}	bam_{i+5}	bam_{i+10}	bam_{i+30}
bam_i	0,932	0,787	0,737	0,628

Man erkennt die Abnahme der Korrelation mit der wachsenden Entfernung der Tage voneinander. Die Erfahrung lehrt allerdings, daß man mit solchen Korrelationstabellen sehr vorsichtig umgehen muß, wenn nicht überprüft wurde, ob noch weitere relevante Strukturen in den Daten enthalten sind.

Hiermit soll die Darstellung des Beispiels abgeschlossen werden, das zwar nur elementare statistische Begriffe der Sekundarstufe I verwendet hat, jedoch neue Anforderungen hinsichtlich des Umgangs mit Graphiken und Daten sowie hinsichtlich der Komplexität der Datenanalyse aufwirft. Das folgende Beispiel verwendet darüber hinaus die Grundbegriffe der beurteilenden Statistik und Wahrscheinlichkeitstheorie aus der Sekundarstufe II.

3. Statistische Analyse von realen »Zufallszahlen«: Taufen im London des 18. Jahrhunderts

3.1 Einführung

In unserem zweiten Beispiel wollen wir uns exemplarisch einem Datensatz zuwenden, der in der Entwicklung der Wahrscheinlichkeitsrechnung eine wichtige Rolle gespielt hat. Es handelt sich um die Anzahl der Jungen- und Mädchentaufen in London von 1629 - 1710 nach Graunt.[11] An diesem Datensatz sollen mit einfachen Mitteln einer Computer-Arbeitsumgebung und einigen zentralen Begriffen der beurteilenden Statistik und Wahrscheinlichkeitstheorie wie empirisches Gesetz der großen Zahl, Binomialverteilung, Normalapproximation und Konfidenzintervall Analysen durchgeführt werden, die zugleich auch einem Verständnis dieser Begriffe dienen können. Die Verwendung realer statt simulierter Daten hat dabei verschiedene Vorteile (s. auch Biehler 1988). Aus Platzgründen können manche der Analysen nur angedeutet werden.

Gegenüber den in 2. benutzten Funktionen einer Computer-Arbeitsumgebung benötigen wir hier zusätzlich:

- *Algebraisch-numerische Verarbeitung:* Möglichkeiten zu arithmetischen Operationen und Transformationen von Tabellenspalten;
- *Simulation und Modellbildung:* Simulation von Zufallszahlen einschließlich Weiterverarbeitung mit den anderen Funktionen des Systems, Wiederholungsmöglichkeiten.

Für 82 Jahre t = 1629, 1630 ... 1710 liegen vor:

N (t): Anzahl der getauften Kinder im Jahr t
J (t): Anzahl der getauften Jungen im Jahr t
M (t): Anzahl der getauften Mädchen im Jahr t

In Abbildung 12 findet sich ein erstes Streudiagramm für die zeitliche Entwicklung.

An diesen Daten kann man verschiedene Fragestellungen mit elementaren Mitteln untersuchen, z. B.:

- Wie hat sich die Gesamtheit der getauften Kinder in den 82 Jahren entwickelt? Welche Trends und Besonderheiten kann man feststellen? Welche Ursachen liegen zugrunde – z. B. Kriege, Epidemien usw.? (Streu- und Liniendiagramme, evtl. mit Glättung)[12]
- Was läßt sich über das numerische Verhältnis von Jungen- und Mädchentaufen aussagen? Wie konstant und wie variabel war es in den 82 Jahren? (Streudiagramm für Mädchen/Jungenanteil gegen Zeit, Histogramm und Boxplots für die Verteilung des Anteils)

- Gilt ein »Gesetz der großen Zahl«, stabilisieren sich die relativen Häufigkeiten? (Streudiagramm für den kumulierten Mädchen/Jungenanteil gegen die Zeit, s. Abb. 13)
- Gibt es im Verhalten der Daten bedeutsame Abweichungen von der Modellannahme einer konstanten Wahrscheinlichkeit für eine Mädchen/Jungentaufe?

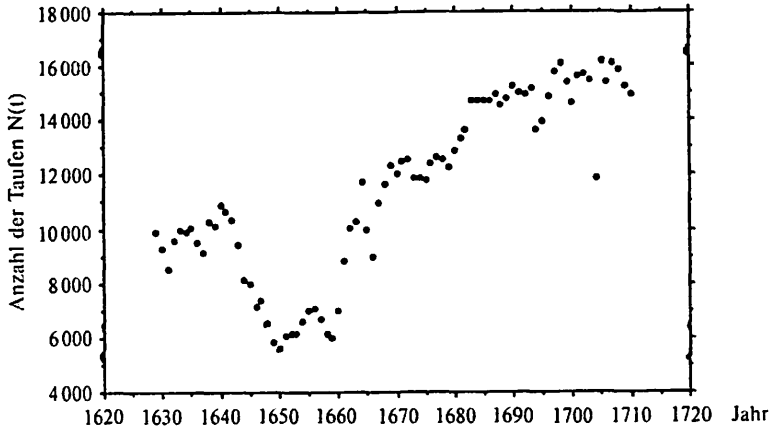

Abb. 12: Taufen in London 1629 - 1710 (Daten nach Graunt)
(Quelle der Daten: Die Daten sind enthalten in dem Brief von N. Bernoulli an de Montmort vom 23. Januar 1713, in: de Montmort, P. R.: 1713. Essay d'analyse sur les jeux de hazard Seconde Edition. Revûe et augmentée de plusieurs lettres. Paris: Jacques Quillan. Reprint 1980 durch Chelsea Publishing Company, New York, 388-394. Eine Tabelle der Daten findet sich auch im Anhang dieses Beitrags.)

Wir wollen uns vor allem mit der Entwicklung des Jungen- bzw. Mädchenanteils befassen. In dem gesamten Zeitraum finden sich 938212 Taufen, davon 484382 Jungentaufen, das entspricht einem Anteil $p_j = 0{,}51628$. Diese Zahl ist auch mit dem Wert für 1710 in Abbildung 13 identisch.

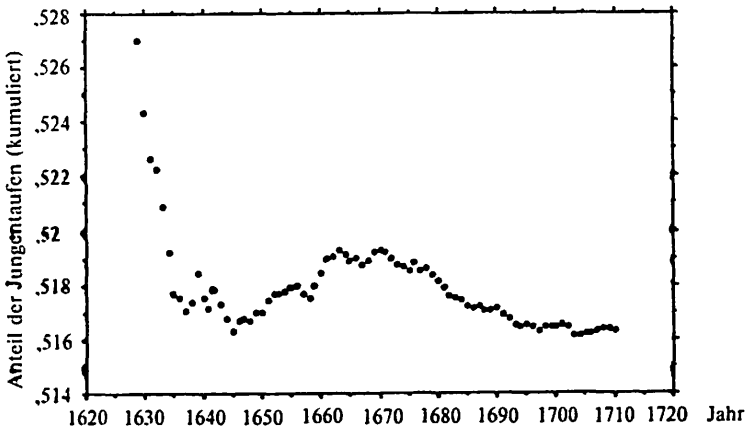

Abb. 13: Taufen in London (1629 - 1710) (kumulative relative Häufigkeit für Jungentaufen)

Hier sehen wir eine Stabilisierung, ähnlich wie man sie bei Münzwerfen vorfinden würde. Die hohe Vergrößerung der y-Achse macht aber auch bestehende Schwankungen sichtbar, die bei einer Achse von 0 bis 1 kaum wahrzunehmen wären. Die Möglichkeit mit Skalenwechsel zu explorieren und zu manipulieren, ließe sich hieran gut demonstrieren.

Die eigentlich interessante Frage besteht nun aber darin, wie konstant bzw. wie variabel der Jungenanteil auf der Basis der *einzelnen* Jahre gewesen ist. Mit Hilfe der algebraisch-numerischen »Tabellenarithmetik« rechnet man sich den Jungenanteil im Jahr t aus. Er soll mit $h_j(t) = \frac{J(t)}{N(t)}$ bezeichnet werden. Diesen neuen Datensatz kann man jetzt mit Mitteln der beschreibenden Statistik beschreiben und untersuchen (Histogramm, Boxplot, Kennzahlen). Der Jungenanteil $h_j(t)$ liegt immer über 50 %, seine maximale relative Abweichung von p_j beträgt z. B. weniger als ca. 0,02 d. h. weniger als 4 % relative Abweichung. Das sind interessante Entdeckungen, die auch historisch bedeutsam waren. Wir wollen hier gleich die Abweichungen von dem Schätzwert p_j untersuchen. In Abbildung 14 sind die *Residuen*, die »Modell-Daten-Abweichungen« $h_j(t)-p_j$ geplottet.

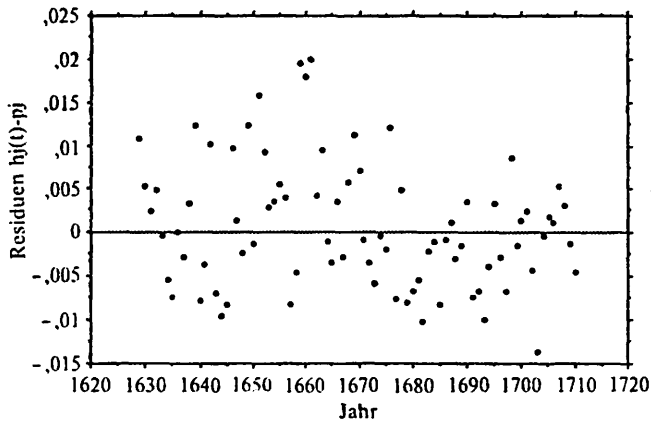

Abb. 14: Jungenanteil bei Taufen in London (Residuen $h_j(t)-p_j$)

Bei der hohen Auflösung der Skalierung in Abbildung 14 sieht man eine deutliche Variation. Wir entdecken in Abbildung 14 überraschenderweise einen gewissen Trend zu niedrigeren Jungenanteilen im Laufe der Zeit. Nach 1680 kommen »hohe« Jungenanteile praktisch nicht mehr vor.

Spricht diese Entdeckung nun gegen die Annahme der Hypothese einer konstanten Wahrscheinlichkeit für eine Jungentaufe oder kann auch »rein zufällig« solch ein Muster entstehen? Diese Problemstellung, die eine Herausforderung an die stochastische Modellierung stellt, ist eine für die schulische beurteilende Statistik vergleichsweise offene Fragestellung. Es sollen zwei Wege aufgezeigt werden, wie man dies Problem bearbeiten kann.

3.2 Konfidenzintervalle und Hypothesentests in wiederholter Anwendung

Mit den schulüblichen Begrifflichkeiten kann man an die Daten von einer Perspektive herangehen, die zugleich auch das Konzept des Konfidenzintervalls, des Hypothesentests und der Häufigkeitsinterpretation der Irrtumswahrscheinlichkeiten verdeutlichen kann.

Wir gehen von der Annahme aus, daß für jedes Jahr t der Jungenanteil binomialverteilt ist mit unbekannter Wahrscheinlichkeit $p_j(t)$ und Stichprobenumfang $N(t)$. Für alle 82 Jahre ließe sich jetzt die Hypothese $p_j(t) = p_j$ vs. $p_j(t) \neq p_j$ testen, z. B. auf dem 5 %-Signifikanzniveau. Mit unserem Datenanalysesystem oder auch nur mit einem einfachen Tabellenkalkulationssystem kann man die nötigen 82 Rechnungen durch nur *eine* Rechnung mit Tabellenspalten ersetzen. Was erwartet man als Testergebnisse, selbst wenn $p_j(t) = p_j$ immer zutrifft? Man erwartet in etwa 5 % von 82 Jahren fälschliche Verwerfungen. Wir wollen dies hier nicht durchführen, sondern den Weg über Konfidenzintervalle beschreiten.

Das Konzept des 95 %-Konfidenzintervalls wird oft dadurch erläutert, daß hiermit ein Verfahren gegeben ist, das mit 95 % Wahrscheinlichkeit ein Intervall erzeugt, das die wahre Wahrscheinlichkeit überdeckt. Bei wiederholter Anwendung erwartet man auf lange Sicht, daß etwa 5 % der Intervalle die Wahrscheinlichkeit nicht überdecken. Dies wird manchmal durch Simulation illustriert. Mit den vorliegenden Daten kann man nun ein retrospektives Realexperiment durchführen. Für jedes der 82 Jahre nehmen wir an, daß wir die theoretische Wahrscheinlichkeit für eine Jungentaufe $p_j(t)$ nicht kennen, sondern vielmehr durch ein 95 %-Konfidenzintervall schätzen. Für alle 82 Jahre t errechnet sich für das Konfidenzintervall für die unbekannten Wahrscheinlichkeiten $p_j(t)$ die Bedingung:

$$|h_j(t) - p_j(t)| \leq 1{,}96 * \sqrt{\frac{p_j(t) * (1 - p_j(t))}{N(t)}}$$

Wie auch in den meisten Schulbüchern üblich, rechnen wir mit der rechentechnisch einfacheren Variante der Bedingung[13]:

$$(*) \qquad |h_j(t) - p_j(t)| \leq 1{,}96 * \sqrt{\frac{h_j(t) * (1 - h_j(t))}{N(t)}}$$

In Abbildung 15 finden wir ein Diagramm, in das die Punkte $(t, h_j(t))$ eingetragen sind; um diese Punkte herum ist das jeweilige Konfidenzintervall gemäß (*) gekennzeichnet, ähnlich dem Mittelwert-Streuungs-Diagramm aus dem ersten Beispiel.[14] Unter der Annahme einer konstanten Wahrscheinlichkeit $p_j = 0{,}51628$ würde man erwarten, daß etwa 5 % von 82, also etwa vier Intervalle die theoretische Wahrscheinlichkeit nicht enthalten. Durch Auszählen in der Graphik oder Berechnung mit dem Datenanalysesystem stellt man fest, daß hingegen *14* von 82 Intervallen dieses p_j nicht enthalten. Starke Zweifel an unserer Hypothese von einem konstanten p_j sind also angebracht!

Man kann weitere Fragen zum Verständnis von Konfidenzintervallen an die Graphik richten, z. B. warum die Konfidenzintervalle um 1655 deutlich länger sind als diejenigen nach 1700. Es fällt auch auf, daß die Konfidenzintervalle in drei aufeinanderfolgenden Jahren bei 1660 p_j nicht überdecken: der notwendige Jungenüberschuß nach Ende des Bürgerkrieges?

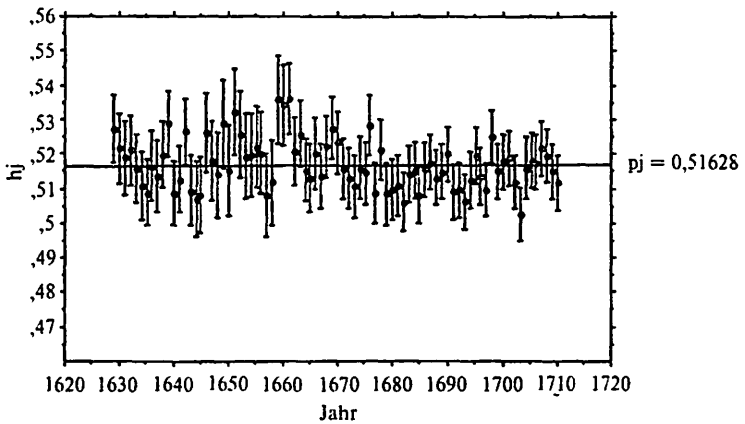

Abb. 15: Konfidenzintervalle für die unbekannten Wahrscheinlichkeiten für eine Jungentaufe

3.3 Beurteilende Statistik als kritische Statistik

Wir wollen wieder auf Abbildung 14 zurückgehen und einen zweiten Weg beschreiten, der stärker genetisch an die Grundidee der beurteilenden Statistik als »kritischer Statistik« anknüpft, die einmal folgendermaßen formuliert wurde:

»... zufallsbedingte Variation [chance variation] kann ein Muster produzieren, das eine Regularität simuliert. Wenn der Forscher hierbei getäuscht wird, kann ein Trugschluß entstehen. Die Wissenschaft Statistik wurde mit dem Ziel entwickelt, die Trennung von Zufallseffekten und wirklichen Regularitäten zu unterstützen.« (Wilson 1952, 169; Übers. R. B.)

»Die meisten Statistiker des frühen 19. Jahrhunderts, die primär praktische Leute waren, würden sofort über die Gründe spekuliert haben für die Differenz der Geburtenraten in Paris und in Frankreich als ganzem, ohne sich Gedanken zu machen, wie es Laplace tat, ob der Unterschied ausreichend war, eine ›Untersuchung zu autorisieren‹ ... es war leichter und verführerischer, mögliche konstante Ursachen für beobachtete Unterschiede vorzuschlagen, als zu berechnen, ob die Unterschiede selber wirklich wichtig waren.« (Hilts 1973, 211; Übers. R. B.)

Beide Zitate beziehen sich auf eine Erkenntnissituation, wo zunächst eine Auffälligkeit entdeckt wurde, die danach kritisch aus statistischer Sicht beurteilt werden soll. Es gibt allerdings auch viele Anwendungssituationen, wo es notwendig oder zweckmäßig ist, statistische Hypothesen ohne Voruntersuchung der Daten aufzustellen und zu prüfen.

Zu erleben, wie der Zufall Muster produziert, ist eine wesentliche statistische Erfahrung, die insbesondere wichtig für eine kritische Beurteilung von Graphiken ist. In diesem Fall wäre es also interessant, die Prozesse unter der Annahme eines konstanten p_j durch Simulation zu reproduzieren und für die simulierten Daten Graphiken herzustellen, die Abbildung 14 entsprechen.

Man würde so einen qualitativen Eindruck gewinnen können und könnte durch mehrfache Wiederholung auch sehen, ob Muster mit abnehmendem Trend wie in Abbildung 14 häufiger vorkommen. Auf diesem Hintergrund kann sich dann genetisch die Grundidee anschließen, ein numerisches Kriterium zu finden, daß gleichsam die Stärke der Abweichung, den Trend, durch eine Funktion aus den Daten, eine Statistik beschreibt. Mißt man etwa den Trend durch einen Korrelationskoeffizienten τ und erhält einen Wert τ_{beob}, so kann man für die statistische Beurteilung durch Simulation feststellen, wie groß die Wahrscheinlichkeit $P(|\tau| \geq |\tau_{beob}|)$ ist. Aus Platzgründen können wir dies hier nicht durchführen.

Wenn ein Datenanalysesystem Möglichkeiten zur wiederholten Simulation und Weiterverarbeitung der Daten enthält, so ist es möglich, auch für situationsspezifische ad hoc Kriterien, für die gar keine Theorie zur Verfügung stehen muß, zu einer statistischen Beurteilung zu kommen!

Die Beurteilung von Abbildung 14 wird nun noch durch weitere Umstände erschwert. Eine statistische Beurteilung einer Abweichung von z. B. 0.01 ist direkt nicht möglich, man muß dazu erst Rechnungen durchführen. Hinzu kommt das folgende: Wegen der unterschiedlichen Stichprobenumfänge für die einzelnen Jahre sind die (theoretischen) Varianzen in den einzelnen Jahren unterschiedlich; qualitativ und grob gesprochen: in den Jahren mit weniger Taufen können größere Abweichungen eher vorkommen. Zur Verbesserung der diagnostischen Möglichkeiten von Abbildung 14 kann man Datentransformationen vornehmen.

Die Anzahl der Jungentaufen $J'(t)$[15] ist eine Zufallsgröße, deren Realisierung $J(t)$ wir beobachtet haben. Die dem Anteil $h_j(t)$ korrespondierende Zufallsgröße bezeichnen wir mit $h_j'(t)$. Wir gehen nun also von der Annahme aus, daß $J(t)$ die Realisierung einer binomialverteilten Zufallsgröße $J'(t)$ mit den Parametern p_j und $N(t)$ ist.[16] Wegen der Größe von $N(t)$ kann man die Normalapproximation für $J'(t)$ anwenden. Also gilt:
$J'(t)$ ist approximativ normalverteilt mit

$$E(J'(t)) = N(t) * p_j \text{ und } \operatorname{var}(J'(t)) = p_j * (1 - p_j) * N(t).$$

Die standardisierte Zufallsgröße

$$Z'(t) = \frac{J'(t) - N(t) * p_j}{\sqrt{p_j * (1 - p_j) * N(t)}}$$

ist normalverteilt mit Erwartungswert 0 und Varianz 1. Man transformiert nun die Folge der

beobachteten Daten für die Jungenanzahl $J(t)$ nach $Z(t)$ gemäß

$$Z(t) = \frac{J(t) - N(t) * p_j}{\sqrt{p_j * (1 - p_j) * N(t)}} = \sqrt{N(t)} * \frac{1}{\sqrt{p_j * (1 - p_j)}} * \left(\frac{J(t)}{N(t)} - p_j\right)$$

$$= \sqrt{N(t)} * \frac{1}{\sqrt{p_j * (1 - p_j)}} * (h_j(t) - p_j) = \frac{h_j(t) - p_j}{\sigma_j(t)}, \text{ wobei}$$

$$\sigma_j(t) = \sqrt{\frac{p_j * (1 - p_j)}{N(t)}}$$

der (theoretischen) Standardabweichung der Zufallsgröße $h_j'(t)$ entspricht. $Z(t)$ bezeichnen wir als die standardisierten Residuen und stellen sie in Abbildung 16 dar.

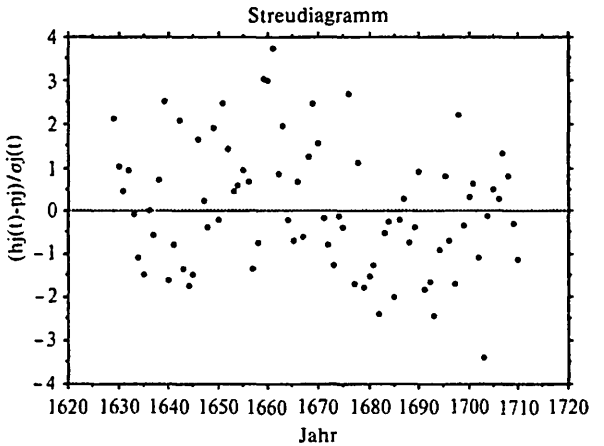

Abb. 16: Standardisierte Residuen

$$Z(t) = \frac{h_j(t) - p_j}{\sigma_j(t)}$$

Welches ist die Wirkung der Transformation im Vergleich zu Abbildung 14? Nimmt man die niedrigen $N(t)$ in den Jahren 1645 - 1660 als Vergleichsmaßstab (s. Abb. 12), so führt die Transformation zu einer relativen Vergrößerung der Absolutwerte der standardisierten Residuen in den anderen Zeiträumen. Besonders stark ist dies für die hohen $N(t)$ ab 1680 festzustellen. Dies kann man im genauen visuellen Vergleich der Diagramme wiederfinden.

Wenn weiterhin angenommen wird, daß die $J'(t)$ stochastisch unabhängig sind, müßten sich diese 82 Daten $Z(t)$ wie eine Zufallsstichprobe aus einer Standardnormalverteilung verhalten.

Mit dieser Transformation ist nun ein beachtlicher Fortschritt erreicht worden:

- Für den Vergleich mit der theoretischen Erwartung müssen jetzt nur einfache Stichproben aus einer Normalverteilung simuliert werden.
- Der visuelle Vergleich mit solcherart simulierten Daten läßt sich besser durchführen, wenn man ein gewisses »Gefühl« für das Verhalten einer Stichprobe aus einer Normalverteilung entwickelt hat.
- Auf der y-Achse steht eine Standardskala zur Verfügung, die den Vergleich und die statistische Bewertung der Residuen erleichtert: im Bereich ± 1 liegen theoretisch 68,3 % der Daten, im Bereich ± 2 liegen theoretisch 95,5 % der Daten, im Bereich ± 3 theoretisch 99,7 %.

Einen Test der Hypothese $p_j(t) = p_j$ vs. $p_j(t) \neq p_j$ zum Signifikanzniveau 5 % kann man jetzt graphisch durchführen: Sie wird genau dann verworfen, wenn $|Z(t)| > 1{,}96$ ist. Im vorliegenden Fall ist das genau 14 mal der Fall, bei Annahme konstanter Jungenwahrscheinlichkeit würde man im Mittel nur vier Ablehnungen erwarten.

Geringeren Signifikanzniveaus entsprechen weitere Grenzen im Diagramm. Neben den praktischen Vorteilen von Abbildung 16 wird hier auch das Konzept einer *statistisch signifikanten Abweichung* vielfältig visualisiert. Die Jahre mit statistisch signifikanten Abweichungen sind außerdem direkt identifizierbar und die Struktur dieser Jahre ist sichtbar.

Man kann jetzt auch noch anders untersuchen, welche signifikanten Abweichungen es von der Normalverteilung gibt: wenn man z. B. die 82 Daten unter Abstraktion von der zeitlichen Dimension als Verteilung analysiert, so ergibt sich als arithmetischer Mittelwert und Standardabweichung:

$$m = \frac{1}{82} \sum_{t=1629}^{1710} Z(t) = 0{,}07; \quad s = \sqrt{\frac{\sum_t (Z(t) - m)^2}{81}} = 1{,}45$$

Man könnte zur weiteren Beurteilung ein Histogramm herstellen lassen, hierin geeignete Normalverteilungen graphisch einzeichnen, die Daten mit dem Computer in ein Wahrscheinlichkeitspapier eintragen lassen, um die Art der Abweichung von der »Normalität« besser zu diagnostizieren usw.

»Intuitiv« hält man eine derartig große Abweichung der Standardabweichung von der Theorie $\sigma = 1$ kaum für möglich. Ein direkter Vergleich von s und σ ist problematisch, da in die statistische Beurteilung noch der Stichprobenumfang (hier 82) einbezogen werden muß. Zu einer statistischen Beurteilung könnte man kommen, indem man auch hier wieder Zufallsstichproben der Länge 82 aus einer Standardnormalverteilung erzeugt und daraus schätzt, wie groß die Wahrscheinlichkeit für Werte $s \geq 1{,}45$. Eine Theorie über die Verteilung der Stichprobenstandardabweichung ist damit durchaus verzichtbar.

Abschließend sei die Frage aufgeworfen, in welchem Verhältnis Abbildung 16 und die dort identifizierten statistisch signifikanten Jahre mit denen aus Abbildung 15 stehen. Rein empirisch kann man feststellen, daß es sich um dieselben 14 Jahre handelt! Man manche sich dazu klar, daß das Kriterium zu Abbildung 15 mathematisch genau – und nicht nur approximativ – äquivalent zu dem aus Abbildung 16 ist, wenn man dort bei den standardisierten Residuen nur

$\sigma_j(t)$ durch $s_j(t) = \sqrt{\frac{h_j(t) * (1 - h_j(t))}{N(t)}}$ ersetzen würde.

4. Schlußbemerkung

Die Beispiele haben deutlich gemacht, daß vom »Anwender« schwierige neue Kompetenzen gefordert werden, die sich beziehen auf

- die aktive Mitgestaltung und Auswahl von Graphiken und Methoden;
- das In-Beziehung-Setzen und Interpretieren der mit verschiedenen Methoden und Graphiken gewonnenen Ergebnisse;
- der Vergleich und die Bewertung von Methoden und Graphiken im Hinblick auf bestimmte Zwecke.

Diese Kompetenzen müßten geeignet in Zusammenhang mit der Anwendung auf reale Datensätze entwickelt werden. Dabei ist sicher eine wichtige Komponente, Daten und Graphiken nicht nur auf dem Bildschirm zu sehen, sondern auch in verschiedener Weise direkt für Papier und Bleistift zugänglich zu haben. Das gilt insbesondere, wenn man gleichzeitig mit mehreren Graphiken arbeiten will.

Was Unterricht hier unter normalen Bedingungen leisten kann und soll, muß dabei als noch offen angesehen werden. Die Beispiele sollten Anregungen bieten für weiteres Nachdenken und Ausprobieren. Dabei kann man sicher manche Aspekte auch ohne komplexe Datenanalysewerkzeuge realisieren, wenn man auf entsprechend vorbereitete Materialien zurückgreifen kann. Verloren gehen wird dabei leicht die durch die Offenheit einer Software gegebene Möglichkeit, alternative Analysewege zu gehen, dabei die von Schülern und Schülerinnen angeregten Ideen aufzunehmen sowie Graphik und rechenintensive Statistikmethoden so selbstverständlich und natürlich in die Arbeit zu integrieren, wie das von der Sache angemessen ist und eigentlich auch von den begrifflichen Voraussetzungen der Schüler und Schülerinnen möglich zu sein scheint.

5. Anhang

Anzahl der Taufen in London (1629 - 1710)

Jahr	Jungen	Mädchen	Jahr	Jungen	Mädchen
1629	5218	4683	1670	6278	5719
1630	4858	4457	1671	6449	6061
1631	4422	4102	1672	6443	6120
1632	4994	4590	1673	6073	5822
1633	5158	4839	1674	6113	5738
1634	5035	4820	1675	6058	5717
1635	5106	4928	1676	6552	5847
1636	4917	4605	1677	6423	6203
1637	4703	4457	1678	6568	6033
1638	5359	4952	1679	6247	6041
1639	5366	4784	1680	6548	6299
1640	5518	5332	1681	6822	6533
1641	5470	5200	1682	6909	6744
1642	5460	4910	1683	7577	7158
1643	4793	4617	1684	7575	7127
1644	4107	3997	1685	7484	7246
1645	4047	3919	1686	7575	7119
1646	3768	3395	1687	7737	7214
1647	3796	3536	1688	7487	7101
1648	3363	3181	1689	7604	7167
1649	3079	2745	1690	7909	7302
1650	2890	2722	1691	7662	7392
1651	3231	2840	1692	7602	7316
1652	3220	2908	1693	7676	7483
1653	3196	2959	1694	6985	6647
1654	3441	3179	1695	7263	6713
1655	3655	3349	1696	7632	7229
1656	3668	3382	1697	8062	7757
1657	3396	3289	1698	8426	7626
1658	3157	3013	1699	7911	7452
1659	3209	2781	1700	7578	7061
1660	3724	3247	1701	8102	7514
1661	4748	4107	1702	8031	7656
1662	5216	4803	1703	7765	7683
1663	5411	4881	1704	6113	5738
1664	6041	5681	1705	8366	7779
1665	5114	4858	1706	7952	7417
1666	4678	4319	1707	8379	7687
1667	5616	5322	1708	8239	7623
1668	6073	5560	1709	7840	7380
1669	6506	5829	1710	7640	7288

Anmerkungen

1) GRAPHDAS: Graphische Darstellungen zur Datenanalyse im Stochastikunterricht, R. Biehler / H. Steinbring, IDM Universität Bielefeld.
2) S. hierzu die ausführlichen Analysen bei Biehler/Rach (1990).
3) Zum Beispiel die Software TIMES für die Analyse von Daten, die aus bestimmten Reaktionszeit-Tests hervorgehen oder die Software C.I.A. für Daten aus der Simulation einfacher stochastischer Prozesse, s. Biehler/Winkelmann (1988) für eine kritische Darstellung dieser beiden Programme. Kleine Datenanalysesysteme gibt es auch für andere Schulfächer, z. B. für quantitative Eigenschaften der chemischen Elemente im Periodensystem oder für wirtschaftsgeografische Daten, z. B. das System RUDI RUHR (Raum- und Daten-Informationssystem Ruhr. Version 3.0. Braunschweig: Westermann 1989).
4) Dieses Beispiel der Wetterdaten wird bei Biehler/Rach (1990) genutzt, um die Software STATVIEW vorzustellen, zu bewerten und allgemeine Gesichtspunkte für die Gestaltung von statistischen Softwaretools zu entwickeln.
5) Diese Begriffe werden weiter unten genauer erläutert.
6) Aus Platzgründen kann hier nur eine Auswahl der tatsächlich verwendeten Graphiken abgedruckt werden.
7) S. hierzu auch DIFF 1982.
8) S. Tukey (1977, 205 ff.) für verschiedene einfache robuste Glättungsverfahren auf Medianbasis. Ein wichtiges Verfahren mit großer Anwendungsbreite ist das LOWESS-Verfahren (s. Chambers et al. 1983, 91 ff.), das auch bei Biehler/Rach (1990) auf die jahreszeitliche Entwicklung von Verkehrstoten angewendet wurde.
9) Bei der Interpretation muß man vorsichtig sein und berücksichtigen, wie der Mittelwert zustandegekommen ist. Man findet viele Veröffentlichungen, wo »glatte« Kurven in die Daten gelegt werden, ohne daß genau angegeben wird, wie diese zustande kamen. Die Kurve in Abbildung 9 darf auch nicht verwechselt werden mit den in der Geographie üblichen langjährigen Mittelwertkurven: hier ist nur *ein* Jahr in die Mittelung eingegangen.
10) In die Korrelationsrechnung wurden immer die letzten 30 Tage des Jahres nicht einbezogen, da »Bamberg + 30 Tage« hierfür keine Werte besitzt. Bei der Berechnung wurde der (übliche) Pearsonsche Produkt-Moment-Korrelationskoeffizient zugrundegelegt.
11) Vgl. Biehler/Rach 1990 und Biehler/Steinbring 1990 für weitere Analysen zu diesen Daten.
12) Die Abnahme von 1645 bis 1660 hängt mit dem Religions- und Bürgerkrieg zusammen.
13) Diese gilt wegen $\sqrt{\frac{p_j(t) * (1 - p_j(t))}{N(t)}} \approx \sqrt{\frac{h_j(t) * (1 - h_j(t))}{N(t)}}$.
14) Die Berechnung kann mit den algebraisch-numerischen Funktionen zur Tabellenarithmetik erfolgen oder aber einer eingebauten statistischen Funktion entsprechen, deren 82 Ergebniswerte aber dann leicht der Graphik übergeben kann.
15) $J^*(t)$ bezeichnet die der Zahl $J(t)$ korrespondierende Zufallsgröße.
16) Im Grunde ist $N(t)$ selber zufällig, davon abstrahieren wir.

Literatur und Software

[1] *Biehler, R.* [1982]: Explorative Datenanalyse: Eine Untersuchung aus der Perspektive einer deskriptiv-empirischen Wissenschaftstheorie. IDM Materialien und Studien 24, Bielefeld: Universität Bielefeld
[2] *Biehler, R.* [1988]: Computers in Probability Education. IDM Occasional Paper 108, Bielefeld: Universität Bielefeld
[3] *Biehler, R.* [1990]: Datenanalyse und Computer im Stochastikunterricht: Erfahrungen und Entwicklungsperspektiven. In: Die Zukunft des Mathematikunterrichts, hrsg. v. LSW Soest, Soest: Soester Verlagskontor, 92-101
[4] *Biehler, R./Rach, W.* [1990]: Softwaretools zur Statistik und Datenanalyse: Beispiele, Anwendungen und Konzepte aus didaktischer Perspektive. Reihe Neue Medien im Unterricht, hrsg. vom LSW Soest. Soest: Soester Verlagskontor. ISBN 3-8165-1739-O

[5] *Biehler, R./Steinbring, H.* [1990]: Entdeckende Statistik im Mathematikunterricht: Materialien aus dem GRAPHDAS-Projekt. Bielefeld: Universität Bielefeld IDM
[6] *Biehler, R./Winkelmann, B.* [1988]: Mathematische Unterrichtssoftware: Beurteilungsdimensionen und Beispiele. Der Mathematikunterricht 34, Heft 4, 19-42
[7] *Borovcnik, M./Ossimitz, G.* [1987]: Materialien zur Beschreibenden Statistik und Explorativen Datenanalyse. Wien: Hölder-Pichler-Tempsky
[8] *Chambers, J. M. et al.* [1983]: Graphical Methods for Data Analysis. Belmont: Wadsworth
[9] *DIFF* [1982]: Wahrscheinlichkeitsrechnung und Statistik unter Einbeziehung von elektronischen Rechnern: Beschreibende Statistik. Studienbrief. Tübingen: Deutsches Institut für Fernstudien
[10] *Hills, V. L.* [19973]: Statistics and social science. In: Giere, R. N. & Westfall, R. S. (Hrsg.): Foundations of Scientific Method: The Nineteenth Century, Bloomington: Indiana University Press, 206-233
[11] *Tukey, J. W.* [1977]: Exploratory Data Analysis. Reading: Addison-Wesley
[12] *Wilson, E. B. jr.* [1952]: An Introduction to Scientific Research. New York: Mc Graw-Hill

Software
[1] *Datadesk V 3.0:* 1989, Odesta Corporation. Inc. 4084 Commercial Ave., Northbrook, Illinois, 60062 (für Apple Macintosh Computer)
[2] *Statview SE+Graphics:* 1988, Abacus Concepts, Inc., 1984 Bonita Ave., Berkeley, Calif. 94704 (für Apple Macintosh Computer)