P. Naeve, D. Trenkler und H.P. Wolf

Sehen und Glauben
Über Experimente in der Statistikausbildung

Der nachstehende Satz gehört sicher zu den Sätzen, die jedem angehenden Statistiker recht bald begegnen:

> **Satz:** *Sei F die Verteilungsfunktion einer kontinuierlichen Zufallsvariablen. Ist U gleichverteilt auf $[0,1]$, so hat $X = F^{-1}(U)$ die Verteilungsfunktion F.*

Er und auch seine Umkehrung bestechen den Neuling bestimmt zuerst durch die »Einfachheit« des Beweises. Aber vielleicht ist es gerade diese Einfachheit, die es dem Neuling häufig so schwer macht zu verstehen, daß mit diesem Satz das Instrumentarium geliefert wird, um das Problem der Erzeugung von Zufallszahlen zu lösen.

Das nachstehende Bild ist eines aus einem ganzen Film, in dem die in dem Satz enthaltenen Konzepte (Inverse Verteilungsfunktion, Gleichverteilung, Stichprobe, ...) in Aktion gezeigt werden. Der Film ist das Ergebnis eines Experimentes, in dem der Satz in jeweils unterschiedlichen Situationen (Verteilung F, Stichprobenumfang n, usw.) angewandt wird.

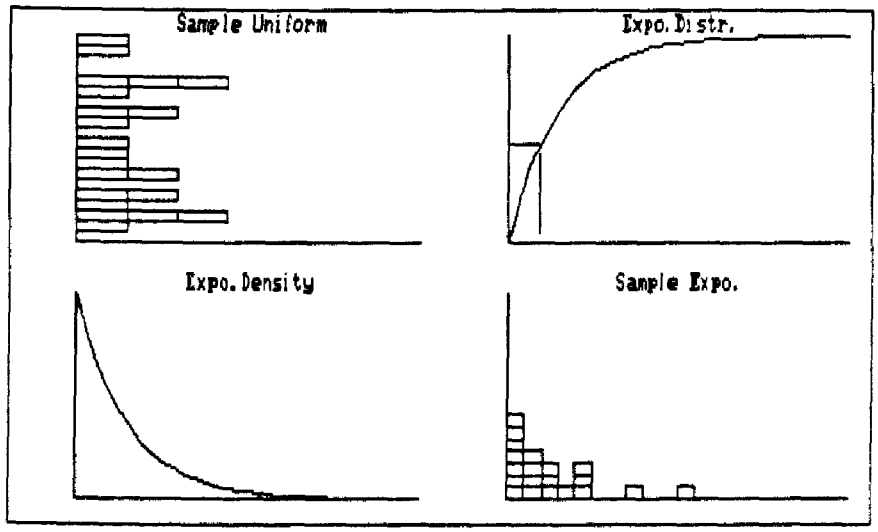

Abb. 1: Erzeugung exponentialverteilter Zufallsvariablen

Auch dieser Miniausschnitt läßt erkennen, daß der Film (das Experiment) dazu beiträgt zu sehen, was in dem Satz enthalten ist.

Wir unterscheiden grundsätzlich zwei Typen von Experimenten, die in der Statistikausbildung von Nutzen sind. Der eine Typ – nennen wir ihn *Theorie-Verständnis-Experiment* – dient zur Veranschaulichung, zur Verständnisvermittlung von Konzepten und Prinzipien der Statistik. Das gerade kurz vorgestellte Experiment ist von diesem Typ. Der andere – hier als *Strategie-Experiment* bezeichnet – ist gedacht zum Verstehen, Lernen von statistischen Strategien. Wir haben an anderer Stelle diese beiden Typen näher beschrieben und durch Beispiele erläutert, wobei dort die Bezeichnung Typ-1- bzw. Typ-2-Experiment verwendet wurde (Naeve/Trenkler/Wolf 1991).

In dieser Arbeit geht es uns nicht vorrangig um die Motivation für Experimente in der Statistik. Diese als gegeben vorausgesetzt, geht es uns hier vor allem um einen Versuch, so etwas wie die Kunst des guten Experiments zu entwickeln. Dies führt, unserer Meinung nach notwendige allgemeine Betrachtungen zu Experimenten eingeschlossen, zu der folgenden Gliederung unseres Papiers:

- Nicht nur historische Bemerkungen zur Rolle des Experiments in den Wissenschaften
- Experimente und Statistik
- Regeln für die Kunst des Experimentierens
- Ein gelungenes Experiment

Nicht nur historische Bemerkungen zur Rolle des Experiments in den Wissenschaften

Wie angekündigt, wollen wir uns in diesem Abschnitt kurz mit der Rolle von Experimenten in den modernen Wissenschaften befassen. Es ist schon fast ein Gemeinplatz, daran zu erinnern, daß mit Galileos Experimenten zur Erforschung der Bewegungsgesetze am schiefen Turm zu Pisa Wissenschaft, so wie wir sie heute praktizieren, begründet wurde. Das grundlegende Paradigma sei hier mit den Worten Jevons formuliert (1877: L): »All inductive investigation consists in the marriage of hypothesis and experiment.«

Wissenschaftlicher Fortschritt beruht also auf Experimenten. Jeder Wissenschaftler muß sich daher mit der Kunst des Experimentierens auseinandersetzen. Denn wie Jevons ausführt, ist die Zahl der Probleme groß. Es beginnt mit der Gefahr des (falschen) Glaubens, daß man mit der hypothetisch-deduktiven Methode zu logisch gesicherten Schlüssen kommen kann. Aus der Übereinstimmung von empirischer Evidenz mit der Konklusion der Hypothese folgt eben nicht die Gültigkeit der Hypothese, da aus der empirischen Evidenz eben nicht die Hypothese logisch gefolgert werden kann. Weiter wissen wir in der Regel nicht, ob die empirischen Daten den ganzen Umfang der relevanten Daten umfassen. Nicht untersuchte Fakten können also immer noch in einem Widerspruch zur Hypothese stehen. Last but not least sind empirische Daten fast nie mehr als eine Approximation der angenommenen Gesetze. Die zentrale Rolle des Experiments für die

wissenschaftliche Arbeit und die hier nur kurz beleuchteten Schwierigkeiten beim Experimentieren machen es verständlich, warum zumindest in der Ausbildung eines Naturwissenschaftlers die Einführung (das Lernen) in die Kunst des Experimentierens einen großen Raum einnimmt. Jevons (1877) verwendet ein ganzes Kapitel auf die Darstellung der Schwierigkeiten beim Experimentieren.

Aber auch aus ganz anderen Gründen sind Experimente wichtig. Warum wohl werden bekannte Experimente in der Schule immer wieder noch einmal durchgeführt? Niemand zweifelt an der Gültigkeit der Bewegungsgesetze. Dennoch wird Jahr auf Jahr der schiefe Turm von Pisa in das Klassenzimmer geholt, und viele kleine Galileos machen die alten Experimente zum n-ten Male – für sie ist es aber das erste Mal! Und diese Experimente verhelfen ihnen zu Einsichten. Sie sehen nun auch, was sie (hoffentlich) vorher intellektuell eingesehen hatten.

Um den knappen Platz für die Hauptsache zu verwenden, seien die wichtigsten Argumente für Experimente kurz (noch einmal) zusammengefaßt.

1. Viele Menschen tendieren dahin, eher Dinge zu glauben, die sie sehen oder anfassen können. (Man denke nur an die Reaktionen auf das Schild: *Frisch gestrichen!*) das »Sehen wissenschaftlicher Gesetze« kann mit Hilfe von Experimenten ermöglicht werden.
2. Wissenschaftliche Gesetze sind in der Regel Abstraktionen von beobachtbaren Fakten. Die Visualisierung des Unterschieds der idealtypischen Situation, wie sie im Gesetz beschrieben wird, und der Realität eines Datensatzes, der gemäß diesem Gesetz generiert wurde, kann mit Experimenten geleistet werden.
3. Neue Gesetze zu formulieren bedeutet oft, diese Gesetze sich »formen« zu sehen bei der Analyse verschiedener, aber verwandter Situationen. Auch dies kann durch Experimente betrieben werden.

Experimente und Statistik

Wie steht es nun um die Rolle des Experimentes in der Statistik? Nun, da ist zuerst das – leider oft in anderen Disziplinen gar nicht ausreichend bekannte – Bemühen der Statistiker, sich mit den geschilderten Schwierigkeiten des Experimentierens auseinanderzusetzen und Lösungen anzubieten. An erster Stelle muß hier auf den von Sir R.A. Fisher gegründeten Bereich »Experimental Design« hingewiesen werden. Die Fülle an unmittelbar praktisch nutzbarer, gleichwohl theoretisch fundierter Literatur hätte sicher das Herz von Jevon höher schlagen lassen und dem erwähnten Kapitel über die Probleme beim Experimentieren ein ganz anderes Aussehen verliehen. Auch dem Statistiker so geläufige Überschriften wie »Varianzanalyse« stehen für ein Füllhorn an nützlichen Werkzeugen für den geplagten Experimentator. So gesehen können wir Statistiker mit erhobenem Haupte im Kreis der (uns leider gar nicht dankbaren) anderen Wissenschaftler verweilen.

Wie aber sieht es mit Experimenten in der Statistik aus? Da haben wir die eher anekdotischen Hinweise auf die Münzwurfexperimente von Buffon, De Morgan und Pearson – hatten *die* Zeit! Es gibt wohl auch noch einige von uns, die einmal Experimente auf einem Galton-Brett in der Vorlesung sahen. Aber sonst ist nicht viel zu berichten. Dabei gelten natürlich auch alle bisher gebrachten Argumente und Ausführungen in der Statistik, sie alle lassen sich auf die Statistik anwenden. An anderer Stelle (Naeve/Trenkler/Wolf 1991) wurde dies breiter ausgeführt. Das ist beileibe keine neue Erkenntnis. Schon Jevons hatte sie. Lassen wir daher Jevons mit einem etwas längeren Zitat zu Wort kommen: »I have made a series of experiments in a third manner, which seemed to me even more interesting, and capable of more extensive trial. Taking a handful of ten coins, usually shillings, I threw them up time after time, and registered the numbers of heads which appeared each time. Now the probability of obtaining 10, 9, 8, 7, &c., heads is proportional to the number of combinations of 10, 9, 8, 7, &c., things out of 10 things. Consequently, the results ought to approximate to the numbers in the eleventh line of the Arithmetical Triangle. I made altogether 2048 throws, in two sets of 1024 throws each, and the number obtained are given in the following table[1]:

Character of Throw				Th.N.	S. 1	S. 2	Ave.	Diverg.
10	Heads	0	Tails	1	3	1	2	+1
9	"	1	"	10	12	23	17.5	+7.5
8	"	2	"	45	57	73	65	+20
7	"	3	"	120	129	123	126	+6
6	"	4	"	210	181	190	185.5	−25.5
5	"	5	"	252	257	232	244.5	−7.5
4	"	6	"	210	201	197	199	−11
3	"	7	"	120	111	119	115	−5
2	"	8	"	45	52	50	51	+6
1	"	9	"	10	21	15	18	+8
0	"	10	"	1	0	1	0.5	−0.5
Totals				1024	1024	1024	1024	−1

The whole number of single throws of coins amounted to 10x2048, or 20,480 in all, one half of which or 10,240 should theoretically give head. The total number of heads obtained was actually 10,353, or 5222 in the first series, and 5131 in the second. The coincidence with theory is pretty close, but considering the large number of throws there is some reason to suspect a tendency of favour of heads.

The special interest of this trial consists in the exhibition, in a practical form, of the results of Bernoulli's theorem, and the law of error or divergence from the mean to be afterwards more fully considered. It illustrates the connection between

1) Dabei wurden die folgenden Abkürzungen verwandt: Theoratical Number (Th.N.), First Series (S. 1), Second Series (S. 2), Average (Ave.) und Divergence (Diverg.).

combinations and permutations, which is exhibited in the Arithmetic Triangle, and which underlies many important theorems of science.«

Besser kann man ein Theorie-Verständnis-Experiment nicht beschreiben.

Regeln für die Kunst des Experimentierens

Bevor das Experimentieren in der Statistik munter beginnen kann, müssen die Ziele genauer beleuchtet werden, um hieraus Anforderungen an Experimente zu formulieren. Dieses soll im folgenden geschehen.

Die Ausbildung in der Statistik erfordert eine Vermittlung von Konzepten der Statistik wie auch das Einüben des Umgehens mit diesen. Voraussetzung für den Erfolg ist also Verständnis, das sich dann in einem angemessenen Umgang zeigt. In dem Bewußtsein, daß bei späteren Aufgaben das »Tun« im Vordergrund steht, ist es naheliegend, Elemente der Aktion – Experimente – in die Ausbildung zu integrieren.

Hierdurch sind zwei Wege vorgezeichnet: Betrachten wir zuerst die Phase der Vorstellung von Konzepten. Hier können Mißverständnisse die nächsten Schritte eines Lernenden erheblich behindern. Deshalb muß das Bemühen groß sein, das richtige Verständnis zu erzeugen. Ein Medium, das es gestattet, allgemeine zentrale Erkenntnisse oder Sätze anhand konkreter (Daten-)Situationen zu untersuchen, kann hierbei sehr hilfreich sein – ein Medium, das es gestattet zu experimentieren. Durch Experimentieren mit konkreten Fällen wird das Verständnis für Konzepte der Statistik überprüft und dadurch zusätzlich vertieft. Diese Fälle werden bisweilen als realisierte Musterbeispiele in der Erinnerung haften bleiben. Als Ziel ergibt sich, Experimente zu ersinnen, die wesentliche Zusammenhänge der Statistik zum Ausdruck bringen und das richtige Verständnis für statistische Theorie untermauern. Solche Experimente sollen *Theorie-Verständnis-Experimente* genannt werden.

Angemessener Umgang mit Konzepten der Statistik offenbart sich in gelungenen Datenanalysen. Hier liegt der Ausgangspunkt nicht mehr in einem Konzept, sondern in einem Datensatz begründet. Konzepte der Statistik werden zu Bausteinen der Analyse. Ziel der Auseinandersetzung ist, statistische Strategie begreifbar zu machen: Was ist wann, wo, wie, warum zu tun? Zur Beantwortung dieser Frage dienen *Strategie-Experimente*. Es drängt sich fast auf, wohlüberlegte Situationen zur Analyse anzubieten und Hilfestellung bei Vorgehensfragen in Form eines Plans zu geben. Auch hier können durchgeführte Analysen als Musterbeispiel überdauern. Damit die Kreativität der Entdecker nicht behindert wird, müssen Freiheiten für die Ausgestaltung der konkreten Analyse vorhanden sein. Andererseits sollte ein Rahmenplan als Orientierungspunkt existieren. Dieser Ansatz darf nicht mit dokumentierten Fallstudien verwechselt werden, die nur noch passiv betrachtet werden können.

Durch diese beiden kurz skizzierten Extremsituationen ist ein riesiges Feld abgesteckt, in dem Experimente entworfen werden können. Es ist an der Zeit, eine Arbeitsdefinition für Experimente vorzuschlagen.

> **Definition:** *Ein statistisches Experiment ist ein Plan, um, basierend auf einer zu wählenden konkreten Situation, zu einem Urteil über theoretische Aussagen bezüglich eines Datensatzes zu gelangen.*

Also wird ein Experiment erst durch die Wahl einer konkreten Situation zu einem Beispiel. Vorher ist ein Experiment nur ein Plan, um zu einer Lösung zu gelangen. Ein Experiment verkörpert damit Vielfältigkeit. Gleichzeitig ist es eine wohlüberlegte Prozedur, die bei der Urteilsfindung helfen soll. Das vom Experimentator aufgrund eines realisierten Plans gebildete Urteil kann nicht als Beweis angesehen werden, sondern bleibt – wie alle Aussagen über unsere Welt – mit einer Restunsicherheit behaftet. Ein Experiment unterscheidet sich von einem statistischen Test, der auch eine Prozedur des Vorgehens darstellt, da durch ein Experiment die Restunsicherheit nicht quantifiziert wird. Ein Experiment spiegelt gleichfalls das Dilemma des empirisch arbeitenden Forschers wider und hilft deshalb, datenanalytisches Arbeiten zu erfassen.

Experimente sind grundsätzlich im Rahmen von sehr unterschiedlichen Umgebungen denkbar. So sind Experimente zum Beispiel mit real existierenden Münzen oder aber als reine Gedankenexperimente vorstellbar. Eine herausragende Rolle spielen Experimente, die mit Hilfe eines Rechners bzw. an einem Rechner durchgeführt werden. »Computer-Experimente« ermöglichen es, sonst nur sehr mühsam umsetzbare Pläne zu realisieren. Natürlich liegt dieses an der Eigenschaft von Rechnern, langweilige, zeitraubende Routinearbeiten, die eine experimentierfreudige Ausbildung bisher verhinderten, in Sekundenschnelle zu erledigen.

Doch wo Licht ist, ist auch Schatten! Mit dem Einsatz von Rechnern entstehen neue Probleme, so daß es angemessen erscheint, hierüber eine Diskussion zu entfachen. Die hier folgenden Bemerkungen versuchen, zu dieser Diskussion einen Beitrag zu leisten, indem für eine geordnete Struktur bei Computer-Experimenten geworben wird. Wir hoffen, daß dadurch vor vorschnellen, computergestützten Ad-hoc-Ideen Mißtrauen erzeugt und der Blick von den blendenden Outputs wieder zurück zu dem Wünschbaren und Machbaren gelenkt wird.

Zwei verschiedene Personengruppen lassen sich unterscheiden. Eine Gruppe – im folgenden *Designer* genannt – entwirft Pläne, bzw. Experimente. Die zweite Gruppe – im folgenden als *Experimentatoren* bezeichnet – wählt die konkrete Situation aus und realisiert die Pläne. Es ist zu vermuten, daß die Gruppen unterschiedliche Interessen und einen unterschiedlichen Wissensstand aufweisen. Die Designer müssen viel über Statistik wissen, um geeignete Pläne zu entwerfen. Das Implementieren dieser Pläne in einen Rechner erfordert natürlich mehr chirurgische Kenntnisse als das Betreiben der entstehenden Experimente. Demnach gibt es Gemeinsamkeiten:

Forderung 1: *In der geschlossenen Welt eines Experimentes sind drei Phasen zu unterscheiden:*
- *die Phase vor der Durchführung des Experimentes*
- *die Phase während der Durchführung eines Experimentes*
- *die Phase nach der Durchführung eines Experimentes.*

Die verschiedenen Eigenschaften dieser drei Phasen müssen den Designern und den Experimentatoren bewußt sein.

Die erste Phase ist die Vorbereitungsphase, die dritte die Nachbereitungsphase eines Experimentes. Ein Experiment lenkt die Konzentration auf die zweite Phase. Trotzdem kann diese erst dann beginnen, wenn die erste erfogreich abgeschlossen worden ist. Was ist für die erste Phase zu beachten?

Forderung 2: *Die durch das Experiment zu beantwortende Frage muß klar gestellt sein.*

Ohne Mangellage benötigt man kein Experiment! Andererseits hängt eine Frage nie in der Luft, sondern setzt immer ein (auch statistisches) Vorwissen voraus.

Forderung 3: *Die (statistische) Wissensbasis,*
die für die Durchführung eines Experimentes erforderlich ist, muß klar definiert sein.

In der Vorbereitungsphase muß eine Vorstellung vom Ablauf des Experimentes ausgebildet werden. Der Experimentator muß also den Plan kennen, mit dessen Hilfe eine Antwort gesucht wird.

Forderung 4: *Die Semantik eines Experimentes muß bekannt sein.*

Der Experimentator muß also um den Inhalt und die Bedeutung der Experimentkomponenten wissen, so daß er seine Beobachtungen angemessen deuten kann. Damit der Einstieg in die zweite Phase gelingt, dürfen keine sich auf die Syntax beziehenden Fragen offen bleiben. Syntax soll auch alle Aktionsmöglichkeiten umfassen, die der Experimentator während der Durchführung besitzt:

Forderung 5: *Die Syntax eines Experimentes muß bekannt sein.*

Diese Forderungen bedeuten für den Experimentator, daß er sich für die Durchführung des Experimentes angemessen vorzubereiten hat, und für den Designer, daß er neben der Erschaffung und Implementierung des Plans, eine große Dokumentationslast zu tragen hat. Dabei sollte nicht zuviel implizit bleiben. Für die zweite Phase ergibt sich als zentrale Forderung:

Forderung 6: *Ein Experiment muß das gestellte Problem klar und direkt angehen.*

In Erinnerung an den »data ink ratio« von Tufte (1983) sollte die zur Diskussion stehende Frage ungestört von Nebensächlichkeiten angegangen werden und nicht in einem Gewirr von Outputs untergehen. So macht es keinen Sinn, dort Beweise durch Experimente zu ergänzen oder gar zu ersetzen, wo Experimente für den Experimentator viel komplizierter sind als die theoretischen Ableitungen selbst. Ein Experimennt, das zeigt, daß die Stichprobenvarianz nie negativ ist, wäre von fraglicher Bedeutung. Damit ein Experiment sich deutlich von einem Lehrbuchbeispiel abhebt, ist es erforderlich, das besondere Wesen eines Experimentes als einen Plan mit unterschiedlichen Ausprägungsmöglichkeiten zu berücksichtigen. Als Forderung formuliert erhalten wir:

Forderung 7: *Ein Experiment muß für einen Bereich von Situationen durchführbar sein.*

Dieses bedeutet zweierlei. Sofern als Input ein Datensatz dient, muß das Experiment mit verschiedenen strukturähnlichen Datensätzen durchführbar sein. Hierdurch kann eine Frage in verschiedenen Datensituationen überprüft werden. Zweitens sollte die Chance bestehen, wesentliche Parameter, die den Plan charakterisieren, zu modifizieren. Mit Hilfe einer solchen »Situations-Variablen« erreicht man eine Allgemeinheit des Plans, die den Begriff »Experiment« gerechtfertigt erscheinen läßt.

Mit diesen Bemerkungen ragt die Diskussion bereits in die dritte Phase hinein. Betrachtet man die Nachbereitungsphase, so wird sie geprägt sein von der Reflexionsarbeit des Experimentators. Er wird seine Vorstellungen überarbeiten (bezüglich der Statistik, des Datensatzes, ...), und er wird neue Ideen und Fragen entwickeln. Letztere können dahin gehen, das Experiment in der gewählten Konkretisierung zu wiederholen. Oder um die Sensitivität des Ergebnisses festzustellen, kommt der Wunsch nach Änderung einiger Situations-Variablen auf. Ist die letzte Forderung angemessen erfüllt, wird sich der Experimentator auf einer bunten Wiese voller Möglichkeiten tummeln können.

Dieses klingt leichter als es ist, denn der Designer muß hierzu die Idee der Experimentatoren vorausahnen können. Hierbei sind Grenzen gesetzt. Deshalb wird sich ein kreativer Experimentator trotz des sorgfältigsten Designs eingeengt vorkommen. Ein zweiter Punkt der Kritik ist dadurch gegeben, daß der Lernende möglicherweise die Ergebnisse mit seinen Vorstellungen nicht in Einklang bringen kann. Solche Mißverständnisse oder »breakdowns« (Winograd/Flores 1986) können in dem Unwissen des Experimentators begründet sein, mögen aber auch auf Kommunikationsprobleme, an denen der Designer beteiligt ist, zurückzuführen sein. In solchen Fällen kann der Experimentator die entstandene Problemlage selbst als Forschungsgegenstand nehmen und analysieren. Hier kann nur die Konsequenz für Experimente zu beleuchten sein. Um den Experimentator nicht zu behindern, müssen die Experimente offen gestaltet sein:

Forderung 8: *Experimente müssen offen und zugänglich und lesbar sein* (readability)

Lesbarkeit bedeutet mehr als die technische Chance, irgendwelche Codefetzen erhaschen zu können. Vielmehr setzt Lesbarkeit voraus, daß die Experimente aus Sprachelementen zusammengesetzt sind, die für den Experimentator zugänglich sind. Experimente können immer nur aus Elementen einer formalen Sprache, die einer festen Syntax genügen muß, bestehen. Deshalb wird es nie möglich sein, diese Elemente direkt aus der Alltags-Sprache des Experimentators zu bilden. Dennoch sollten für die Konstruktion von Experimenten Elemente eines hohen Abstraktionsniveaus verwendet werden, das der alltäglichen Sprache nahekommt:

> Forderung 9: *Ein Experiment sollt aus »high level«-Sprachelementen aufgebaut sein, die der Sprache des Statistikers so nahe wie möglich kommen.*

Folglich darf ein Experiment zur Mittelwertberechnung nicht »X37« heißen; auch dürften Passagen der Form:

```
{
float sum=0;
int i=1;
for(;i<=n;i++)
sum=sum+x[i];
xq=sum/n;
}
```

ausscheiden, so daß nur Vorschläge wie

MEAN X

in Einklang mit Forderung 9 sind. Diese Forderung ist letztlich das Kondensat einer Binsenweisheit aus den Computerwissenschaften. Trotzdem kann sie nicht oft genug wiederholt werden. Es sei angemerkt, daß jedes Sprachelement wieder aus Sprachelementen zusammengesetzt ist, die wiederum von einem wohlüberlegten Abstraktionsniveau stammen sollten. Durch wiederholte Anwendung des Gedankens kommt man zu einer Schichtenvorstellung für den Aufbau der Experimente. Dabei darf der Schritt von Schicht zu Schicht nicht zu hoch sein, um so einen Analytiker, der seinen »breakdown« auflösen will, nicht ins Bodenlose fallen zu lassen.

Aus dem Aspekt der Lesbarkeit von Experimenten läßt sich leicht eine an dieser Stelle letzte Forderung anfügen – hierdurch wird gleichzeitig die nächste Potenz erreicht:

> Forderung 10: *Experimente sollen durch Austausch von Sprachelementen weiterentwickelbar sein* (writeability)

Wenn ein Experimentator im Rahmen der vorüberlegten Veränderungsmöglichkeiten seine Kreativität noch nicht ausgereizt und erkannt hat, welche Elemente an welchen Stellen geeignet sind, warum sollte ihm ein Austausch der Elemente

verwehrt werden? Warum sollte er sich nicht auch in der Sprache ausdrücken dürfen, in der die Experimente formuliert sind? Zugangsbegrenzung erzeugt Einengung, Offenlegung schafft Möglichkeiten.

Es soll noch einmal betont werden, daß die Möglichkeiten abgestuft angeboten werden müssen. Für den Newcomer muß ein sicheres Terrain existieren, für den Fortgeschrittenen muß es möglich sein, seine Fesseln abzuwerfen und einen Schritt in Richtung »Designer« zu tun. Die Erwartung ist hoch, daß durch Experimente das Vergnügen der Designer an ihren netten Experimenten auf die Experimentatoren und späteren Designer vererbt wird.

Ein gelungenes Experiment

Im letzten Abschnitt dieses Papiers geht es darum, die vorangegangenen Ausführungen an einem Beispiel zu konkretisieren. Ausgewählt wurde ein Experiment zum Theorie-Verständnis, das anhand einer Fragestellung im Zusammenhang mit dem χ^2-Anpassungstest verschiedene Aspekte statistischen Experimentierens verdeutlichen soll, die eingangs abstrakter formuliert wurden. Ein Beispiel für ein Strategie-Experiment findet man bei Neave/Trenkler/Wolf 1991.

Das Folgende läßt sich zwei Ebenen zuordnen. Einerseits wird ein Problem in einer einem Statistiker teilweise vertrauten Form abgehandelt, andererseits sind in den Text Meta-Überlegungen eingestreut, um eine Verbindung zu den allgemeinen Ausführungen über Experimente herzustellen. Bemerkungen auf der Meta-Ebene sind zur Verdeutlichung *kursiv* geschrieben.

Technische Momente sollen nicht in allen Einzelheiten behandelt werden, so daß zum Beispiel Fragen zur Syntax – Forderung 5 – hier nicht erörtert werden. Außerdem wird nicht auf die bedeutsamen Forderungen 9 und 10 eingegangen. Zu ihrer angemessenen Behandlung wäre eine umfangreichere Beschreibung der verwendeten Sprachelemente unumgänglich.

Statistische Inferenzen basieren häufig auf der Möglichkeit, Verteilungen durch andere zu approximieren. Dieses wird deutlich am Satz von DeMoivre-Laplace. In der Ausbildung bereitet die Formel:

$$P(X = x) \approx \Phi\left[\frac{x + 0.5 - np}{\sqrt{np(1-p)}}\right] - \Phi\left[\frac{x - 0.5 - np}{\sqrt{np(1-p)}}\right]$$

in der Regel keine größeren Schwierigkeiten, wenn sie anhand eines Vergleichs von einer zu einem Histogramm modifizierten Dichte einer $N(np, np(1-p))$-Verteilung und der entsprechenden Wahrscheinlichkeitsfunktion einer $B(n,p)$Verteilung erläutert wird. Die folgende Abbildung zeigt die Wahrscheinlichkeitsfunktion f einer $B(20,0.4)$-Verteilung und ein modifiziertes Histogramm $\hat{f}$, welches im Intervall $[x - 0.5, x + 0.5]$ die Höhe

$$\hat{f}(x) = \Phi\left[\frac{x + 0.5 - 8}{\sqrt{4.8}}\right] - \Phi\left[\frac{x - 0.5 - 8}{\sqrt{4.8}}\right], \quad x = 0,1,\ldots,20$$

aufweist.

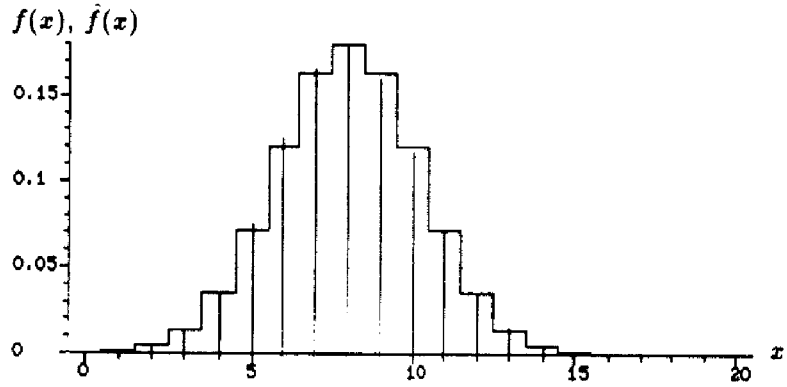

Abb. 2: Approximation einer B(20,0.4)- durch eine *N*(8,4.8)-Verteilung.

Ein weiteres Beispiel stellt der χ^2-Anpassungstest dar, welcher vermutlich eines der am häufigsten angewendeten statistischen Verfahren ist. Hier fällt das Verständnis schwerer, da die Realisierungsmöglichkeiten für

$$T = \sum_{j=1}^{k} \frac{(N_j - e_j)^2}{e_j}, \quad e_j = np_j$$

nicht unmittelbar deutlich werden. Die Ursache ist darin zu sehen, daß das stochastische Modell in einer Einführungsveranstaltung selten angesprochen wird. Der Vektor $(N_1,\ldots,N_k)$ hat eine Multinomialverteilung mit Parametern n und $p_j = P(A_j)$, so daß gilt

$$P(N_1 = n_1,\ldots,N_k = n_k) = \binom{n}{n_1,\ldots,n_k} \prod_{j=1}^{k} p_j^{n_j}$$

Dieser Ausdruck ist sicherlich beeindruckend, zumal die Zahlen $n_1,\ldots n_k$ mit $\sum n_j = n$ nicht leicht zu bestimmen sind.

Zusätzlich herrscht oftmals große Unsicherheit darüber, ob die Anwendung des χ^2-Anpassungstests überhaupt hinreichend gerechtfertigt ist. Insbesondere die Frage, inwieweit die Prüfgrößenverteilung von T durch eine χ^2-Verteilung mit $k-1$ Freiheitsgraden approximiert werden kann, wird in der Literatur uneinheitlich behandelt.

Kreyszig (1965: 229) weist darauf hin, daß im Fall von $e_j < 5$ für ein j »das Ergebnis des Tests mit Vorsicht« zu verwenden ist. Rohatgi (1984: 628) behaup-

tet, daß die Approximation auch dann brauchbar sein kann, wenn ein $e_j \approx 1$ ist. Lindgren (1976: 424) entnimmt man: »No simple answer can be given, but it has been found that when the sample size is, say four or five times the number of cells, the approximation is rather good, even if some of the expected frequencies $n\pi_j$ are quite small (as small as 1, or smaller).«

Damit ist eine Motivation für ein Experiment gegeben. Das gewählte Beispiel hat den Vorzug, daß der theoretische Hintergrund jedem Statistiker geläufig ist. Damit erübrigen sich weitere Ausführungen zur Angleichung der Wissensbasis an dieser Stelle – siehe Forderung 3.

Im folgenden wird versucht, die Problematik mittels eines statistischen Experiments zu beleuchten, indem untersucht wird, unter welchen Umständen die Approximation gültig ist. Konkret soll danach gefragt werden, welchen Einfluß die Parameter n und p_j besitzen.

Damit ist die Untersuchungsfrage formuliert – Forderung 2. Zur Vorbereitungsphase gehört es auch, den Plan des Experiments zu erklären (vgl. auch Forderung 1).

Welche Werkzeuge bieten sich an, um die Approximationsgüte zu verdeutlichen? Maßzahlen lassen sich erst mit viel Erfahrung deuten. Naheliegend ist der direkte Vergleich zwischen den Verteilungsfunktionen, also zwischen $P(T \leq t)$ und $P(\chi^2_{k-1} \leq t)$. Konkret betrachten wir den Fall $n = 10$ und $p_1 = p_2 = p_3 = p_4 = 0.25$. Die vorstehende Tabelle zeigt die Realisationen t von T für $P(T \leq t)$ und $P(\chi^2_3 \leq t)$.

t	$P(T \leq t)$	$P(\chi^2_3 \leq t)$
0.4	0.1442	0.0598
1.2	0.2804	0.2470
2.0	0.5688	0.4276
3.6	0.7394	0.6920
4.4	0.8331	0.7786
5.2	0.8908	0.8423
6.8	0.9485	0.9214
7.6	0.9629	0.9450
8.4	0.9821	0.9616
10.0	0.9836	0.9814
10.8	0.9887	0.9871
11.6	0.9970	0.9911
13.2	0.9983	0.9958
16.4	0.9994	0.9991
17.2	0.9999	0.9994
22.8	1.0000	1.0000
30.0	1.0000	1.0000

Eine solche Tabelle ist mühsam zu begutachten. So erinnert man sich an: »There is no single statistical tool that is as powerful as a well-chosen graph« (Chambers et al. 1983). Ein Scatterplot der durch diese Tabelle definierten Punkte hat das folgende Aussehen.

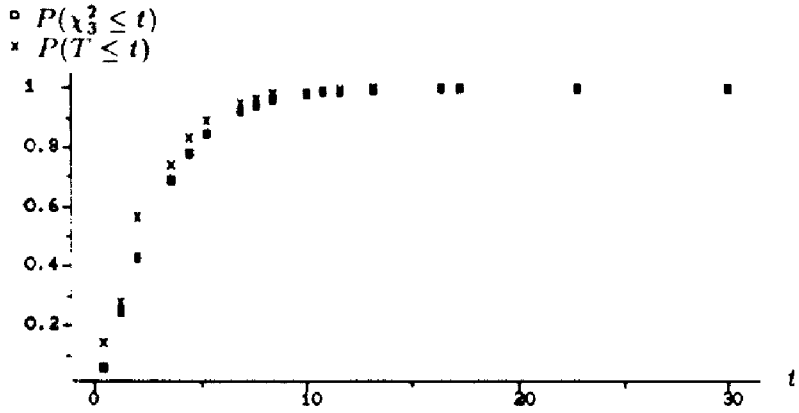

Abb. 3: Vergleich von Verteilungsfunktionen

Ziel an dieser Stelle ist es, eine dem Problem gerechte Graphik zu entwerfen – Forderung 6.

Um die Qualität der Approximation herauszustellen, befriedigt der Plot nicht ganz. Eine Alternative ist durch einen P-P-Plot gegeben (Fisher 1983). Dieser erscheint jedoch zur Beantwortung der hier interessierenden Frage immer noch nicht geeignet. Es geht darum, den Fehler zu erkennen, den man begeht, wenn statt mit der exakten Verteilung mit deren Approximation gearbeitet wird. Deswegen bietet es sich an, Approximation und Differenz aus Approximation und exakten Werten darzustellen.

Die folgende Modifikation scheint den Bedürfnissen besser zu entsprechen. Sei X eine diskret verteilte Zufallsvariable mit Verteilungsfunktion F und Y eine stetig verteilte Zufallsvariable mit Verteilungfunktion G. Es wird der Zusammenhang $F(x) \approx G(h(x))$ postuliert, wobei h eine geeignet gewählte Funktion ist. Beispielsweise ist im Zusammenhang mit der Approximation der Binomialverteilung an die Standardnormalverteilung

$$h(x) = (x + 0.5 - np)/\sqrt{np(1 - p)}$$

zu verwenden. Zur Beurteilung der Approximationsgüte kann ein Plot von $H(x) = F(x) - G(h(x))$ gegen $A(x) = G(h(x))$ dienen. Je besser die Anpassung ist, desto weniger weichen Punkte von der Geraden $y = 0$ ab. Für die Punkte $(A(x), H(x))$, die über der Geraden liegen, gilt $G(h(x)) < F(x)$.

Damit ist der Plan für ein Experiment vorgezeichnet: Wähle einen passenden Satz von Parametern und erstelle daraufhin den beschriebenen modifizierten P-P-Plot – Forderung 4.

Das Experiment besteht nun darin, den folgenden Plot von $(A(t), H(t))$ mit $A(t) = P(\chi_3^2 \leq t)$ und $H(t) = P(T \leq t) - P(\chi_3^2 \leq t)$ zur Beurteilung zu erstellen. *Dieser Plot ist das Resultat des Experiments.*

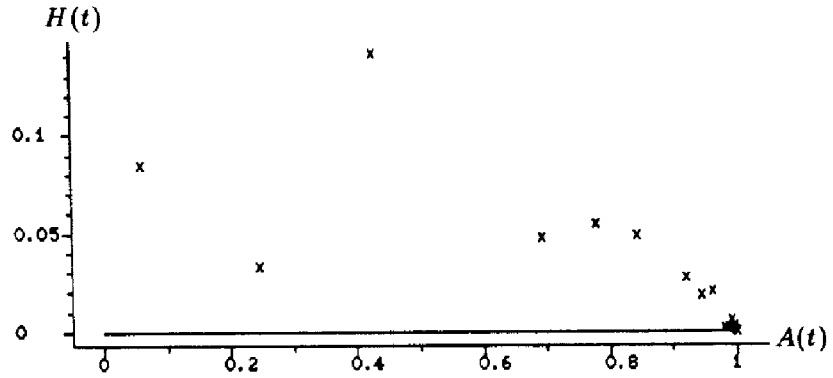

Abb. 4: Plot von H(t) gegen A(t) für $p_1=p_2=p_3=p_4=0.25$, $n=10$

Es folgt die Nachbereitung des Experiments. Wir sehen, daß fast immer $P(T \leq t) > P(\chi_3^2 \leq t)$ gilt, der χ^2-Test also dann konservativ ist, und daß die Approximation für $A(t) > 0.9$ durchaus brauchbare Resultate liefert.

Es war naheliegend, den symmetrischen Fall $p_1 = p_2 = p_3 = p_4 = 0.25$ zuerst zu untersuchen. Wie aber steht es mit anderen Konstellationen? Ergeben sich dann auch so zuverlässige Ergebnisse? Ist der Ausgangspunkt eine konkrete Testsituation, so wird die Wahl der Parameter demgemäß ausfallen. Eine zweite Motivation könnte in dem Wunsch nach Überprüfung von Literaturempfehlungen liegen.

Ideengenerierung ist ein wichtiger Bestandteil der Nachbereitungsphase. Es zeigt sich, ob für die ganz naheliegenden Folgefragen das Experiment auch eine Antwort liefern kann. Hier gilt es, das Experiment mit neuen Parameterwerten zu starten – Forderung 7.

Deshalb wird das Experiment nun noch einmal mit den Parameterwerten $p_1=p_2=p_3=0.1$, $p_4=0.7$ durchgeführt. Hierbei gilt nämlich $e_1=e_2=e_3=1$, so daß gegen die Empfehlung Kreyszigs (1965) verstoßen wird. Wir erhalten wieder einen modifizierten P-P-Plot.

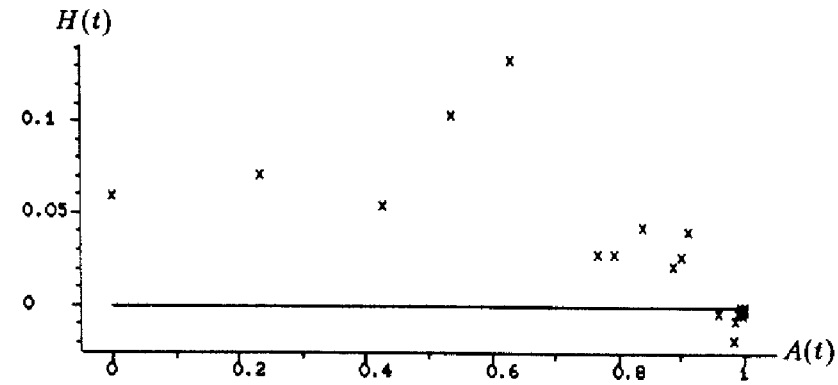

Abb. 5: Plot von H(t) gegen A(t) für $p_1=p_2=p_3=0.1$, $p_4=0.7$, $n=10$

Hier gibt es Fälle, wo $P(T \leq t) < P(\chi_3^2 \leq t)$ ist, jedoch sind dann die Differenzen gering. Für Approximationswerte $A(t) > 0.9$ ist die Approximation sehr gut, denn die Fehler liegen unter 3 %.

Als Zwischenergebnis können wir festhalten, daß der χ^2-Anpassungstest zu besseren Ergebnissen führt als das auf Grund der Empfehlungen zu erwarten gewesen wäre.

Mit wachsenden Werten für k oder n wird die Berechnung der exakten Verteilung von T immer aufwendiger. Sollen jedoch gerade solche Parameterkonstellationen untersucht werden, wird sich der Experimentator nicht mit der Antwort *Parameter unzulässig!* abspeisen lassen. Als Ausweg können Simulationen herangezogen werden.

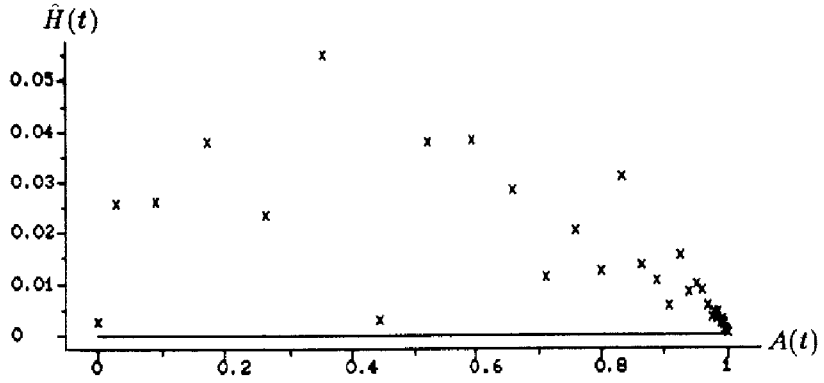

Abb. 6: Plot von $\hat{H}(t)$ gegen $A(t)$ für $p_1=p_2=p_3=p_4=p_5=0.2$, $n=20$

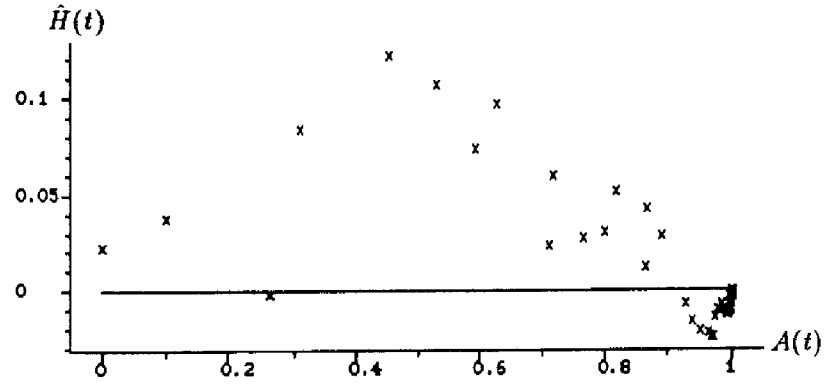

Abb. 7: Plot von $\hat{H}(t)$ gegen $A(t)$ für $p_1=p_2=p_3=p_4=0.05$, $p_5=0.8$, $n=20$

Um einen Eindruck hiervon zu geben, zeigen die vorstehenden Abbildungen einen Plot von $\hat{H}(t) = \hat{F}(t) - P(\chi_4^2 \leq t)$ gegen t, wobei $\hat{F}(t)$ eine Schätzung der Verteilungsfunktion von T ist, die auf 10.000 simulierten Werten basiert.

Ein Vergleich der Abbildungspaare 5 und 7 sowie 4 und 6 wirft die Frage auf, welchen Einfluß die Symmetrie bzw. Unsymmetrie der Verteilung auf das Verhalten des χ^2-Anpassungstests haben. Schon werden weitere Experimente angeregt.

Literatur

Chambers, J.M. et al., 1983: Graphical Methods for Data Analysis, Duxbury Press, Boston.

Fisher, N.I., 1983: Graphical Methods in Nonparametric Statistics: A Review and Annotated Bibliography. In: Int. Stat. Rev. 51: 25–58.

Jevons, W.S., 1877: The Principle of Science, a Treatise on Logic and Scientific Methods, Dover publ. 1958.

Kreyszig, E., 1965: Statistische Methoden und ihre Anwendungen, Vandenhoek & Rupprecht, Göttingen.

Lindgren, B.W., 1976: Statistical Theory, 3. Auflage, Collier Macmillan, New York.

Naeve, P./Trenkler, D./Wolf, H.P., 1991: How to Make the Teaching of Statistics Roar. In: Computational Statistics Quarterly, Vol. 6, Iss. 4: 325–353.

Rohatgi, V.K., 1984: Statistical Inference, John Wiley, New York.

Tufte, E.R., 1983: The Visual Display of Quantitative Information, Graphic Press, Cheshire.

Winograd, T./Flores, F., 1986: Understanding Computers and Cognition, Ablex Publishing Corporation, New Jersey.