Interaktive Statistik

MANFRED BOROVCNIK, KLAGENFURT

Modellbildung bereichert das Bild von Mathematik und hat eine formative Kraft; Begriffe werden dadurch in ihrer Bedeutung über die mathematischen Zusammenhänge hinaus noch einmal geprägt und verändert. Experimente und Fallstudien in diesem Beitrag sollen dies unterstreichen. Modellierung ist ein Handwerk, man lernt es durch Tun. Die ausführliche Behandlung der Fallstudien soll die Stärken der Modellierung hervorheben.

1. Die Perspektive der Modellierung

Das englische ‘modelling’ mag etwas mehr aussagen als die deutsche Form ‚Modellierung’ oder ‚Modellbildung’. Vielleicht hat das auch mit Unterrichtskultur oder Philosophie im Allgemeinen zu tun. Die deutsche Philosophie hat sich doch sehr dem Erforschen des „Wie die Begriffe sind“ verschrieben und das „Wozu die Begriffe dienen“ ein wenig vernachlässigt. Modellierung hat sehr viel damit zu tun, *wozu* die Begriffe dienen können. Tatsache jedoch ist, dass Modellbildung zwar in der didaktischen Diskussion einen hohen Stellenwert einnimmt, den Unterrichtsalltag jedoch wenig bestimmt. Unbestritten ist, dass die Kompetenzen der Beteiligten rasch überfordert werden. Man vermittelt aber ein verzerrtes Bild von Mathematik, wenn man Anwendungen und Modellbildung weglässt. Stärker akzentuierte Ansätze zur Mathematik würden den realen Situationen sogar den Vorrang geben und zum Ausdruck bringen, dass die realen Phänomene erst die mathematischen Begriffe wirklich organisieren. In einer wechselweisen Befruchtung liegt sicherlich viel Potential für Mathematiklernen und Verstehen. Blum (2012) führt folgende Begründungen für Anwendungen und Modellbildung ins Treffen:

- *Pragmatisch*: Um Situationen in der realen Welt verstehen und beherrschen zu können.
- *Formativ*: Kompetenzen werden durch Modellierungsaktivitäten wesentlich begründet.
- *Kulturell*: Bezüge zur realen Welt sind unerlässlich für ein adäquates Bild von Mathematik.
- *Psychologisch*: Zur Motivation aber auch zur Strukturierung mathematischer Inhalte.

1.1 Modellbildung in Wahrscheinlichkeitsrechnung und Statistik

Wahrscheinlichkeitsmodelle haben eine große Verbreitung gefunden. Physiker bauen ihre Theorien von kausalen Paradigmen auf den Zufall um. Risiko war und ist ein grundsätzliches Element menschlichen Daseins. Statistische Inferenz hat sich als Standard zur Verallgemeinerung von Schlüssen aus empirischen Daten durchgesetzt. Stochastische Modellierung, Modellbildung und die zugehörigen Prozesse sind daher ein lohnendes Ziel für den Unterricht.

Wir werden die Schritte der Modellierung in empirischer Forschung exemplarisch in Fallstudien illustrieren und dabei auf authentische Daten, die von den Lernenden selbst erzeugt wurden, zurückgreifen. Die Hypothesen, die untersucht werden, ergeben sich natürlich aus dem Zusammenhang. Diese Anbindung hebt einerseits die Motivation, erleichtert andererseits aber auch die Chancen, die jeweils weiteren Schritte der Modellierung und die späteren Ergebnisse besser zu verstehen.

Auch in der Wahrscheinlichkeitsrechnung kann der Modellierungsgedanke mit Vorteil eingesetzt werden; Orientierungen dazu findet man in Borovcnik und Kapadia (2011). Etwa kann man mit Aussagen wie „Die Wahrscheinlichkeit beim Würfeln für einen Sechser IST 1/6“ aufräumen und stattdessen dies als Ergebnis eines Modells hinstellen. Klar, dass auch andere Modelle in Frage kommen. Man wird aber aus Gründen der Einfachheit (Modellökonomie) vielleicht bei dem Modell der Gleichwahrscheinlichkeit bleiben. Etwa deswegen, weil ein feineres Modell, das die Asymmetrien eines speziellen Würfels berücksichtigt, doch nur unwesentlich andere Ergebnisse liefern würde.

Schriftenreihe zur Didaktik der Mathematik der Österreichischen Mathematischen Gesellschaft (ÖMG), Heft 45, 2012, S. 1 – 18

Statt der mechanischen Berechnung von weiteren Wahrscheinlichkeiten aus gegebenen – eventuell unter der Annahme einer bestimmten Verteilung – könnte man den Stellenwert der Wahrscheinlichkeitsrechnung aufwerten, indem man zeigt, dass man damit eine Entscheidungssituation verbessern kann. Etwa wenn man aus mehreren Handlungen eine auswählt, die sich nach Kriterien wie erwartete Kosten oder Wartezeit als am besten begründbar herausstellt. Dazu muss das Modell noch gar nicht immer ganz gut passen. Es reicht oft, die Wirklichkeit in Szenarien nachzuspielen, damit man kritische und entscheidungsrelevante Parameter herausfiltert. Ein Szenario ist dabei ein ‚Modell', das heuristisch verwendet wird, um die Situation auf der Basis ‚was wäre, wenn?' zu explorieren. Es kann sein, dass ein Szenario relevante Einsichten in ein Problem ermöglicht, obwohl es eigentlich nicht perfekt passt (Borovcnik, 2009).

Im Prozess der Modellierung braucht man viel mehr Mathematik als angenommen. Man muss das Problem adäquat erfassen, normalerweise muss man zwischen verschiedenen Modellen abwägen, also die Vor- und Nachteile dieser kennen und bewerten. Auch der Kontext der Probleme erfordert Kenntnisse, sodass es angezeigt ist, die mathematischen Details gelegentlich durch Simulation und animierte Visualisierungen zu ergänzen oder teilweise zu substituieren. Hauptaugenmerk soll auf die Anwendungen und die zugrunde liegenden Ideen gerichtet sein, weniger auf technische Feinheiten. Ziel des Modellierungsprozesses ist, geeignete Modelle zu finden und sie schrittweise zu verbessern. Die Schritte umfassen dabei

- einmal ein vernünftiges Modell zu bekommen, das die Situation adäquat repräsentiert,
- dann innerhalb des Modells eine Lösung abzuleiten,
- und schließlich diese Lösung auf das Ausgangsproblem zu übertragen.

Üblicherweise sind mehrere Zyklen erforderlich, um eine zufriedenstellende Modellierung zu erreichen. Um die Güte eines Modells zu beurteilen, sind Übersetzungsmuster zwischen den Welten nötig. Eines nutzt die externe Struktur des stochastischen Modells und prüft – über einen statistischen Test – die Anpassung der relativen Häufigkeiten an die modellmäßigen Wahrscheinlichkeiten. Eine andere nützt die interne Struktur eines Modells. Hierbei wird der Gedanke einer fundamentalen Idee hinter den Verteilungen wesentlich. Anregungen dazu findet man in Borovcnik und Kapadia (2011).

1.2 Kompetenzen zum Modellieren

Anwendungen sind auch ein Motor zur Entwicklung von Unterrichtseinheiten. Jedoch gilt zu beachten, dass genuine Anwendungen sich für den Unterricht nur bedingt eignen. Die Anforderungen an Lehrer als auch an Schüler[1] sind sehr vielfältig. Wissen über den Kontext ist wesentlich. Man muss mehrere mathematische Modelle miteinander vergleichen, will man die Situation ausreichend erfassen. Die Arbeitsweisen schließen die Organisation der Zusammenarbeit und die Übernahme von Ergebnissen anderer Gruppen mit ein. Technologie ist unerlässlich und hilfreich, verlangt aber auch einen Grad an Vertrautheit damit.

Um die Ideen und die Arbeit zu fokussieren, hat man Projektarbeit oder projektorientierten Unterricht vorgeschlagen. Dabei werden Fragen entwickelt, man arbeitet in unterschiedlichen Gruppen, findet partielle Antworten und benutzt bestimmte mathematische Verfahren, über die man sich oft erst orientieren muss. In wirklich authentischen Anwendungen ist die anfängliche systemanalytische Phase jedoch ziemlich vage; oftmals zu vage, um weiter zu kommen. Viele Projekte scheitern in dieser Phase. Das kann man im Unterricht nicht riskieren. Da sind gezielte Fallstudien besser geeignet. Wesentlich für den Fortgang der Arbeit ist die Herausarbeitung von ‚Forschungsfragen' in den frühen Phasen eines Projekts, auf die man sich in der darauf folgenden Arbeit entsprechend beziehen kann. Ein sorgfältig gewählter Kontext trägt zur Motivation bei und hilft bei der Interpretation von Ergebnissen und bei den Entscheidungen über die weitere Vorgangsweise. Ein Ziel zu formulieren ist entscheidend für die straffere Ausrichtung der Anstrengungen.

1.3 Modellieren mit Lehrern

Der Autor setzt Modellieren, Experimente und Fallstudien seit Jahren in der Lehrerfortbildung ein. Ein wesentliches Merkmal eines interaktiven Zugangs zur Stochastik ist die zyklische Arbeitsweise. Zur Motivation und zur leichteren Interpretation der Arbeitsschritte und der Ergebnisse ist es förderlich, wenn die Hypothesen sich aus dem Kontext direkt ergeben und Interesse erwecken. Während der einzelnen Schritte der Datenanalyse ergeben sich dann Zwischenergebnisse, welche erste Ansätze verändern lassen und neue Gedanken einbringen.

Im Folgenden wird ein Gedächtnistest eingehend analysiert. Die Ergebnisse stammen aus einem Workshop, der von John O'Donoghue an der Universität Limerick mit 46 Lehrern organisiert wurde (27 Frauen, 19 Männer). Neben diesem Test wurden dabei "Spaghetti brechen", "Motivation in Wettbewerben" und "Placebo-Effekt und Regression zur Mitte" behandelt. In der Mathematik gilt es als Regel, dass eine Aufgabe eine (richtige) Antwort hat. In der Statistik gibt es einen Spielraum und üblicherweise ist die Frage nach richtig oder falsch falsch gestellt. Während sich diese Sicht doch verbreitet hat, ist es dagegen für viele völlig überraschend, dass die Modellierung auch für Wahrscheinlichkeitsprobleme letztlich Fragen offen lässt. Diese Offenheit kann nicht nur bei Lehrern Schwierigkeiten verursachen. Letztlich geht es auch um die Frage, wie sich die Autorität von Lehrern in der Klasse ausprägt.

Die durchwegs positiven Reaktionen nähren die Hoffnung, dass bekannte Hindernisse für Modellierung im Unterricht bewältigt werden können. Gruppenarbeit wurde für effizient befunden, die Rahmenbedingungen und der Kontext der Fallstudien trugen dazu bei, die Ideen fließen zu lassen und die Möglichkeiten zu explorieren. Technologie war wichtig zur Bestandsaufnahme, zur Analyse der Daten und zur graphischen Gestaltung der Ergebnisse, welche die weiteren Aktivitäten wesentlich beeinflussten.

Die anhaltende Konzentration der Teilnehmer und ihr Engagement zeigten von Interesse am Ansatz, wie man auch an der Videoaufzeichnung ersehen kann. Die Lehrer beurteilten die Materialien als authentisch, interessant und geeignet. Allerdings sind für echten Unterricht mit Schülern entsprechende Anpassungen wohl noch zu machen. Abschließend äußerten sich die Lehrer optimistisch über die Möglichkeiten, den Ansatz auf ihre Schulklassen zu transferieren. Der experimentelle Ansatz wird insgesamt trotz seiner Herausforderungen als vielversprechend eingeschätzt.

2. Fallstudien in Modellbildung

In der Statistik gibt es mindestens zwei Ansätze: Der eine besteht darin, die Ergebnisse einer Stichprobe gegen Hypothesen zu vergleichen, das führt auf statistische Tests und Vertrauensintervalle. Der andere ist, Daten zu ‚produzieren' – in der Form von so etwas wie Stichproben – um zu untersuchen, wie eine Zielvariable von Einflussfaktoren abhängt; das Ziel kann dabei sein, diese Zusammenhänge zu beschreiben, Einsicht in die dahinter stehenden Prozesse zu gewinnen und sie eventuell später zielgerichtet zu beeinflussen.

Wir illustrieren die dabei verwendeten Methoden anhand von Fallstudien, wobei der statistische Modelleur in den Prozess des Modellbildens miteinbezogen wird; die Hypothesen beziehen sich auf diesen, die Daten sind persönlich verankert und die Ergebnisse werden auf ihn selbst bezogen. Das erhöht die Motivation der Lernenden und bietet gleichzeitig mehr Einsicht in die statistischen Methoden und die daraus gewonnenen Aussagen. Es zeigt auch den interaktiven Prozess des Entstehens neuer Vermutungen, wie er durch Zwischenergebnisse beeinflusst wird, und erleichtert die Interpretation der Schlussfolgerungen, die gleichzeitig weitere Fragen aufwerfen. Dies zeigt, wie Statistik zur Erweiterung empirisch begründeten Wissens – Schlagwort ‚evidence-based knowledge' – beiträgt.

2.1 Ein Gedächtnistest

Das Experiment besteht aus der Anforderung, Items, die hintereinander vorgestellt wurden, korrekt wiederzugeben. Die zu untersuchende Variable ist Erfolg – wie viele Items kann man richtig angeben? Gemäß den zwei Zugängen zur Statistik stehen zwei Wege offen zur Analyse.

Der eine vergleicht die Ergebnisse mit denen anderer Gruppen oder mit Ergebnissen aus psychologischen Studien. In der Tat hat Miller (1956) aus vergleichbaren Experimenten ein Gesetz der magischen 7 herausgefiltert; danach können sich Leute ca. 7 Items aus vollständig zusammenhanglosen Items merken, mit einer Marge von plus oder minus 2. Dieses psychologische Gesetz wird nun als Aufhänger benutzt, um die Gruppe zu motivieren, ihr Bestes zu geben. Dieser Strang führt in den Kern der beurteilenden Statistik. Ein Vorteil der Versuchsanordnung besteht darin, dass die Hypothese direkt – aus dem Kontext heraus – verständlich ist und gleichzeitig die Neugierde geweckt wird, der Analyse zu folgen, um eine Antwort auf die Frage zu bekommen: „Sind wir denn nun besser als ein solches Gesetz vorhersagt?“

Der andere untersucht potentielle Faktoren, welche den Erfolg im Gedächtnistest beeinflussen. Dies führt zunächst in die Statistik als Systemanalyse: Was sind entscheidende Faktoren, welche den Erfolg beeinflussen? Solche Faktoren können sich auf die Items selbst, auf die Personen oder auf die Testsituation als solche beziehen. Können wir die Niveaus oder Werte dieser Faktoren verlässlich messen und wie können wir die Ergebnisse unserer Untersuchungen verallgemeinern? Wesentlich ist dabei die Frage nach Störgrößen, welche den Einfluss der Faktoren mitbestimmen, welche aber nicht erfasst sind, sodass sie weder kontrolliert (ausgeschlossen) noch erfasst werden können. Als Folge solcher Störgrößen könnten die Ergebnisse verzerrt werden und daher keinerlei allgemeine Schlüsse aus dem Experiment zulassen.

Störgrößen können sich auch auf den beurteilenden Teil von oben auswirken. So ist es viel leichter, die einmal gemerkten Items aus dem Gedächtnis abzurufen, wenn sie entgegen der Voraussetzung doch irgendwie zusammenhängen. Als Folge würde die Gruppe viel besser sein im Vergleich zum Gesetz (zur Hypothese, gegen die man vergleicht). Die Interpretation, dass die Gruppe besser wäre, ist aber nicht schlüssig – auch wenn das Ergebnis signifikant ist – denn das gute Abschneiden wäre dann wohl dem leichteren Test als der größeren Fähigkeit der Gruppe zuzuschreiben. Die Suche nach potentiellen Störgrößen (auch Confounder genannt) ist daher von fundamentaler Bedeutung für die Angewandte Statistik und muss in der Phase der Systemanalyse, d. h., *vor* der ‚Produktion’ der Daten, sorgfältig abgehandelt werden. Fehlen nämlich Daten dazu, kann man sie normalerweise nicht mehr nachverfolgen, wenn später in der Analyse ein Verdacht nach Confoundern auftaucht.

2.2 Vergleichen der Ergebnisse mit Hypothesen oder anderen Gruppen

Wir beginnen mit der beurteilenden Statistik, war doch dies das ursprüngliche Motiv, den Kontext des Tests der Gruppe vorzustellen. Das Ziel dabei ist zweierlei: i. Zu hinterfragen, ob unsere Gruppe besser als die Vergleichshypothese ist. ii. Methoden für einen solchen Vergleich zu entwickeln, die noch dazu leicht zugänglich sind, was den Bezug zur inhaltlichen Situation noch verstärkt.

Wie gut sind Leute allgemein und wie gut sind wir im Memorieren von Items? Wir könnten die Ergebnisse mit anderen Gruppen vergleichen. Ein wichtiger Vergleichsmaßstab ist der Leistungspegel einer allgemeinen Gruppe. Wir nutzen hier mit Vorteil die Ergebnisse von Psychologen aus den 1950ern (Miller, 1956): Menschen können 7 Items, die *keinerlei* Zusammenhang aufweisen, korrekt aus dem Kurzzeitgedächtnis abrufen. Natürlich kann es sich als schwierig erweisen zu beurteilen, ob diese Voraussetzung der fehlenden Zusammenhänge für die verwendeten Items zutrifft. Solche Zusammenhänge zu rekonstruieren ist ein wichtiges Element für Lernen. Ein weiteres Ergebnis der Forschungen ist, dass die Zahl der angebotenen Items die Zahl der richtig aus dem Gedächtnis abgerufenen Items nur wenig beeinflusst – es macht kaum etwas aus, ob das 15 oder gar 100 sind.

Wir entschieden uns für 15 Worte – nicht Buchstaben oder Zeichen – auch um das Experiment motivierender zu machen (die Liste der Worte ist aus den Abbildungen weiter unten ersichtlich). Jedes der Worte wurde eine Sekunde lang auf die Leinwand projiziert, dazwischen gab es eine kurze Pause. Die Anleitung war klar, keine Notizen waren erlaubt, das Ziel war, sich möglichst viele zu merken, um anschließend die Frage zu analysieren, ob wir besser als das Gesetz der magischen 7 sind; damit war auch der Ehrgeiz der Teilnehmer angestachelt. In unserem Experiment folgen wir einem Vorschlag von Richardson und Reischman (2011).

Die Analyse unserer ‚Forschungsfrage' wurde erst mit informellen Mitteln vorangetrieben, wir bezogen dann zunehmend formalere Methoden in die Arbeit ein: zuerst wurde ein Stamm-und-Blatt-Diagramm beurteilt, dann wurden einzelne Parameter berechnet und interpretiert, schließlich versuchten wir, den Vorzeichentest einzuführen, um die Resultate von vorher zu erhärten.

Sind wir besser als das magische Gesetz der 7?

Wenn wir den Spielraum plus oder minus 2 des Gesetzes der 7 im Stamm-und-Blatt-Diagramm hervorheben, sehen wir eine deutliche Verschiebung nach oben gegenüber diesem Vergleichsmaßstab (Abb. 1):

Nur eine Person liegt unterhalb, 19 sind darüber; 56 % passen zur Regel. Das ist ein deutliches Indiz, dass unsere Gruppe besser ist! Die zentrale Tendenz der Leistung (Mittelwert 8,93, Median 9,00) fällt mit der oberen Grenze aus dem Gesetz, das ist 7 plus 2, zusammen. Bei einer Standardabweichung von 2,37 können wir sagen, dass wir das Gesetz der 7 beinahe um eine Standardabweichung nach oben überragen.

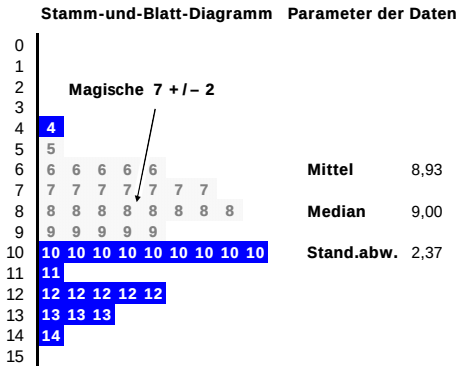

Abb. 1: Stamm-und-Blatt-Diagramm der Daten.

Dennoch, wir müssen die Aussage noch besser abstützen, denn bislang haben wir die Größe unserer Gruppe außer Acht gelassen; wir wissen aber, dass – zumindest wenn es um die Schätzung unbekannter Wahrscheinlichkeiten geht – der Spielraum des Zufalls lediglich mit $1/\sqrt{n}$ kleiner wird. Die Größenordnung hier ist schwieriger herzuleiten, aber der Hinweis zieht: wir brauchen weitere Analysen.

Sind wir besser als hypothetische Daten auf der Basis des *Modells* der magischen 7?

Wir verfeinern die Methoden, indem wir die Frage „Was können wir unter dem Gesetz der magischen 7 erwarten?" drehen und wenden. Damit betreten wir das Reich der Hypothesen. Die Hypothese, die sich hinter dieser Aussage verbirgt, ist eine über die Fähigkeit, die Worte korrekt aus dem Gedächtnis abzurufen, welche nun eine zufällige Variable wird. Wir könnten diese durch eine Normalverteilung modellieren und einen *t*-Test anwenden. Unser Zugang ist aber informeller.

Wenn wir die 7 des Gesetzes als Median unserer (zufälligen) Fähigkeit auffassen, können wir jeder Person eines der drei Attribute zuordnen: *höher*, *niedriger* oder *gleich* der 7. Diejenigen, die genau 7 Items korrekt wiedergegeben haben, tragen zur Entscheidung über die Frage, ob wir *besser* oder *schlechter* als das Gesetz der 7 sind, nichts bei – es liegt nahe, sie einfach wegzulassen. Damit vereinfachen wir die Modellsituation: wir stellen nur mehr fest, ob eine Person *über* oder *unter* der 7 liegt. Die Forschungsfrage lautet nun: Haben wir mehr Personen über der 7 als unter den Bedingungen des Gesetzes der 7 zu erwarten sind? Was können wir denn erwarten, oder wie können wir ein Modell entwickeln, um zu beurteilen, was wir erwarten können, auch unter extremeren Fällen, die von Zeit zu Zeit ja doch passieren? Können wir einen Schwellenwert in diesem Modell angeben, welcher die Trennung markiert zwischen dem normalen Spielraum der Fluktuation und den zu extremen Fällen?

Die wesentliche Annahme für den Vergleich unserer Gruppe mit dem Gesetz lautet nun: Können wir unsere Gruppe als zufällige Stichprobe aus einer Verteilung für unsere theoretische Fähigkeitsvariable auffassen, welche zudem einen Median von 7 hat? Wenn das der Fall ist, so kann das Modell durch wiederholtes Münzwerfen mit einer gewöhnlichen Münze (mit p = ½ für Kopf) verkörpert werden. Die Zuordnung von *über* (1) und *unter* (0) der magischen 7 erfüllt die Voraussetzungen einer Bernoulli-Kette, insbesondere sind die Versuche unabhängig.

Kommen wir auf die Frage zurück: „Sind wir besser als die magische 7?“ Wir vergleichen die Daten unserer Gruppe, *als ob* sie unter den Bedingungen des entworfenen Modells entstanden wären. Mit derselben Wahrscheinlichkeit ½ ist eine Person über wie unter dem Median von 7. Die Wahrscheinlichkeit für die Anzahl der *über* folgt daher einer Binomialverteilung mit n = 39 (wir lassen die 7 Personen weg, welche genau 7 Worte richtig wiedergeben konnten) und p = ½. Der erwartete Wert in diesem Modell ist 19,5; dagegen haben wir 32 beobachtet. Das ist ganz weit weg; wir wollen nun die Logik ausbauen um zu beurteilen, *wie* weit das weg ist.

Wir stellen die Basis für den Vergleich und die Berechnung der Schwellenwerte bereit

Mit Hilfe des Binomialmodells präzisieren wir, was erwartet werden kann. Wir wissen, dass der Erwartungswert nur eine Richtzahl ist, was passieren kann; er bezeichnet das Zentrum und gibt keinerlei Aufschluss über die zufällige Variation. Wir berechnen stattdessen ein zentrales Intervall von möglichen Ergebnissen, wenn die Hypothese zutrifft. Die Abbildung 2 (links) zeigt, wie weit der Wert unserer Gruppe draußen liegt; das wirft erhebliche Zweifel am Zutreffen der zugrunde liegenden Annahmen auf. Die Hypothese ist als Wahrscheinlichkeitsverteilung formuliert, was eine entsprechende Interpretation von Wahrscheinlichkeiten erfordert.

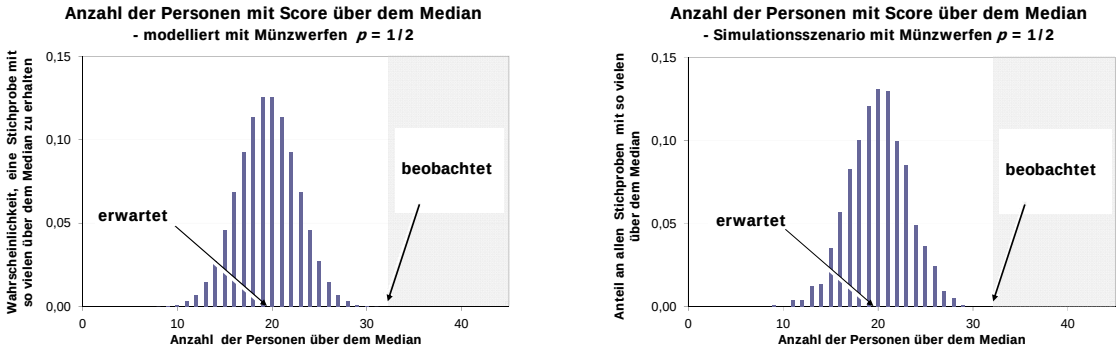

Abb. 2: Links. Potentielle Ergebnisse. Rechts. Simulierte Ergebnisse – beide unter der Hypothese des Münzwerfens.

Weit anschaulicher ist es, diese Wahrscheinlichkeit auf wiederholte Studien zu beziehen. Was passiert, wenn die Annahme des Münzwerfens immer zutrifft und wir öfter eine solche Studie machen? Wir simulieren eine Studie durch Simulation von 39 Daten und bestimmen damit die Anzahl der Personen über dem Median 7. Dann wiederholen wir das Szenario sehr oft. Wir erhalten bei 3 000 fiktiven Wiederholungen – siehe Abb. 2 (rechts) – keinen einzigen Wert von der Höhe 32. Das zeigt, wie extrem unser Wert ist.

Wir berechnen aus dem Binomialmodell eine Wahrscheinlichkeit von 0,000 04 dafür. Wir erwarten eine Gruppe mit solch exzellentem Score in etwa in einer von 30 000 vergleichbaren Studien. Noch immer ist so ein Ergebnis möglich, aber wir entscheiden uns dafür, zu sagen, dass unsere Gruppe signifikant besser als das Gesetz der 7 ist. Die Methode ist als Vorzeichentest bekannt; jedes Datum bekommt ein Vorzeichen (plus oder minus), das anzeigt, ob es über oder unter dem angenommenen Median ist, was der definierenden Eigenschaft des Medians entspricht.

Logik eines Signifikanztests

Eigentlich kann man auf der Basis einer Wahrscheinlichkeitsverteilung kaum irgendeinen Wert für eine Stichprobe ausschließen, außer es gibt Bereiche, die gar nicht auftreten können; daher ist es auch umgekehrt nicht möglich, die Daten zu verwenden, um eine Hypothese logisch zu Fall zu bringen. In der Praxis muss man sich aber *für* oder *gegen* ein Modell entscheiden.

Ein üblicher Ansatz geht von der Idee aus, dass ein Zielwert (das Zentrum der Verteilung) durch eine Reihe zufälliger Einflüsse ‚verfehlt' wird. Je weiter der festgestellte Wert im äußeren Bereich der Verteilung liegt, desto größer müssen diese Einflüsse sein. Umso weniger glaubwürdig dann, dass sie tatsächlich auf das Geschehen eingewirkt haben. Je weiter draußen ein Wert liegt, desto eher spricht er also dagegen, dass die Hypothese tatsächlich zutrifft.

Man berechnet daher ein zentrales Intervall in der Verteilung und nimmt eine Beobachtung, die aus diesem heraus fällt, die also über die Schwellenwerte (die Randpunkte des Intervalls) hinausgeht, als Anlass, die Hypothese in Frage zu stellen und sie aus der weiteren Betrachtung auszuschließen. Jedenfalls markiert eine solche Beobachtung ein seltenes Ereignis und das Risiko wäre sehr klein, gerade dies zu beobachten, sollten die Daten unter regulären Bedingungen ‚erzeugt' worden sein. Ein weit draußen liegendes Ergebnis wird also als *Indiz* gegen die Hypothese verstanden und zwar monoton: je weiter draußen, umso stärker ist das Indiz. Liegt solch ein Indiz vor, werden die Daten als empirischer Beleg gewertet, dass die Bedingungen (und damit die zugrunde liegende Hypothese) verletzt worden sind.

Beim Gedächtnistest fanden wir (zumindest) 32 unter den 39 Personen über dem Median; das entspricht 32 Köpfen in 39 Würfen mit einer fairen Münze, was ein sehr seltenes Ereignis ausmacht. Die Hypothese, welche zum Modell des Münzwerfens führte, war: „Die Lehrer sind mit dem Gesetz der magischen 7 im Einklang." Entsprechend verwerfen wir diese Hypothese und stellen stattdessen fest, dass unsere Teilnehmer besser als dieses Gesetz sind. Wir ziehen diese Interpretation dem vor, dass wir nach wie vor daran festhalten, dass uns tatsächlich ein so seltenes Ereignis passiert ist. Das ist die Logik des Signifikanztests.

Der Schluss, wonach die Gruppe besser als das Gesetz der 7 ist, muss in unterschiedlicher Weise in Frage gestellt werden. Warum ist die Gruppe keine zufällige Stichprobe aus der Population mit einem Median von 7? Das kann an den Bedingungen des Experiments oder am Verhältnis Experiment-Personen liegen. Was ist, wenn die Worte (einigen) schon vorher bekannt waren? Was ist, wenn die Teilnehmer während der Präsentation der Worte Notizen gemacht haben. Dann würde die gezeigte Leistung nicht ihre Fähigkeit widerspiegeln, sich die Worte merken zu können. Die Bedingungen des Experiments müssen klar festgehalten und ihre Einhaltung aufgezeichnet werden. Eine weitere Störgröße ist, wenn die Personen voneinander abgeschrieben und damit ihre Leistung kombiniert haben.

Es könnte auch sein, dass zwei Personen wesentlich besser als der Rest der Gruppe sind, welche ganz dem Gesetz entspricht. Das ist hier nicht der Fall, aber würde – falls es zuträfe – eine Zuerkennung von Exzellenz (als Folge der Ablehnung des Gesetzes der magischen 7) für die ganze Gruppe bedeuten, während es eigentlich nur auf diese beiden zutrifft. Da solche Confounder hier ausgeschlossen werden, kommen wir zum Schluss: Die Gruppe der Lehrer kann als signifikant besser als das Gesetz der magischen 7 betrachtet werden.

Findet man Daten, die zu einer signifikanten Abweichung von der Hypothese führen, so wird das häufig überinterpretiert als ein empirischer Beleg, dass diese Hypothese falsch ist und daher das logische Komplement (hier: wir sind besser als die magische 7) *bewiesen* ist. Solch ein signifikantes Ergebnis bedeutet i. A. aber lediglich, dass man danach erst über die Frage im Kontext des Problems nachdenken und nach Gründen suchen muss, warum das so gilt und wie man das Ergebnis besser erklären kann – eventuell auch alternativ durch Verweis auf Störgrößen.

Es bleibt noch die Frage, ob das gefundene ‘Muster’ auf alle Personen heute im Vergleich zu Millers Versuchspersonen 1950 verallgemeinert werden kann. Sind Leute heute generell besser? Eine Klärung ergibt sich aus dem Kontext. Das exzellente Abschneiden kann dadurch erklärt werden, dass sich i. die Lehrer durch den Besuch einer Fortbildungsveranstaltung während der Sommerferien vom Durchschnitt abheben, und ii., dass es sich bei ihnen um so genannte ‚Schneeball’-Lehrer handelt, die speziell ausgewählt wurden, damit sie neue Inhalte als Multiplikatoren an ihren Schulen weiter verbreiten. Das rechtfertigt die Feststellung, dass diese Gruppe als Ganzes für sich in Anspruch nehmen darf, besser als das Gesetz von der 7 zu sein, da sie offenbar geübt ist, Stoff zu memorieren. Gleichzeitig macht diese Beschreibung klar, dass die Gruppe keineswegs als repräsentativ für die heutige Bevölkerung gelten kann und die Ergebnisse daher ebenso wenig auf diese übertragen werden können.

Signifikanztests können am besten als Filter verstanden werden, um den Anteil jener Hypothesen zu erhöhen, über welche es sich vom Kontext der Sache her lohnt nachzudenken.

Epilog

Wenn Modellbildung nicht eine reine Übung für sich bleiben soll, wird man praktische Phasen einschließlich der ‚Datenproduktion’ einschließen. In der Realität werden nicht nur die Voraussetzungen einer zufälligen Stichprobe verletzt – schon in früheren Stufen – gibt es viele Fallen. Fehlende Daten zählen dazu. Die wird es immer geben, noch dazu fehlt meist eine Möglichkeit, sie zu ergänzen. Ein Ansatz hierzu ist, alle statistischen Einheiten mit fehlenden Daten aus der Studie zu entfernen; neuerdings scheint sich durchzusetzen, sie aufgrund mehr oder weniger plausibler Annahmen zu schätzen. Beide Ansätze haben ihre Meriten und Nachteile.

Wir haben die Daten aus dem Gedächtnistest analysiert, indem wir jene Arbeitsblätter, die in diesem Abschnitt leer waren, aus der Betrachtung ausgeschlossen haben. Der Schluss, den wir gezogen haben, war, die untersuchte Gruppe von Lehrern ist signifikant besser als es das Gesetz der 7 besagt. Zur Erinnerung: die mittlere Zahl der richtig erinnerten Worte war 8,93 und wir hatten 39 schlüssige Daten mit 32 über dem Median. die Wahrscheinlichkeit für eine solche extreme Beobachtung (32 und mehr) in einer einzelnen Stichprobe war mit 0,000 04 berechnet worden – das sollte in 30 000 vergleichbaren Studien ein einziges Mal erwartet werden, sollte das Gesetz doch zutreffen.

Wir müssen anmerken, dass 7 der Arbeitsblätter in diesem Abschnitt leer waren (6 weiblich, 1 männlich). Was machen wir mit diesen Personen? Im vorhergehenden Ansatz haben wir sie einfach weggelassen. Was aber ist, wenn zwischen „Nicht Ausfüllen“ und Erfolg ein Zusammenhang besteht? Sind einige der Teilnehmer, die sich der Härte des Konzentrationstests nicht stellen wollten, schlechter als der Durchschnitt? Waren einige der besonders guten zu nervös und wollten ihre Ergebnisse den Kollegen nicht zeigen? Haben einige ihre Ergebnisse nicht abgegeben, weil sie zu schlecht waren? Wir haben keine Ahnung, was wirklich dahinter steckt.

Können wir die Resultate der Studie trotz dieser Unbestimmtheit aufrechterhalten? Können wir einen schlechtesten Fall annehmen und dessen Auswirkungen erfassen? Nehmen wir an, 6 Personen liegen unter dem Median und einer genau bei 7 (fast der schlechteste Fall), dann hätten wir 45 gültige Daten mit 32 über dem Median. Wie zuvor errechnen wir die Wahrscheinlichkeit für solch ein extremes Ergebnis (32 und mehr) mit 0,0033, falls die Hypothese des Münzwerfens zutrifft. Noch immer ist das sehr klein; wir würden solch ein Ergebnis einmal in rund 300 vergleichbaren Studien erwarten.

Als ernster Vorschlag für die Praxis kann gelten, fehlende Daten tunlichst zu vermeiden, indem man Kontrollmechanismen in die Datenproduktion einplant. Trotz Einhaltung solcher Kontrollen sind fehlende Daten ein alltägliches und kritisches Problem. Oftmals müsste man die Ergebnisse eines Projekts deswegen eigentlich wegwerfen. Gelegentlich hilft die Analyse eines schlechtest anzunehmenden Falles: Wir konnten dadurch die ersten Ergebnisse aufrechterhalten, wenngleich sie nun etwas weniger spektakulär erscheinen. Unsere Lehrer sind signifikant besser als nach dem Gesetz der 7.

2.3 Faktoren, welche den Erfolg und das Verhalten im Gedächtnistest beeinflussen

Zielvariable ist der Erfolg, d.h., die Anzahl der korrekt aus dem Gedächtnis wiedergegebenen Worte. Einflussfaktoren zu kennen mag dazu beitragen, effizient zu lernen. Solche gesetzesähnliche Zusammenhänge zu kennen, ist an sich interessant. Im Rahmen einer Systemanalyse erfasst man dazu zunächst die beteiligten Objekte und ihre wechselweisen Abhängigkeiten. Diese Phase ist der eigentlichen statistischen Analyse vorgeschaltet und zeichnet sich durch große Unbestimmtheiten aus; sie ist aber von eminenter Bedeutung für die Aussagekraft der späteren Ergebnisse.

Die Personen: Alter (Kinder, Erwachsene), Geschlecht, Bildung einschließlich Vertrautheit mit Mnemotechniken und dem Einsatz von Gedächtnis überhaupt.

Items. Einzelne Items stehen in einem Kontext (sie mögen darüber hinaus lustig sein oder blutrünstige Assoziationen begünstigten etc.), sind kurz oder lang oder mögen fremdländisch klingen (Worte aus dem Griechischen); auch wenn die Items völlig unverbunden sein sollen, kann es Zusammenhänge zwischen ihnen geben, die mehr oder weniger offensichtlich sind, was dann die Merkbarkeit stark erhöht. Eine Bezugsgröße ist jedenfalls *Zeit*, denn die Items müssen in eine Reihenfolge gebracht werden und sie werden zeitlich nacheinander aus dem Gedächtnis abgerufen.

Testsituation. Die experimentelle Situation mag direkt einen Einfluss auf die Teilnehmer ausüben; sie mögen die Situation ernst nehmen oder können überhaupt kein Interesse aufbringen und geben willkürliche Antworten. Die Tageszeit kann Einfluss haben (nach dem Mittagessen etwa); das Medium der Präsentation (auditiv oder visuell) kann Erfolgschancen verändern.

Zusammenhänge. Es mag Wechselwirkungen zwischen Personen, zeitlichen Mustern oder dem Kontext geben: Frauen konstruieren sinnvolle Verbindungen zwischen den Worten gänzlich anders als Männer; der Kontext der Worte hat unterschiedliche Auswirkung auf Männer und Frauen. Solange wir Daten über solche Merkmale haben, können wir ihren Einfluss auf andere Merkmale und die Zielvariable auf Richtung und Größe hin untersuchen. Fehlen solche Daten, kann ein Einfluss später in der Studie, falls eine entsprechende Vermutung auftaucht, nicht mehr überprüft werden. Solche Faktoren nennt man Confounder. Verabsäumt man, ein Ergebnis gegen potentielle Störgrößen abzusichern, wird es möglicherweise fälschlich verallgemeinert und muss in Nachfolgestudien berichtigt werden.

Wir werden nun den Einfluss des Kontexts der Worte, ihre zeitliche Anordnung und das Geschlecht der Teilnehmer auf ihren Einfluss auf den Erfolg und das Abrufverhalten hin untersuchen.

Hat der Kontext der Worte einen Einfluss auf den Erfolg?

Aus einer Rangfolge der Worte nach der Größe des Erfolgs erhält man als ersten Eindruck, dass der Kontext zählt, weil es beträchtliche Schwankungen der Erfolgsraten gibt (siehe Abb. 1).

Danach führen die Worte *rigging*, *ear*, und *octopus* die Liste an; *seed* und *legend* markieren das untere Ende, wobei *legend* ganz weit abgeschlagen ist. *Octopus* könnte man oben erwarten, für die anderen Worte ist ihr Einfluss weniger strukturierbar. Auffällig ist, dass der innere Teil den weiten Bereich von 36 bis 65 % abdeckt, wobei wenig Gruppierung zu erkennen ist.

Kontext hat einen Einfluss, aber es ergibt sich kein klares Bild. Das ist auch ein Hinweis darauf, dass die Items doch neutral genug gewählt waren, was die Zuverlässigkeit des Gedächtnistests erhöht.

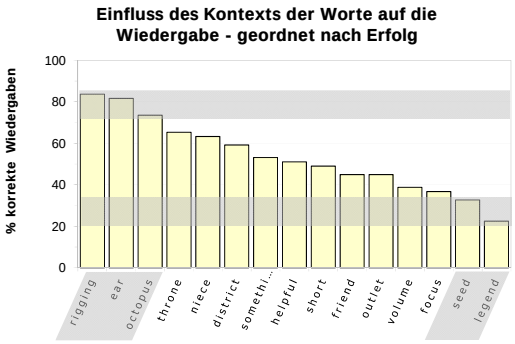

Abb. 3: Rangfolge der Worte nach Erfolg.

Die zeitliche Reihenfolge der Darbietung hat Einfluss auf den Erfolg der Wiedergabe

Es entspricht durchaus dem Hausverstand, dass die zeitliche Reihenfolge der Darbietung der Worte den Erfolg bei der Wiedergabe beeinflusst. Man konnte während des Experiments direkt sehen, wie sich die Teilnehmer konzentrierten, gleich von Anbeginn an. Auch ganz plausibel, der letzte Eindruck ist ein bleibender. Wir wollen diese natürlichen Hypothesen anhand der Daten überprüfen.

Tatsächlich zeigt der Verlauf des Prozentsatzes korrekter Wiedergaben eine klare Abhängigkeit von der Zeit der Präsentation. Mittlere Zeitpunkte weisen die niedrigsten Erfolgsraten auf.

Dasselbe Diagramm mit Tiefe als unabhängiger Variablen würde die angesprochene Abhängigkeit noch deutlicher zeigen. Dabei ist mit Tiefe der Darbietung die Entfernung von Anfang und Ende gemeint: Eine Tiefe von 2 bedeutet, das entsprechende Wort ist an der Stelle 2 oder 14 auf der Leinwand erschienen. Die Darstellung unterbleibt aus Platzgründen; wir kommen auf die Tiefe weiter unten noch einmal zurück.

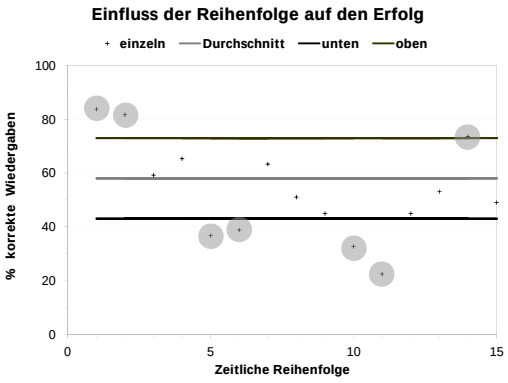

Abb. 4: Erfolgsrate der Wiedergabe in Relation zur zeitlichen Darbietung.

Solche Ergebnisse stehen nicht nur im Einklang mit dem Hausverstand sondern können auch durch psychologische Faktoren erklärt werden. Die anfängliche Konzentration wird zur Mitte hin fallen; die letzten Worte heben sich einfach besser ab und lassen sich besser merken.

Beeinflusst der Zeitpunkt der Präsentation die Zeit des Abrufens aus dem Gedächtnis?

Merken sich die Teilnehmer die Worte sequentiell in der Reihenfolge, in der sie präsentiert werden und – als Folge davon – geben sie die Worte beim Abrufen aus dem Gedächtnis in genau dieser Reihenfolge (allenfalls mit Lücken) wieder? Gibt es beim Memorieren genauso wie bei Körperbewegungen im Sport eine Richtung der – geistigen – Bewegung?

Die Reihenfolge, in der die Teilnehmer die Worte aus dem Gedächtnis abgerufen haben, wurde durch die Anordnung der Worte auf dem Arbeitsblatt rekonstruiert. Es sollte einen starken Zusammenhang geben, dennoch könnte dies aus verschiedentlichen Gründen verletzt sein. Da im Experiment dies nicht ausreichend erklärt und bestimmt worden ist – etwa durch ein entsprechend strukturiertes Arbeitsblatt – kann man die Zuverlässigkeit des gewählten Meßverfahrens anzweifeln. Dennoch werden wir den Einfluss der Zeit der Präsentation (zeitliche Ordnung) auf die Zeit, wann die Teilnehmer die entsprechenden Worte aus dem Gedächtnis abgerufen haben, auf diese Weise untersuchen.

Während in der Analyse des Einflusses des Faktors Zeit (der Präsentation) auf den Erfolg der Wiedergabe die Einführung der Variablen Tiefe etwas künstlich erschienen sein mag, verdeutlicht der Bezug auf die Tiefe hier das Muster: Die Zeit des Abrufens der einzelnen Worte hat mit der Zeit der Präsentation nur einen mäßigen Zusammenhang (R = 0,42), dagegen besteht zur Tiefe der Darbietung eine starke Korrelation (R = 0,63); man kann diese man am besten aus einer *standardisierten* Punktwolke (Abb. 5) erkennen. Dabei sind die Achsen so skaliert, dass die Punkte in ein Quadrat passen.

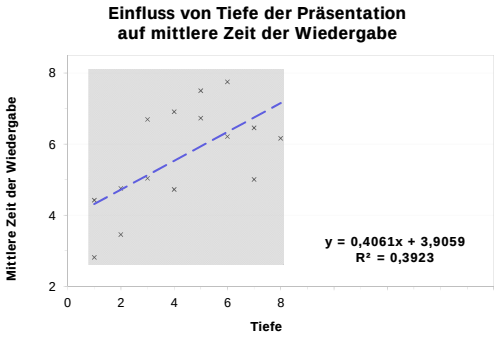

Abb. 5: Einfluss von Tiefe auf Zeit des Abrufens.

Wechselwirkungen zwischen Geschlecht und anderen Einflussfaktoren

Im nächsten Schritt untersuchen wir, ob es Geschlechterunterschiede in der Wiedergabe verschiedener Worte gibt. Mit allen verfügbaren Daten haben wir 25 Frauen und 18 Männer mit einer durchschnittlichen Anzahl von korrekt wiedergegebenen Worten von 9,25 und 8,06, was einem $t = 1{,}54$ entspricht. Obwohl die Frauen um mehr als ein Item besser sind, ist der Unterschied nicht signifikant. Allerdings spielen bei der Beurteilung die fehlenden Daten eine große Rolle, da ja 6 Frauen und nur ein Mann ihr Arbeitsblatt hier leer gelassen haben (bei drei Personen konnte das Geschlecht nicht rekonstruiert werden). Die berichtigte Differenz in der Leistung beträgt 0,36 zugunsten der Frauen, das entspricht einem *t*-Wert von lediglich 0,43 – keine Rede von Signifikanz. Während für alle Daten der Unterschied erstaunlich groß (jedoch nicht signifikant) ist, ergibt der ungünstigste Fall eine vernachlässigbare Differenz von 0,36, die auch gänzlich durch zufällige Schwankung erklärt werden kann.

Beeinflusst der Kontext Männer und Frauen unterschiedlich hinsichtlich des Erfolgs? Welche Worte tragen zur höchsten Differenz bei? Wirkt sich die Zeit der Präsentation unterschiedlich auf das Geschlecht aus? Wird ein solcher Zusammenhang durch den Kontext der Worte verzerrt?

Die Rangfolge nach der Größe der Differenz im Erfolg zwischen Frauen und Männern entnimmt man indirekt aus Abb. 6: Frauen waren einiges schlechter als Männer bei *ear*, *helpful*, *short* und besser bei *friend*, *focus*, *octopus*. Ob die Differenz in der Leistung an der zeitlichen Reihenfolge liegt oder am Kontext der Worte, kann auch anhand derselben Abbildung studiert werden: Frauen sind weniger erfolgreich am Beginn (Tiefe 2, *ear*) und erfolgreicher in der Mitte (Tiefe 7, *friend*). Wir zeigen das Diagramm alternativ mit Tiefe als Einflussfaktor; Abb. 7 zeigt ganz deutlich den geschlechtlichen Unterschied. Ob diese Unterschiede nun tatsächlich vom Kontext oder von der Reihenfolge abhängen, muss man einer Replikationsstudie überlassen. Wir formulieren jedenfalls als Hypothese: Männer und Frauen verfolgen andere Strategien zum Memorieren; dabei orientieren sich Männer eher sequentiell, Frauen stärker am Kontext.

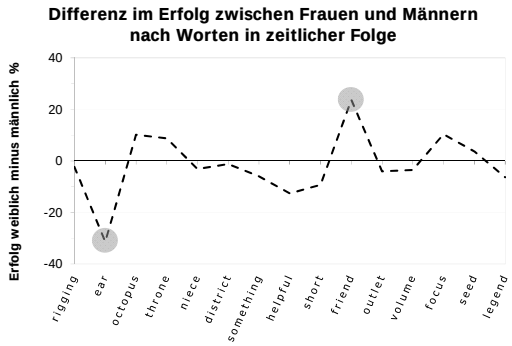

Abb. 6: Geschlechtsdifferenz nach Worten in Serie.

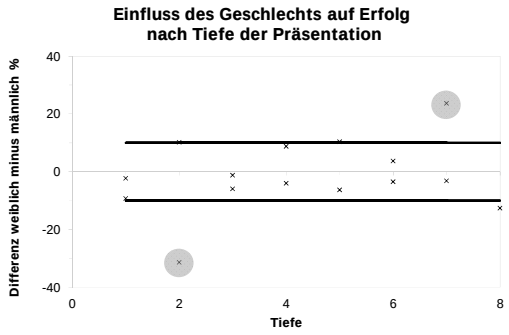

Abb. 7: Wechselwirkungen von Geschlecht und Tiefe auf Erfolg.

Weitere Einflussfaktoren

Im Workshop wurden die teilnehmenden Lehrer auch gebeten, auf ihr Arbeitsblatt eine beliebige dreigliedrige Kennzahl einzutragen. Die Frage ist, ob irgendwelche Eigenschaften dieser Kennzahl einen Zusammenhang mit dem Erfolg beim Abrufen der Worte aus dem Gedächtnis haben. Es war schon ganz gut, dass die Korrelation der ersten Ziffer der Kennzahl nur mäßig ausfiel ($R = 0{,}38$), wobei die Punktwolke auch sonst kein auffälliges Muster auswies. Etwa sinnvoller ist da schon die Frage, ob die Eigenschaft ‚Primzahl' zu sein für die gewählte Kennziffer einen Zusammenhang mit dem Erfolg aufweist. In der Tat, es gibt einen kleinen Einfluss, da aber das Ergebnis nicht signifikant ausfällt, ersparen wir uns nachzudenken, wie und wieso eine Präferenz für Primzahlen einen Einfluss hat auf die Erfahrung oder die Fähigkeit, Worte aus dem Gedächtnis korrekt abzurufen.

Was lernen wir aus den Mustern, die durch die Analyse entdeckt worden sind?

Die Muster, die durch empirische Studien herausgefiltert werden, erheben den Anspruch von gesetzesähnlichen Beziehungen. Man beachte aber, dass normalerweise aus einem angewandten Projekt mehr offene Fragen entstehen als beantwortet werden. Darüber hinaus besteht eine Gefahr, die Ergebnisse überzuinterpretieren. Die Analysen haben interaktiv durch Zwischenergebnisse ihren Fortgang genommen. Insbesondere wurden sie durch die graphische Darstellung mit angeregt. Es ist daher zutreffender von *Hypothesen* zu sprechen, die man als Resultat formulieren kann. Diese warten auf weitere Forschung, einige davon werden zu revidieren sein, andere werden vielleicht erhärtet werden können. Das ist die wesentliche Aufgabe und Ausrichtung empirischer Forschung, welche evidenz-basiertes Wissen anhäuft. In der Praxis aber lohnt es sich selten, Studien zu replizieren: keiner wird dafür die Gelder aufbringen; im Fall der Bestätigung früherer Resultate wird keiner zuhören, im Fall von Widersprüchlichkeiten werden wenige diese zum Anlass nehmen, die ersteren Ergebnisse zu verwerfen, eher wird man nach Mängeln im Design der Replikationsstudie suchen.

Den Abschluss eines Projekts sollte eine Diskussion der Statistiker mit Experten aus dem Kontext des Problems bilden. Diese sollte folgende Punkte tangieren: Eine Evaluation der Ergebnisse; eine Verknüpfung der Ergebnisse mit dem gängigen Wissen der Disziplin; eine Diskussion von möglichen Konflikten der Ergebnisse mit langjähriger Erfahrung; eine Zusammenschau von Fragen, welche durch das vorläufig neue Wissen aufgeworfen werden.

3. Vorschläge für weitere Fallstudien

Wir haben ähnliche Experimente genutzt, um bestimmte Eigenheiten probabilistischer Begriffe und statistischer Verfahren zu illustrieren. Die folgenden Vorschläge für Projekte zeigen einen hohen Grad an Interaktivität und Selbstbezug auf mit jeweils ‚drängenden' Fragen, welche Hypothesen motivieren, die dann auf Klärung warten. Wir beschreiben die Aktivitäten kurz und verweisen auf andere Quellen, wo die Leser mehr Details finden können.

3.1 Spaghettis in drei Teile brechen

Spaghettis brechen ist ein Experiment mit einem überraschenden Ergebnis, weil die relativen Häufigkeiten sehr weit weg von den Modellwahrscheinlichkeiten liegen, was anzeigt, dass etwas schief gelaufen ist. Das führt direkt auf die grundsätzliche Erörterung der Frage „Was ist Zufall?" und zeigt, dass naives Verhalten ohne Präferenz und Zufall nichts gemein haben.

Ein Experiment entwickeln

Das folgende Experiment (Borovcnik und Schenk, 2011) lässt die Annahme der Zufälligkeit explorieren. Ein Stück Spaghetti wird in drei Teile zerbrochen. Erst nachdem alle ihre Spaghettis gebrochen haben, wird aufgefordert, mit den Bruchstücken jeweils eines Spaghettis ein Dreieck zu formen. Üblicherweise erhält man Erfolgsraten von über 70%. Vergleicht man mit nahe liegenden Modellen, so ergeben sich daraus Wahrscheinlichkeiten von 0,25 bzw. 0,3863. Die Modelle kann man durch zwei Zufallszahlen ausdrücken. Modell 1 nützt die zwei Zahlen um die Bruchpunkte auf der Zahlengeraden zwischen 0 und 1 zu markieren. Modell 2 geht hierarchisch vor: erst bricht man nach der ersten Zahl, dann nimmt man einen Teil und bricht diesen nach der zweiten Zahl.

Die Ergebnisse belegen, dass wir die Spaghettis nicht zufällig brechen. Um unser Verhalten besser zu modellieren, beachten wir, dass die meisten für das zweite Brechen das längere Stück wählen und dass man es üblicherweise vermeidet, zu kleine Stücke abzubrechen. Nimmt man dies ins Modell auf, so erhält man – durch Simulation – 0,58. Das Modell ist schon viel besser, es zeigt, dass wir offensichtlich relevante Faktoren für das Brechen entdeckt haben. Aber eben noch nicht alle.

Hintergrund

„Es gibt keinen Zufall" hat de Finetti in seiner bahnbrechenden Arbeit über Wahrscheinlichkeit 1937 festgestellt. Wir fügen hinzu: Zufall ist lediglich eine *Sicht* über die Welt – eine Theorie also. Es gibt natürlich eine starke Identifikation von Zufall mit Situationen, die keinerlei kausale Einflüsse aufweisen, keine Muster, Fairness spielt eine Rolle. Sind Situationen wiederholbar, so bieten relative Häufigkeiten sowohl eine Schätzung als auch ein wesentliches Interpretationsmuster für Wahrscheinlichkeit. Aber, die genauen Bedingungen, unter denen relative Häufigkeiten nützlich sind, sind oft verletzt.

Nicht jeder Prozess ist zufällig und Daten sind nur dann brauchbar, wenn man bei ihrer Entstehung einige Kriterien erfüllt. Sie müssen unabhängig voneinander sein und alle derselben experimentellen Situation (und Verteilung) folgen. Das ist oft verletzt, sodass man – im Sinne eines Modelleurs – dann nicht von Stichproben sprechen kann, auch wenn das immer wieder ins Treffen geführt wird.

Zufall ist nur ein Instrument des *Modellierens*. Sogar Physiker sind im Gebrauch von Zufall sehr schludrig. Sie erklären die Entstehung des Universums durch Zufall und reden davon, dass es keine Kausalität mehr gibt. Dabei vernachlässigen sie, dass ihr Ansatz lediglich eine *Sichtweise* darstellt. In der Tat, es gibt neuerdings Ansätze mit latenten Variablen, welche die Kausalität wieder einbringen.

Wir sollten die Liste, was Zufall ausmacht, erweitern: Zufall assoziieren wir auch mit ‚keine Absicht' oder ‚keine Präferenzen' und erwarten, dass sich dadurch Gleichwahrscheinlichkeit einstellt. Da wir die Spaghettis ohne Absicht brechen, denken viele, dass das Ergebnis zufällig *ist*. Umso größer ihre Überraschung, wenn die Ergebnisse so stark voneinander abweichen. Das ergibt einen kognitiven Konflikt, der zur Klärung des Konzepts beitragen kann.

3.2 Präferenzen für Zahlen

Man kann Menschen nach einer beliebigen Schlüsselzahl mit drei Ziffern befragen. Einmal mehr kann man die Frage untersuchen, ob sich die Wahl der Zahlen wesentlich von zufälligen Experimenten abhebt und man daher von Präferenzen sprechen kann. Eine Inspektion dieser Präferenzen zeigt immer Auffälligkeiten, führt aber mitten in die grundsätzliche Diskussion „Was kennzeichnet ein signifikantes Ergebnis?" oder was meinen wir, wenn ein empirischer Befund als *signifikant* bezeichnet wird.

Hintergrund

Es lohnt sich, zu untersuchen, ob Menschen Zahlen zufällig oder nach Präferenzen wählen. Man könnte Teilnehmer Zahlen so zufällig wie nur möglich wählen lassen. Üblicherweise vermeiden sie längere Runs derselben Zahl und Abschnitte mit ausgeprägtem Muster, besonders in binären Folgen. Wir schlagen eine offenere Aktivität vor und lassen im Arbeitsblatt eine dreigliedrige Kennzahl eintragen. Es ist nicht ganz klar, ob sie diese zufällig wählen oder ihre Lieblingszahl oder einfach ‚irgendeine' Zahl angeben sollen. Die Motive hinter der Wahl sind vielfältig. Wir untersuchen die Kennzahlen deskriptiv und vergleichen dann mit der Zufälligkeitshypothese um zu prüfen, ob man signifikante Abweichungen vom Zufall erkennen kann, was dann rechtfertigt, von Präferenzen zu sprechen.

Einige Ergebnisse

Als Schlüsselzahl nutzen Frauen im Workshop die gesamte Bandbreite aus während Männer die mittleren Zahlen auslassen. Für einzelne Ziffern ist eine Tendenz zu 1en und 3en bei Frauen zu beobachten, wobei 48% auf die Gruppe 1 3 7 fällt; Männer bevorzugen 7 und haben 80% für 7 3 1 (in dieser Rangfolge an Häufigkeit). Die einzelnen Positionen (erste, zweite, dritte) zeigen wenig Abweichung von diesem Muster, lediglich 4 (zweite) und 5 (dritte) spielen bei Frauen noch eine Rolle. Ähnlich zum Gedächtnistest können wir nach Signifikanzen prüfen. Verglichen werden dabei rein zufällige Auswahl aus 000 bis 999 gegen Präferenzen. Die Teilnehmer wählten signifikant mehr (*p*-Werte in Klammern) 1 3 7 – Ziffern (0,0000007), die 7 (0,020), mehr Primzahlen (0,020).

Caveat

Die Ergebnisse haben den Status von gesetzesähnlichen Beziehungen. Ehrenberg (1981) sieht im Herausfiltern solcher ‚Gesetze' und dem Ausloten, unter welchen Bedingungen sie Anspruch auf Gelten haben, eine vornehme Aufgabe der Statistik. Wir müssen für die Gültigkeit der gefundenen Muster noch berücksichtigen, dass wir eigentlich viele Tests gleichzeitig durchgeführt haben und zu den Hypothesen durch erste Analysen aus *denselben* Daten angeregt wurden. Viele Ergebnisse in der empirischen Forschung haben kaum mehr Anspruch auf Verallgemeinerung als unsere über die Präferenz. Zur besseren Absicherung kann man hier eine Wiederholung der Studie empfehlen. Konfirmation durch Replikation ist auch im Einklang mit methodischen Überlegungen in der Wissenschaftstheorie (Popper, 1935), wird aber zu selten eingesetzt.

3.3 Motivation in Wettbewerben

Die Motivation von Menschen wird durch die Anzahl der Mitbewerber stark beeinflusst, so haben Psychologen kürzlich herausgefunden. Zeigt dieses signifikante Ergebnis, dass wir uns irrational verhalten oder können wir dagegen behaupten, dass wir unsere Anstrengungen – zu Recht – auf solche Situationen richten, in denen wir bessere Chancen haben.

Hintergrund

Kürzlich haben Psychologen ein Gesetz über das Verhalten formuliert, wonach Leute sich stärker demotivieren lassen, wenn die Anzahl der Mitbewerber steigt. Aus ihren Experimenten haben Garcia & Tor (2009) diesen so genannten *N*-Effekt herausgefiltert. Der Faktor dahinter ist Motivation und er ist sehr stark durch die Anzahl der ‚Gegner' beeinflusst. Ihre Arbeit hat eine richtige Kontroverse ausgelöst als nämlich Probabilisten die Situation auf der Basis *reinen* Zufalls modellierten und damit zeigten, dass es – im Design von Garcia & Tor – einfach leichter ist, unter den Gewinnern zu sein, wenn es weniger Mitbewerber gibt (Mukherjee & Hogarth, 2010).

Der Kern der Auseinandersetzung ist folgender: Beeinflusst ein psychologischer Faktor (zu Unrecht) unser Verhalten oder agieren die Leute rational, indem sie ihre Bemühungen auf jene Situationen richten, in denen sie leichter zum Erfolg kommen? Tor & Garcia (2010) haben darauf ziemlich wenig überzeugend reagiert. Sie vergleichen einfach die Bereitschaft von Leuten, ihr Bestes zu geben (auf einer geeigneten Skala gemessen), einmal bei wenigen und dann bei vielen Mitbewerbern und stellen fest, dass die Daten einen signifikanten Unterschied zwischen den zwei Bedingungen anzeigen. Die Probabilisten dagegen erstellen ein Modell als Vergleichsmaßstab und kommen zum Schluss, dass die Chancen im kleinen Wettbewerb größer sind, unter den 10% Besten (das experimentelle Kriterium zu gewinnen) zu landen, weshalb es sinnvoll ist, mehr Anstrengung in einen Wettbewerb zu stecken, wenn die Anzahl der Mitbewerber klein ist.

Von der Kontroverse zu einer Fallstudie

Erstens, das Experiment der Psychologen über Motivation und Anstrengung kann mit Lernenden durchgeführt werden, damit man authentische Daten hat. Auf einer 7-Punkt-Likert-Skala werden die Teilnehmer nach dem Grad ihrer Motivation und der Bereitschaft gefragt, ihr Bestes zu geben; die Bedingungen sind einmal, der Beste unter 10 Bewerbern zu sein, das andere Mal, unter 100 Bewerbern zu den 10 Besten zu gehören, mit anderen Worten, in beiden Fällen geht es darum, bei den 10% Besten zu sein. Zusätzlich werden die Teilnehmer gebeten, ihre Antworten zu begründen, ihre Chancen zu schätzen, einen Preis zu gewinnen (unter den Top 10% zu sein) und Faktoren anzugeben, welche diese Chancen beeinflussen könnten.

Zweitens, das rein probabilistische Rennen der Mathematiker wird erklärt. Entweder nutzt man die Binomialverteilung direkt oder bestimmt die Gewinnkurve durch Simulation. Aus dieser Kurve kann

man dann ablesen, dass der Bruchpunkt gerade etwas kleiner als 0,9 ist. Das bedeutet: hat man eine individuelle Stärke (im Sinn von Wahrscheinlichkeit) von mehr 0,9, einen zufälligen Gegner zu schlagen, so ist ein größerer Wettbewerb von Vorteil, liegt man unter 0,9, so gewinnt man im kleinen Wettbewerb mit größerer Wahrscheinlichkeit als im großen.

Das Modell ist rein auf Zufall aufgebaut, es schließt Formschwankungen, Motivation oder irgendwelche unabsehbaren Vorfälle im Bewerb aus. Vielleicht wissen Leute intuitiv, dass ihnen kleinere Gruppen bessere Chancen geben, unter den Besten zu sein?

3.4 Placebo-Effekt und Regression zur Mitte

Der Placebo-Effekt besagt, dass Menschen eine Erleichterung (von Schmerz oder einer Erkrankung) erfahren, wenn sie nur glauben, dass sie eine (wirksame) Therapie bekommen. Ein psychologischer Effekt, der heilen hilft? In diesem letzten Experiment erklären wir das bekannte Phänomen durch ein einfaches probabilistisches Arrangement, welches nur auf den Zufall – im Gegensatz zu einem psychologischen Gesetz – zurückgreift. Eine analoge Erklärung ergibt sich dabei auch für ein ‚Gesetz' der Vererbung – die Regression zur Mitte.

Hintergrund

Nach dem Placebo-Effekt erfahren Menschen einen objektiv feststellbaren Heileffekt von einem Medikament, auch wenn sie nur ein Placebo bekommen. Dabei ist ein Placebo ein Medikament oder eine Therapie, welche äußerlich dem Verum völlig gleich kommt, aber keinen arzneilichen Wirkstoff (oder eine echte medizinische Intervention) enthält. Das Placebo wird auch unter genau denselben Bedingungen verabreicht wie das Verum. Normalerweise wird in medizinischen Studien das Experiment ‚doppelblind' durchgeführt, d.h. weder Teilnehmer noch Arzt wissen, was wirklich verabreicht wird. Zusätzlich werden die Patienten zufällig der Placebo- bzw. Verum-Gruppe zugeordnet. Dies hat sich als goldener Standard in der medizinischen Statistik eingebürgert und soll sicherstellen, dass die Personen in beiden Gruppen so ähnlich wie möglich sind und dass ihre persönlichen Erwartungen, ein Medikament (oder keines) zu erhalten, keinerlei Einfluss hat auf die physisch beobachtbare Reaktion. Das erlaubt dann, die unterschiedlichen Werte aus dem Experiment direkt als Effekt der Behandlung zu interpretieren. Ähnlich wie im Gedächtnistest werden die Daten dazu verwendet, um die Frage zu beantworten „Ist Verum – signifikant – besser als Placebo?"

Regression zur Mitte ist ein weiteres ‚Gesetz'. Wir geben dazu ein Beispiel: Wenn statistische Einheiten unter zwei Bedingungen beobachtet werden – etwa *vor* und *nach* einer Unterrichtsintervention – dann sind zwar normalerweise jene, die vorher zu den Besten gezählt haben, normalerweise noch immer über dem Mittelwert (aller), aber sie sind viel weniger weit von der Mitte entfernt als vorher – sie ‚schreiten zurück zur Mitte'. Für die schlechtesten vorher gilt ähnliches mit umgekehrtem Vorzeichen.

Berühmt geworden sind die Väter und Söhne, die Francis Galton und Egon S. Pearson untersucht haben mit der Absicht zu zeigen, dass physische Merkmale erblich sind. Ein großes Forschungsvorhaben der zweiten Hälfte des 19. Jahrhunderts in England war vom Gedanken der Vererbung gekennzeichnet. Eigentlich wollte man zeigen, dass Intelligenz vererblich ist, aber die Konstrukte zur Messung von Intelligenz waren noch nicht erstellt, sodass man sich auf die Körpergröße (oder Augenfarbe) bezog. Mehr darüber kann man aus Freedman et al. (2007) erfahren. Väter, die nun zu den größten gehörten, hatten Söhne, welche überdurchschnittlich groß, aber weniger vom Mittelwert entfernt waren. Dies wurde als "law of regression" bezeichnet und man nannte den Parameter, mit dem der Grad der Vererblichkeit erfasst wurde, Regressionskoeffizient – ein Koeffizient, der misst, wie stark die Söhne zur Mitte zurück schreiten. Die Methode bekam davon ihren Namen.

Ein Experiment zu einer alternativen Interpretation entwickeln

Das folgende Experiment geht auf Dubben & Beck-Bornholdt (2010) zurück. Es zeigt, dass der Placebo-Effekt zum Teil durch reinen Zufall erklärt werden kann; gleichzeitig ergibt sich daraus, dass die Regression zur Mitte ein rein probabilistisches Artefakt und kein Gesetz über Vererblichkeit ist.

Wir werfen 42 Würfel ein erstes Mal und betrachten nur jene mit extremen Ergebnissen: wir erklären die 6er zu Tops, die 1er zu Flops. Nur mit diesen werfen wir ein zweites Mal. Eine einzelne Simulation des Experiments mag einen Mittelwert von 3,64 und 3,50 für die Flops bzw. Tops ergeben. In diesem Fall sind die Schlechtesten als Gruppe sogar besser als die Besten des ersten Wurfs. Das kann Glück oder Zufall sein. Daher wiederholen wir den Versuch oft (mindestens 500 Mal). Das zweistufige Experiment zeigt, dass die Besten und die Schlechtesten aus dem ersten Wurf im zweiten ziemlich vergleichbar sind mit der Tendenz, dass die Besten schlechter und die Schlechtesten besser werden.

Für den Placebo-Effekt merken wir an, dass Patienten gerade dann zum Arzt gehen, wenn sie sich schlecht fühlen (Flop im ersten Wurf) und sich dann erholen, was immer sie bekommen als Therapie (Verbesserung im zweiten Wurf). Die schlechte gesundheitliche Situation trifft auch auf Teilnehmer medizinischer Studien zu. Für die Regression zur Mitte merken wir an, dass die Besten und Schlechtesten des ersten Wurfes im zweiten näher zum Mittelwert liegen und die Schlechtesten sich verbessern, die Besten sich verschlechtern. Beide Phänomene können durch Bezug auf reinen Zufall erklärt werden. Bei der Regression zur Mitte ist der Effekt damit auf ein stochastisches Artefakt reduziert.

4. Diskussion und Schlussfolgerungen

In diesem Beitrag wurden psychologische Experimente als ein wertvoller und motivierender Hintergrund vorgestellt, um Stochastik zu lernen. Damit werden die behandelten Fallstudien in einen breiteren Rahmen der empirischen Forschung eingebettet. Wir schließen mit einigen Gedanken über die Rolle von Technologie und der begriffsbildenden Kraft des Modellierens.

4.1 Anlass und Argument zum Modellieren: psychologische Gesetze und Zufall

Psychologische Experimente bilden eine Quelle für Unterrichtseinheiten. Aus den Ergebnissen kann man versuchen, menschliches Verhalten durch allgemein gültige Gesetze zu beschreiben. Es ist motivierend, Widerspruch und Interesse, ausgelöst durch solche ‚Gesetze', mit Lernenden aufzuarbeiten.

Psychologen nutzen auch Musik (einzelne Töne hintereinander gespielt), Binärzahlen und Buchstaben, um Kurzzeitgedächtnis zu testen und so allgemeine Feststellungen zu treffen wie „7 plus oder minus 2". Trifft es wirklich zu, dass wir uns, grob gesprochen, nur 7 von zusammenhanglosen Informationseinheiten merken können? Telefonnummern haben üblicherweise 7 Ziffern, die abendländische Musik baut auf Tonleitern mit 7 Intervallen auf, sogar empirische Skalen wie die die populäre 7-Punkte-Likert-Skala verwenden mit Vorzug 7 Ausprägungen zur ‚Messung' von Einstellungen. Steht dahinter ein archaisches Gesetz, das unser Potential, Objekte zu unterscheiden und sich zu merken, beschränkt?

Andere Gesetze gehen gar nur von vier Informationseinheiten aus, was sich vielleicht darin ausdrückt, dass wir Telefonnummern in Gruppen von vier und drei Ziffern zerlegen. Lernen ist zu einem gewissen Grad dadurch charakterisiert, dass wir Verbindungen zwischen Objekten organisieren, egal ob sie versteckt sind und dann offen gelegt oder ob sie überhaupt erdacht werden. Damit überwinden wir Einschränkungen in der Merkfähigkeit. Mit den Memory-Karten etwa haben Kinder einen Vorteil, wenn sie sich statt der einzelnen Karten eine spannende Geschichte *zwischen* den Karten (welche die Lage zueinander berücksichtigt) ausdenken. Für Arbeitsgruppen hat man eine optimale Größe zwischen 8 und 12 ‚herausgefunden', was durch das abnehmende Engagement in größeren Gruppen belegt wird. Das mag mit dem angesprochenen *N*-Effekt zusammentreffen. (Wir haben dagegen dieses ‚Gesetz' durch Zufall erklärt.)

Solche Diskussionen und die Gegenüberstellung von Gesetzen aus einem wissenschaftlichen Hintergrund mit reinem Zufall motivieren und klären gleichzeitig, wie statistische Ergebnisse erhärtet werden, bis man von empirischer Evidenz sprechen kann.

4.2 Die Rolle der Technologie für den Modellbildungsansatz

Modellbildung als Ansatz, probabilistische Begriffe und die statistische Verfahrensweise zu erlernen, legt den Zweck, zu dem sie erdacht worden sind, frei und zeigt auf, wie Forschungsfragen und Fragen von alltäglicher Relevanz beantwortbar sind. Die Komplexität der Mathematik jedoch steigt gegenüber herkömmlichem Unterricht an, denn ein Wesenszug ist, dass die Modelle passen sollen und dass wesentlich unterschiedliche Modelle miteinander verglichen werden müssen. Mathematik schematisch anzuwenden verbietet sich gemäß den Zielen des Ansatzes. Modellbildung soll ja dazu beitragen, Entscheidungen in der realen Welt – nach geeigneten Bewertungskriterien – zu verbessern. Dazu muss man die Begriffe immer auf den Kontext übertragen und die abschließenden Antworten (Plural!) interpretieren und bewerten, auch im Hinblick darauf, ob sie nun den Zweck erfüllen. Das sind schwierige Aufgaben und sie verlangen mehr Kenntnisse über die Mathematik und über den Kontext.

Um den Ansatz zu unterstützen und um die technischen Details einzudämmen, kann und soll man von technologischen Hilfsmitteln Gebrauch machen. Abgestimmt auf das Niveau und die Erfahrungen der Lernenden wird man Optionen wie EXCEL, Fathom, REXCEL (ein Hybrid zwischen EXCEL und R), oder die Programmiersprache R wählen. Die zunehmenden Funktionalitäten des ursprünglich geometrisch ausgerichteten GeoGebra werden vielleicht die Präferenzen ändern. EXCEL hat wegen seiner weiten Verbreitung im Hinblick auf den zukünftigen Arbeitsplatz der Lernenden einige Vorzüge.

Technologie ermöglicht auch umständliche Berechnungen und graphische Darstellungen, welche eine zentrale Rolle als Antriebskraft zur weiteren Analyse übernehmen. Ferner bietet sie Zugang zur Methode der Simulation von unterstellten Hypothesen, um artifizielle Daten daraus zu erhalten. Dieser Ansatz erweist sich immer mehr – auch in Anwendungen – als unverzichtbar, weil Methoden und Aufgaben komplexer werden und sich mittels geschlossener Mathematik nicht mehr lösen lassen. Er hat aber auch eine Schwachstelle: Simulierte Ergebnisse erhalten – verstärkt durch graphische Präsentation – den Charakter von Fakten; man muss immer wieder darauf hinweisen und nachhaltig einwenden, dass diese Daten nur unter Szenarien entstanden sind und ihre Gültigkeit an das Zutreffen von Voraussetzungen gebunden ist. Modelle werden dabei auch immer öfter in einer ‚was wäre, wenn'-Manier verwendet. Man spielt einfach durch, was passiert, wenn man dieses oder jenes Modell unterstellt. Diese Verwendung ähnelt der Idee von Szenarien; mehr dazu findet man in Borovcnik (2009).

4.3 Das formative Potential von Modellbildung

Prozesse zur Modellbildung sind ganz global gesehen ein Verbindungsglied zwischen den Lernenden, den realen Situationen, die damit durchdrungen werden, und der Mathematik, welche die Begriffe dazu bereitstellt. Solche Prozesse lassen die reale Problematik besser verstehen und füllen die Begriffe mit Leben. Wissenschaftliche Konzepte sind ja immer dazu entworfen worden, Zwecke zu erfüllen und sie entsprechend einzusetzen. Diese Zwecke zu kennen, orientiert beim Lernen und Verstehen.

Obwohl Kapadia & Borovcnik (1991) „chance encounters" aus unterschiedlichen Perspektiven analysiert haben, was Mathematik und Verständnis im Lernprozess mit einschließt, haben sie das Hauptaugenmerk (noch) nicht auf Modellieren gelegt. Die Interaktion zwischen Modellbildung und Simulation und das Potential daraus für Lernen von Stochastik ist hingegen ein zentrales Anliegen von Chaput & Girard (2008). Blum (2012) spricht von der *formativen Kraft* des Modellbildens: Durch Modellbildungsprozesse werden wesentliche Kompetenzen aufgebaut, der Wert und die Bedeutung der Begriffe wird nicht nur veranschaulicht sondern gestaltet. Modellbildung als Ansatz zeigt die Beteiligten in Aktion und trägt dazu bei, Lernende zu motivieren, sich den Herausforderungen zu stellen.

Literatur

Blum, W. (2012): *Quality teaching of mathematical modelling – what do we know, what can we do?* Plenary lecture at ICME 12, Seoul.

Borovcnik, M. (o.J.): *Materialien zur Stochastik*. wwwg.uni-klu.ac.at/stochastik.schule/Boro/index_inhalt (Zugriff: 21.9.2012).

Borovcnik, M. (2009): Aufgaben in der Stochastik – Chancen jenseits von Motivation. *Didaktik-Reihe der Österreichischen Mathematischen Gesellschaft*, 42, 1–23. www.oemg.ac.at/DK/Didaktikhefte/ (Zugriff: 21.9.2012).

Borovcnik, M., Kapadia, R. (2011): Modelling in probability and statistics – key ideas and innovative examples. In J. Maaß, J. J. O'Donoghue (Eds.): *Real-World Problems for Secondary School Students–Case Studies* (1–44): Rotterdam: Sense.

Borovcnik, M., Schenk, M. (2011): Simulationen im Stochastik-Unterricht. *Didaktik-Reihe der Österreichischen Mathematischen Gesellschaft*, 44, 1–16. www.oemg.ac.at/DK/Didaktikhefte/ (Zugriff: 21.9.2012).

Chaput, B., Girard, J. C., Henry, M. (2008): Modeling and simulations in statistics education. In C. Batanero, G. Burrill, C. Reading, A. Rossman (Hsg.): *Joint ICMI/IASE Study: Teaching Statistics in School Mathematics. Challenges for Teaching and Teacher Education.* Monterrey: ICMI and IASE. www.stat.auckland.ac.nz/~iase/publications.php (Zugriff: 21.9.2012).

Dubben, H.-H., Beck-Bornholdt, H.-P. (2010): Mit an Sicherheit grenzender Wahrscheinlichkeit. Logisches Denken und Zufall. Reinbek bei Hamburg: Rowohlt.

Ehrenberg, A.S.C. (1981): *Data reduction*. New York: Wiley.

Freedman, D., Pisani, R., Purves, S. (2007): *Statistics*. Fourth edition. London: Norton.

Garcia, S. M. Tor, A. (2009): The *N*-Effect. More competitors, less competition. *Psychological Science*, 20(7), S. 871–877. www.sitemaker.umich.edu/stephen.garcia/files/the_n-effect.pdf (Zugriff: 21.9.2012).

König, G. (2011): Motivation bei Wettbewerben: Stochastische Aspekte einer Diskussion in der Zeitschrift „Psychological Science". In: *Stochastik in der Schule* 31(3), 22–25.

Miller, G. (1956): The magical number seven, plus or minus two: Some limits on our capacity for processing information. In: *The Psychological Review*, 63(2), 81–97. www.musanim.com/miller1956 (Zugriff: 21.9.2012).

Mittag, H.-J. (n.d.): *Virtuelle Bibliothek mit neuen interaktiven Experimenten für die Statistikausbildung*. www.fernuni-hagen.de/jmittag/bibliothek/index.php (Zugriff: 21.9.2012).

Mukherjee, K. Hogarth, R. M. (2010): The *N*-effect: Possible effects of differential probabilities of success. In: *Psychological Science* 21(5), 745–747. pss.sagepub.com/content/21/5/745.extract (Zugriff: 21.9.2012):

Popper, K. (2005, 1935): *Logik der Forschung* (Hrsg. H. Keuth, M. Siebeck), Tübingen. Original: Wien: Springer 1935. https://www.uni-rostock.de/fakult/philfak/fkw/iph/strobach/hroseminare/modul/popper.html (Zugriff: 21.9.2012).

Richardson, M., Reischman, D. (2011): The magical number 7. In: *Teaching Statistics*, 33(1), 17–19. (Deutsch in: *Stochastik in der Schule*, 31 (3), 26–29.)

Tor, A. Garcia, S. M. (2010): The *N*-Effect: Beyond winning probabilities. In: *Psychological Science*, 21, 748–749. www.sitemaker.umich.edu/stephen.garcia/files/the_n-effect_reply.pdf (Zugriff: 21.9.2012).

Verfasser

Manfred Borovcnik
Alpen-Adria-Universität Klagenfurt
Institut für Statistik
Universitätsstraße 65
9020 Klagenfurt
`manfred.borovcnik@uni-klu.ac.at`

[1] Die generische Verwendung von Substantiva macht einen Text einfach viel besser lesbar. Mit Lehrern und Schülern sind selbstverständlich Personen beiderlei Geschlechts angesprochen.